5-1-2009

A Comparative Study of Bayesian Model Selection Criteria for Capture-Recapture Models for Closed Populations

Ross M. Gosky
Appalachian State University, goskyrm@appstate.edu

Sujit K. Ghosh
North Carolina State University, sujit_ghosh@ncsu.edu

A Comparative Study of Bayesian Model Selection Criteria
for Capture-Recapture Models for Closed Populations

Ross M. Gosky   Sujit K. Ghosh
Appalachian State University   North Carolina State University

Capture-Recapture models estimate unknown population sizes. Eight standard closed population models exist, allowing for time, behavioral, and heterogeneity effects. Bayesian versions of these models are presented and use of Akaike's Information Criterion (AIC) and the Deviance Information Criterion (DIC) are explored as model selection tools, through simulation and real dataset analysis.

Key words: AIC, Bayesian inference, capture-recapture models, closed population, DIC, Gibbs sampling, heterogeneity, MCMC, model selection, WinBUGS.

Introduction

For capture-recapture experiments involving closed populations, likelihood-based models based upon the multinomial distribution are commonly used, and a thorough treatment of these models is given by Otis, Burnham, White, and Anderson (1978). These models allow animal capture probabilities to vary based on three types of effects: time effects, heterogeneity effects, and behavioral effects. Time effects occur when capture probabilities vary by capture period. Heterogeneity effects occur when capture probabilities vary by animal. Behavioral effects occur when an animal's capture probability changes after they are captured for the first time.

This effect is called a trap-happy effect when the capture probability increases after initial capture, and is called a trap-shy effect

Ross Gosky is an Assistant Professor in the Mathematical Sciences Department. His research interests are in Bayesian Statistics and Mark-Recapture models. Email: goskyrm@appstate.edu. Sujit Ghosh is a Professor in the Department of Statistics. His research interests include Bayesian Statistics, Spatial Statistics, and Survival Analysis. Email: sujit_ghosh@ncsu.edu.

when the capture probability decreases after initial capture. Denoting subscripts t, h, and b to refer to time, heterogeneity, and behavioral effects, respectively, eight models have been developed, with the model subscripts indicating which effects are present in the modeling of capture probabilities. The goal of each model is to estimate the unknown population size N. The model M_0 denotes a model which has none of the three effects. Model M_t contains time effects, model M_h contains heterogeneity effects, and model M_b contains behavioral effects. Models M_{tb}, M_{bh}, M_{th}, and M_{tbh} are complex models accounting for variation in capture probabilities from each listed effect. Chao (2001) provides an overview of closed population models as well.

Pledger (2000) discussed using mixture models to fit heterogeneity effects in capture-recapture data, and discussed use of Akaike's Information Criterion (AIC) as a model selection tool. Caution in using heterogeneity models is necessary, though, as Link (2003) showed that estimates of N under M_h models are highly dependent upon the assumed distribution of capture probabilities in the population. He refers to the parameter N as non-identifiable in heterogeneity models because different, reasonable, models may fit the data equally well but give very different inferences about N. Link's results imply that distinguishing between different heterogeneity models may never be possible. However, it remains plausible that

estimates of N from M_h models are more accurate than those from Model M_0, for example, in populations with heterogeneity.

Program MARK (see http://welcome.warnercnr.colostate.edu/~gwhite/mark/mark.htm), provides estimates of N for these closed population models as well as end-user flexibility in the specific parameterization of the models. For example, the mixture models of Pledger (2000) can be fit in Program MARK with different numbers of mixture groups specified by the user. Program MARK also provides model selection functionality based on Akaike's Information Criterion (Akaike, 1973).

Bayesian versions of closed population models have also been presented. Early approaches focused on Model M_t, such as Castledine (1981), and George and Robert (1992). Ghosh and Norris (2005) presented a Bayesian version of M_{bh}, and M_h, M_b, and M_0 as special cases of this model. Furthermore, they presented a model selection approach based upon a criterion proposed by Gelfand and Ghosh (1998). Other recent work on Bayesian models have been presented by Durban and Elston (2005) and by King and Brooks (2008). King and Brooks recommended Bayesian Model Averaging and Reversible Jump Markov Chain Monte Carlo (RJMCMC) methods for recapture/recovery data analyses, while Durban and Elston focused on Model M_{th} by adapting a log-linear modeling approach to the models of Agresti and Coull (1999). More recently, Gosky and Ghosh (2011) provide Bayesian estimation methodologies for all eight models.

Methodology

Bayesian Closed Population Capture-Recapture Models

Bayesian statistical modeling requires the development of the likelihood function of the observed data, given a set of parameters, as well as the joint prior distribution of all model parameters. A major benefit of Bayesian models for capture-recapture data is that Bayesian estimates of N, from its posterior distribution, are easily obtainable and this posterior distribution gives appropriate measures of variability for estimating N. Even when non-informative prior distributions are used for model parameters, these estimates of variability are not based on asymptotic criteria and hold when N and the number of capture periods are relatively small (e.g., see Gosky and Ghosh, 2011). Bayesian modeling also allows for the possibility of using informative prior information about model parameters, if available.

The approach to modeling heterogeneity used is identical to that presented in Ghosh and Norris (2005), using a finite-mixture approach to heterogeneity rather than utilizing a continuous distribution to model individual capture probabilities. This basic idea was introduced by Norris and Pollock (1996) and discussed further in Pledger (2000), and has been shown to be effective in modeling heterogeneity.

Let k represent the number of capture periods in the study. Define indicator variables $X_{ij} = 1$ if animal i is captured during capture period j, for i = 1, 2, ..., N and j = 1, 2, ..., k. Also denote $p_{ij} = Pr(X_{ij} = 1)$ as the probability that animal i is captured during capture period j.

Denote the capture matrix **X** with dimensions N x k with entry X_{ij}. Denote X[i,.] as the ith row of **X**, a vector with 2^k possible values because each entry in the vector is zero or one. For simplicity, these outcomes can be ordered as

Outcome 0: capture history (0, 0, 0, ..., 0, 0, 0);

Outcome 1: capture history (0, 0, 0, ..., 0, 0, 1);

Outcome 2: capture history (0, 0, 0, ..., 0, 1, 0);

Outcome 3: capture history (0, 0, 0, ..., 0, 1, 1);

through Outcome 2^k - 1: capture history (1, 1, 1, …, 1, 1, 1).

Each animal in the population has exactly one of the 2^k capture histories. Noting that $(X_{i1}, X_{i2}, ..., X_{ik})$ represents the observed capture history of any animal in the population, Capture History h_i is defined corresponding to the previous ordering of outcomes, as $h_i = \sum_{j=1}^{k} X_{ij} 2^{k-j}$. Notice that each h_i takes values in the set 0, 1, ..., 2^k - 1. Denote Z_l as the number of animals with capture history l, for l = 0, 1, ..., 2^k

- 1, i.e., $Z_l = \sum_i I(h_i = l)$ where I(.) denotes the indicator function which takes the value 1 if $h_i = l$ and takes the value 0 otherwise. Note that Z_0, the number of animals with capture history (0, 0, ..., 0), cannot be observed. Also, note that $\sum_{l=0}^{2^k-1} Z_l = N$. Denote $S = N - Z_0$ as the number of animals observed during at least one capture period. Denote P_l as the probability of animal i having capture history l. Then

$$P_l = \prod_{i:h_i=l} \prod_{j=1}^{k} p_{ij}^{x_{ij}} \left(1 - p_{ij}\right)^{1-x_{ij}}. \quad (2.1)$$

Defining $L = 2^k - 1$, the joint distribution of (Z_1, Z_2, ..., Z_L) is

$$\Pr\left[Z_1 = z_1, ..., Z_L = z_L \mid N, P\right]$$

$$= \frac{N!}{(N-S)! \prod_{l=1}^{L} Z_l!} \prod_{l=1}^{L} P_l^{Z_l} \left(1 - \sum_{l=1}^{L} P_l\right)^{N-S}$$

$$(2.2)$$

where $\mathbf{P} = (P_1, P_2, ..., P_L)$. Note that if N were known, this model would represent a multinomial likelihood function with counts Z_0, ..., Z_L and probabilities P_0, ..., P_L. However N is unknown and it is the main parameter of interest. From equation (2.2) it follows that the likelihood function of (N, **P**) is given by

$$L(N, \mathbf{P}|Z) \propto \binom{N}{S} \prod_{l=1}^{L} P_l^{Z_l} \left(1 - \sum_{l=1}^{L} P_l\right)^{N-S} \quad (2.3)$$

where $Z = (Z_1, ..., Z_L)$ denotes the set of observed counts, which turns out to be the minimal sufficient statistic for this model.

It is of interest to estimate N, treating **P** as a nuisance parameter. The capture probability vector **P** varies depending on the specific model. A Bayesian modeling framework was adopted for each of the eight models, where

$$\Pr(N = n) \quad \propto \quad \frac{1}{n^\delta} \quad , \text{n} = 1, 2, ..., \text{N}_{\max}$$

was used as the prior distribution for N, with $\delta > 0$ fixed at a specific value and $N_{\max}$ fixed at a realistic upper bound for N. A non-informative prior distribution can be obtained with $\delta = 0.5$ (or alternatively $\delta = 1$) and a uniform prior is obtained with $\delta = 0$. The final estimate of N is obtained from the marginal posterior distribution of N by integrating out the parameters corresponding to **P**. The most complex model, M_{tbh}, is introduced first followed by descriptions of each of the other seven models as special cases of M_{tbh}.

Model M_{tbh}

This model allows for individual heterogeneity, time, and behavior effects. For heterogeneity, a finite mixture distribution is used, representing m possibly distinct groups within the population. Behavioral effects are modeled as constant across each of the m groups and across capture periods 2 through k to minimize the number of model parameters and to allow the model to be fit to studies with a minimal number of capture periods.

Denoting $\tau_{ij} = 1$ if animal i has been captured before capture period j, then for each group, the capture probability vector is $p_i = p_{i1} I(\tau_{ij} = 0) + p_{i2} I(\tau_{ij} = 1)$ and $(p_{i1}, p_{i2}) \sim F(.)$, described next. A finite mixture distribution is assumed for the 2k-dimensional distribution function, F, specifically $dF(p) = \sum_{m=1}^{r} \pi_m I(p = \theta_m)$, where π_m denotes the probability at support point $\theta_m = (\theta_{11m}, ..., \theta_{1km}, \theta_{21m}, ..., \theta_{2km})^T$, and $\sum_{m=1}^{r} \pi_m = 1$. The probability of initial capture in capture period j is represented as θ_{1jm}, where j = 1, 2, ..., k within population group m, where m = 1, 2, ..., r. Similarly, θ_{2jm} is the probability of subsequent capture in capture period j within population group m. As previously stated, the behavior effect is constant across the capture periods and the m population groups. Thus $\theta_{2jm} = \theta_{1jm} + c$ for j = 2, ..., k and m = 1, 2, ..., r. Furthermore, $\theta_{2jm} = 0$ for m = 1, 2, ..., r because subsequent capture is impossible in capture period one. Fixing r = 2 mass points representing possibly two distinct population groups implies that $1-\pi_2$

$= \pi_1 = \pi$. Prior distributions for π and θ_{1jm} are π, $\theta_{1jm} \sim$ Beta(a,b) for $j = 1, \ldots, k$ and $m = 1, 2$.

A conditional prior distribution of c given θ_{1jm} for $j = 1, \ldots, k$ and $m = 1, 2$ is Uniform(-min$_{2 \leq j \leq k; 1 \leq m \leq r}$ θ_{1jm}, 1-max$_{2 \leq j \leq k; 1 \leq m \leq r}$ θ_{1jm}). This mixture model requires restrictions for identifiability of all model parameters, so $\theta_{1j1} \leq \theta_{1j2}$ for $j = 1, 2, \ldots, k$ and $\theta_{2j1} \leq \theta_{2j2}$ for $j = 2, 3, \ldots, k$ is set.

Model M_{tb}

Restrict $\pi_1 = 1$ from Model M_{tbh}.

Model M_{th}

Restrict $\theta_{2jm} = \theta_{1jm}$ for $j = 2, \ldots, k$ and $m = 1, \ldots, r$, from Model M_{tbh}.

Model M_{bh}

Restrict $\theta_{11m} = \theta_{12m} = \ldots = \theta_{1km}$ and $\theta_{22m} = \theta_{23m} = \ldots = \theta_{2km}$ for $m = 1, 2, \ldots, r$ from M_{tbh}. Rather than modeling $\theta_{2jm} = \theta_{1jm} + c$, choose prior distributions θ_{11m}, $\theta_{22m} \overset{i.i.d.}{\sim}$ Beta(a, b) for $m = 1, \ldots, r$. Fixing $r = 2$ mass points as described in M_{tbh}, restrict $\theta_{111} \leq \theta_{112}$ and $\theta_{221} \leq \theta_{222}$ for identifiability of all model parameters.

Model M_t

From Model M_{th}, restrict $\pi_1 = 1$.

Model M_h

From Model M_{bh}, restrict $\theta_{22m} = \theta_{11m}$ for $m = 1, 2, \ldots, r$.

Model M_b

From Model M_{bh}, restrict $\pi_1 = 1$.

Model M_0

Restrict $\pi_1 = 1$ from Model M_h.

The number of parameters in each model as a function of r, the number of support points of the finite mixture distribution F and k, the number of capture periods, is determined from the preceding model descriptions. For example, Model M_{tbh} has parameters N, π_1, ..., π_{r-1}, θ_{111}, ... , θ_{1k1}, θ_{112}, ... , θ_{1k2}, θ_{11r}, ... , θ_{1kr}, and c. The number of parameters is thus $1 + (r-1) + kr + 1 = r(k + 1) + 1$. Similarly it is established that M_{th} has $r(k + 1)$ parameters, M_{bh} has 3r parameters, M_{tb} has $k + 2$ parameters, M_h has 2r parameters, M_b has 3 parameters, M_t has $k + 1$ parameters, and M_0 has 2 parameters.

Posterior distributions of the model parameters for all eight models can be closely approximated using Markov Chain Monte Carlo (MCMC) methods available in the WinBUGS V1.4 software package (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml).

Model Selection Methods

Because eight possible models exist for a given closed population data set, definitive methods for model selection are necessary in such analyses. The eight models are generally (though not exclusively) nested, ranging from very simple models (M_0) to complex models (M_{tbh}). Model selection criteria allow the best model of the eight to be fit to the data. Striking a balance and finding a model that neither under-fits nor over-fits the data is the motivation for model selection criteria (Burnham & Anderson, 2002).

Akaike's Information Criterion (AIC) is one such method of model selection, and seems to be the most commonly used criterion for model selection. The intent of AIC is to measure the mathematical distance between the true population and the fitted model, by using the so-called Kullback-Leibler discrepancy. To differentiate between models with different numbers of parameters, AIC adds two times the number of model parameters to the estimated Kullback-Leibler discrepancy. Thus, when two models of differing complexity fit a data set equally well AIC chooses the simpler model by penalizing the complex model for having more model parameters. The rule of parsimony says that a researcher should choose the simplest model that adequately describes the behavior of the population. Use of AIC generally supports this rule.

As the models are nested, AIC is examined as model selection tool. However, AIC is not asymptotically consistent in the sense that the probability that it chooses the correct model (given that the data has been generated from the correct model) does not converge to one as the sample size tends to infinity (Schwartz, 1978). A modified version of AIC within the Bayesian framework is used for model selection.

The Deviance Information Criterion (DIC) is strictly a Bayesian model selection criterion (Spiegelhalter, et al., 2002), which is structured similarly to AIC. The main difference between AIC and DIC is in the penalty term added to the estimated Kullback-Leibler discrepancy. DIC adds two times the effective number of parameters to the estimated Kullback-Leibler discrepancy. The effective number of parameters is a Bayesian concept. It recognizes that the number of parameters in a Bayesian model is influenced by the prior distributions of these parameters. The DIC criterion presents a methodology to measure this number of parameters. The DIC is, then, the difference between the estimated mean KL distance, and the KL distance estimated at the posterior mean of each of the model parameters.

DIC is also examined as a model selection criterion for these models. Use of DIC does not require the models to be nested. However, the modeling herein uses a mixture approach for heterogeneity models, and there are some questions about use of DIC for mixture models. Some recent suggestions have been made regarding these problems (see Celeux, et al., 2006).

Use of the Bayesian Information Criterion (BIC) was considered for model selection, but, for capture-recapture models the sample size is unclear (as N itself is a parameter and k, the number of capture periods, is usually much smaller than necessary for asymptotic properties to work). Therefore, AIC and DIC are focused on as potential model selection criteria; specifically, it is assessed whether AIC and DIC choose the correct model for a given data set.

A model selection criterion proposed by Gelfand and Ghosh (1998) is based upon minimizing the squared predictive error of the observed data, where the predictive distribution of the observed data is based partially upon the posterior distribution of the parameters, given the observed data, rather than on the prior distribution of the parameters. Ghosh and Norris (2005) discussed using this method for Model M_{bh}, and their findings were promising. This criterion is an area of future research, as it easily allows non-nested models to be directly compared and it balances between model fit and model complexity.

Results

Data Generation Process and Bayesian Analysis Method

This simulation consists of eight experiments. Experiment one contains 100 data sets generated under each modeling assumption (M_0, M_t, M_h, ..., M_{tbh}). Experiments two through eight each contain 50 data sets generated under each modeling assumption. Each experiment uses Markov Chain Monte Carlo (MCMC) methods to fit each data set using each of the eight models. Thus, experiment one consists of 6,400 analyses (800 data sets each analyzed under eight models). Experiments two through eight consist of 3,200 analyses (400 data sets analyzed under eight models). Each data set is a simulated capture-recapture study with $k = 5$ capture periods. The methodology used to generate p_{ij} values is illustrated in Table 1, and detailed information regarding the data generating parameters is provided in an Appendix available at http://www.mathsci.appstate.edu/~rmg/.

Calculations of p_{ij} for M_{tbh}, the most complex model, are computed as $F(\mu + \beta_j + \eta\tau_{ij} + \kappa Z_i)$, where F is the Logistic distribution function

$$F(x) = [1 + e^{-1}]^{-1}, Z_i \overset{i.i.d.}{\sim} N(0,1),$$

where $\tau_{ij} = 1$ if the animal has been previously captured, and $\tau_{ij} = 0$ otherwise.

The approach in Table 1 resembles a 2^{5-2} fractional factorial design with factors N, Average p_{ij}, and magnitude of time, behavioral, and heterogeneity effects. Means and standard deviations of the p_{ij} for each simulation experiment are listed in the Appendix.

For each data set, and under each model, an estimate of the posterior density of N was constructed using WinBUGS Version 1.4. The median of this posterior distribution, denoted $\hat{N}$, was chosen to estimate N and AIC and DIC were also computed. For these simulation experiments, non-informative prior distributions were chosen for the model parameters. Specifically, $\delta = 0.5$ was chosen as the hyperparameter for the prior distribution of N; r = 2 was selected for the number of support points for F, and a = b = 0.5 for hyperparameters

72

Table 1: Data Generating Assumptions for Simulation Experiments 1 to 8

Experiment Number	N	Average p_{ij}	Time Effects	Behavioral Effects	Heterogeneity
1	500	0.2	Large	Positive	Large
2	500	0.2	Small	Positive	Small
3	500	0.4	Large	Negative	Large
4	100	0.4	Large	Positive	Small
5	100	0.4	Small	Positive	Large
6	100	0.2	Large	Negative	Small
7	500	0.4	Small	Negative	Small
8	100	0.2	Small	Negative	Large

for the prior distributions of all the capture probabilities.

A burn-in period of 3,000 samples was used to allow convergence of the MCMC processes to a stable distribution. After the burn-in period, 2,000 samples were selected from each of three MCMC chains with dispersed starting values for the model parameters. Therefore, posterior distribution estimates are based upon 6,000 total samples. Convergence of the models was checked through the Gelman-Rubin statistic in WinBUGS. Table 2 shows means and percentages of times each model was selected by the MCMC estimates of AIC for Experiment one.

Analysis of AIC as a Model Selection Criterion

AIC (Akaike, 1973) has been used extensively as a model selection tool. Calculation of AIC adds a parameter penalty to the estimated Kullback-Leibler Discrepancy between the fitted model and the true model. Using θ as a general term to represent all the model parameters (e.g. θ = (N, **P**) as in Equation 2.3), X as a general term to represent the observed data (e.g. X = (Z_1, …, Z_L) as in Equation 2.3), p' as the number of model parameters, and LogL as the log likelihood function, a form for calculation of AIC is given by

$$AIC = -2LogL(\hat{\theta} \mid X) + 2p' \qquad (4.1)$$

where $\hat{\theta}$ is the MLE of θ under the assumed model. However, the AIC calculation used here is different from the usual form of AIC. Defining

$$D(\theta) = -2LogL(\theta \mid X), \qquad (4.2)$$

use AIC = E[D(θ)|X] + 2p' where E[D(θ)|X] is the mean of the posterior distribution of D(θ).

Analysis of Table 2, which gives the MCMC AIC means and model selection percentages for simulation experiment one, indicates that overall the AIC is effective in determining the correct model. For the first seven columns in the table, the minimum AIC mean occurs when the fitted model matches the data generating assumptions. This suggests that AIC is capable of identifying the correct model, on average.

Perhaps more indicative of the performance of AIC is the percentage of times it chooses the correct model. For this analysis, a model with the minimum posterior mean of AIC was chosen for a given data set. When a tie occurs between two models, the simpler model is chosen. Ideally, the diagonal entries in the table should have the highest percentages of selections by AIC. The table columns represent the true model generating assumptions.

Selection of a different model from the data generating assumptions may be called a model selection error, and the percentage of AIC

Table 2: Simulation Experiment One Average AIC Posterior Mean and AIC
Model Selection Percentages

Model Fit	True Model							
	M_0	M_h	M_t	M_b	M_{bh}	M_{tb}	M_{th}	M_{tbh}
	Avg AIC (Top Line) AIC% (Bottom Line)							
M_0	134.3 92%	197.2 0%	320.7 0%	191.8 0%	252.4 0%	533.4 0%	354.3 0%	539.0 0%
M_h	138.4 0%	159.2 95%	325.1 0%	194.4 0%	196.6 0%	537.2 0%	328.5 0%	501.4 0%
M_t	142.4 1%	205.4 0%	136.2 87%	173.2 0%	236.2 0%	164.6 2%	198.1 0%	239.7 0%
M_b	136.4 7%	197.9 0%	208.0 0%	157.9 99%	201.6 0%	227.2 0%	193.6 0%	181.3 0%
M_{bh}	142.5 0%	163.0 5%	214.8 0%	163.6 1%	165.5 99%	234.1 0%	184.5 2%	174.5 13%
M_{tb}	144.2 0%	203.8 0%	138.0 13%	165.8 0%	207.1 0%	151.1 97%	186.1 3%	174.5 11%
M_{th}	153.8 0%	175.0 0%	147.7 0%	177.0 0%	181.1 1%	161.9 2%	166.7 87%	165.1 36%
M_{tbh}	155.6 0%	176.6 0%	149.4 0%	177.1 0%	179.8 0%	162.1 0%	168.5 8%	163.5 42%
Error %	8%	5%	13%	1%	1%	3%	13%	58%

model selection errors is also listed in the last row of Table 2. In this respect, for seven of the eight models, AIC performs well. Among these seven models, for M_t and M_{th}, the percentage of selections is 87 percent, which is somewhat lower than for the other models. When M_t and M_{th} are not selected by AIC, though, AIC selects a similar model, but with more effects. This is better than the selection of an unrelated model. Model M_{tbh} does not perform as well. Data generated under the assumptions of M_{tbh} only had a 42% selection rate by AIC. When M_{tbh} was not selected in this column, the model selected was one of the sub-models containing two of the effects (M_{th}, M_{bh}, and M_{tb}).

Some of this could be due to relative weighting of the time, behavioral, and heterogeneity effects within M_{tbh}, as AIC may be picking the model based on the most significant of these effects present in any particular data set

Furthermore, with five capture periods, model M_{tbh} may be somewhat over-parameterized. Thirty-one distinct capture histories were observed, and model M_{tbh} includes 13 parameters for such data, which may lead to the estimation of effects due only to random chance. However, from an overall look at this table, it is concluded that AIC performs well as a model selection tool.

A summary of the AIC selection rates from Experiments two through eight is given in Table 3, which lists only whether AIC chose the correct model in each experiment. Thus, the 94% entry in the first row and column shows that in experiment two, AIC chose model M_0 correctly 94% of the time for the data sets generated with constant capture probability.

In column one of Table 3, a strong AIC selection rate is observed for M_0 data for all

Table 3: Selection Rates for AIC for Simulation Experiments 2 to 8 Selection
Rates for Data Sets Generated Via Listed Model Assumptions

Experiment	M_0	M_b	M_t	M_h	M_{bh}	M_{tb}	M_{th}	M_{tbh}
2	94%	96%	78%	0%	86%	0%	0%	0%
3	90%	96%	90%	92%	98%	100%	94%	70%
4	92%	96%	82%	6%	62%	0%	0%	0%
5	94%	92%	36%	78%	12%	52%	2%	0%
6	84%	48%	62%	2%	8%	0%	0%	0%
7	96%	100%	88%	0%	76%	4%	0%	0%
8	92%	44%	18%	54%	6%	20%	0%	0%

experiments. Column two shows strong AIC selection rates for the M_b data sets, with the exception of experiments six and eight. These two experiments had relatively small capture probabilities, smaller population sizes of N = 100, and negative behavioral effects. This combination of factors makes detection of behavioral effects difficult due to small observed numbers of recaptures.

Column three shows strong selection rates for M_t data, except for experiments five and eight. However, both of those experiments have small-magnitude time effects and smaller population sizes N = 100, indicating a simpler model such as M_0 may be more appropriate for the data. Experiments two and seven, which also have small time effects, but a larger population size N = 500 show larger AIC selection percentages of 78% and 88% respectively.

Column four shows that M_h data has reasonably high selection percentages for experiments three, five, and eight, and small selection percentages for experiments two, four, six, and seven. The low selection rates occur in experiments with small heterogeneity effects in the data. It appears AIC selects a heterogeneity model when the heterogeneity effects are large, but not when they are relatively small.

Column five shows that M_{tb} data has a high AIC selection rate in experiments two, three, and seven. M_{tb} data has a moderate AIC selection rate of 62% in experiment four, and has low selection rates in experiments five, six, and eight. The low selection rates occur in

experiments with small time effects. The moderate selection rate in experiment four occurs when time effects in the data are large. Some examination of this case shows that when M_{tb} is not chosen, one of the submodels M_b or M_t is chosen by AIC.

The low selection percentages for experiments two, four, six, and seven for the heterogeneity models M_{bh}, M_{th}, and M_{tbh} occur due to the small heterogeneity effects in those experiments, and this again reflects that comparable models without heterogeneity effects can adequately fit the data. Low selection rates for M_{bh}, M_{th}, and M_{tbh} are seen in experiment eight, and for the M_{th} and M_{tbh} data for experiment five. These low rates occur for experiments where heterogeneity effects are large. However, due to the small magnitude time effects in experiments five and eight, for the M_{th} data AIC chooses M_b and M_{bh} most often as the best model, reflecting adequate fit for these data sets by simpler models. Some examination of the underlying results (not available in Table 3) shows that the penalty term for the number of parameters is the reason that M_{th} has a higher AIC value for these data sets.

For the M_{tbh} data sets, AIC chooses Model M_{bh} most commonly, followed by model M_b. The choice of M_{bh} again reflects the small magnitude of the time effects in these data sets. The choice of model M_b is surprising given that the heterogeneity in the data is strong in experiment five. However, a behavioral effect and a heterogeneity effect are not completely

unrelated. For capture periods two through k, the behavioral effect creates two distinct groups in the population: those which have been previously captured and those which have not been previously captured. Each group has separate capture probabilities. Although group membership is changing with each capture period, Model M_b could provide a reasonable fit to data with heterogeneity in some instances.

Finally, for experiment eight, for the M_{bh}, M_{th}, and M_{tbh} data sets, no particular model is selected overwhelmingly, and the true model is also rarely selected for these data sets. This may reflect the combination of small population size of $N = 100$, the negative behavioral effects, the average capture probabilities being 20%, and the large degree of heterogeneity in the data. For a small population, it is difficult to have one data set reflect all those sources of variation, causing problems for a selection criterion such as AIC.

Overall the performance of AIC as a model selection method for these models is encouraging. However, it is recommended that AIC be a guide to select a subset of suitable models for further analysis. Although AIC performed well in selecting the true model when the degree of underlying time, heterogeneity, or behavioral effects was large, the performance when these effects were small means that another model is selected. For this reason, it is recommended that AIC be used to narrow the set of eight models down to a smaller number of candidate models. A more detailed analysis involving other factors, such as the opinion of a subject matter expert, should be used to make the final model choice.

Analysis of DIC as a Model Selection Criterion

The DIC criterion is a recent development in model selection. DIC can be expressed similarly to AIC. Given the common use of AIC, this feature allows users to quickly understand the form and use of DIC. Additionally, DIC is easy to calculate, as it is a function of the posterior parameters and the model deviance (where deviance is related to the log-likelihood).

Using the same notation as in the definition of AIC, and again denoting $D(\theta) = -2LogL(\theta |X)$, and defining $pD = E(D(\theta)|X) - D($

$\hat{\theta})$, where $E(D(\theta)|X)$ represents the posterior mean of $D(\theta)$, it is possible to calculate

$$\text{DIC} = D(\hat{\theta}) + 2pD, \qquad (4.3)$$

where $\hat{\theta}$ is a posterior estimate of θ, e.g., $\hat{\theta} = E[\theta |X]$ or Median[θ |X]. As stated previously, the pD term in DIC represents an effective number of parameters. The pD term measures the decrease in the deviance (increase in the likelihood) obtained by using posterior estimates of the parameters θ. Note that although DIC is structured to look like AIC, the penalty term is actually a function of the model fit, not simply a discrete number of parameters.

For computational purposes, Dev(θ) is defined as the MCMC computed deviance for any particular data set and model combination and $\overline{D}$ as the MCMC mean of the deviance statistic, pD is computed as $pD = \overline{D} - \text{Dev}(\hat{\theta})$ and computationally, results in DIC = Dev($\hat{\theta}$) + 2pD.

In the simulations, DIC did not perform as well as AIC in model selection. For several of the models, most notably M_{bh}, the pD penalty term in the DIC criterion was frequently negative in the simulations. Although pD is typically positive for most Bayesian statistical models, pD can be negative for a particular model and data set if the likelihood function is not log-concave. A negative pD rewards, rather than penalizes, a model for model complexity. When pD is negative, then for simple data sets (M_0, for example), DIC selects a more complex model in the majority of cases. Of particular concern was the disproportionately large number of selections of model M_{bh} across all data sets, due to the frequency of the penalty term pD being negative. Detailed data tables regarding the performance of DIC across the eight simulation experiments are available at http://www.mathsci.appstate.edu/~rmg/.

Spiegelhalter, et al. (2002) stated that alternative choices for $\hat{\theta}$ could be the posterior median or posterior mode. So, pD can be calculated with these alternatives to the posterior mean of $\hat{\theta}$. Because DIC performed poorly with the posterior mean as $\hat{\theta}$, the performance of DIC

was examined when the posterior median was used for $\hat{\theta}$ instead. Ultimately, with this change, the problem of negative pD values improved, but still persisted with this change. Overall, performance of DIC in model selection for these models is inferior to that of AIC; based upon this simulation study, use of AIC as a model selection tool is recommended over DIC.

Analysis of Real Data Sets: Cottontail Rabbit Data

In Edwards and Eberhardt (1967), a capture-recapture experiment involving 135 cottontail rabbits was performed. The rabbits were released into a forty acre rabbit-proof area, and eighteen capture periods followed after a four day waiting period which gave the rabbits familiarity with their surroundings. Bayesian Models with the Program MARK models are compared. The data and a Program MARK analysis of the data are included with the MARK software package, and Pledger (2000), among others, has analyzed this data set. A total of seventy-six animals were captured at least once during the eighteen capture periods. Forty-three animals were captured once during the study, sixteen were captured twice, eight were captured three times, six were captured four times, two were captured six times, and one rabbit was captured seven times.

Using the models provided by Program MARK, Table 4 gives the estimate of N for each model, and the upper and lower limits of ninety-five percent confidence intervals for N, and the frequentist AIC statistic for each model. For Model M_{th}, the data was analyzed under two specifications, once with constant difference in capture probabilities between the two mixture groups across the time periods, and once without this restriction. Estimates and confidence limits for N are rounded to the nearest integer.

Using AIC, it is found that the M_{th} model with additive capture probability difference across the r = 2 groups across capture periods, and Model M_{tbh} have comparably small AIC values. The point estimator from the chosen M_{th} model is more accurate than the M_{tbh} model and the confidence interval for M_{th} is narrower.

Using WinBUGS v.1.4, the Bayesian

Table 4: Program MARK Results for Cottontail Dataset

Estimator	$\hat{N}$	LCL	UCL	AIC
M_0	96	87	114	379.6
M_h (2 mixture groups)	136	96	256	369.6
M_h (3 mixture groups)	157	89	593	373.5
M_b	94	82	129	381.6
M_t	95	86	112	354.6
M_{bh}	113	86	214	369.1
M_{th} (2 mixture groups; additive)	133	96	241	341.3
M_{th} (2 mixture groups; unrestricted)	98	88	117	367.0
M_{tb}	162	117	260	343.3
M_{tbh}	270	100	1698	341.9

models were fit to the cottontail data, using the non-informative prior distributions for N and the capture probabilities described in Section 4.1. For each model, Table 5 lists $\hat{N}$ (the posterior median), the AIC posterior mean, and ninety-five percent, equal-tailed posterior interval bounds from the MCMC posterior distribution of N. Figure 1 shows the MCMC posterior density of N for Model M_{tb}.

Table 5: Bayesian Model Results for Cottontail Dataset

Model	$\hat{N}$	2.5%	97.5%	AIC Posterior Mean
M_0	97	86	114	461.5
M_h	145	99	615	453.3
M_b	96	81	153	464.4
M_t	92	83	106	453.2
M_{tb}	104	88	138	452.8
M_{bh}	119	85	581	457.2
M_{th}	98	86	120	475.8
M_{tbh}	107	88	151	477.2

Note that AIC chooses Model M_{tb}. Other candidate models with comparable AIC values are M_h and M_t.

Figure 1: Posterior Density of N for Model Mtb for Cottontail Dataset
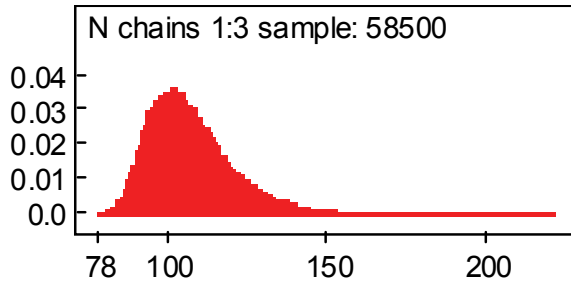
Model M_{tb} underestimates the true N = 135 but a 95% equal-tail interval from the posterior distribution of N contains the true N. The $\hat{N}$ from the Bayesian Model M_t underestimates the true N = 135, as does Model M_t in Program MARK. The interval estimates produced by the two methods are similar, which is not surprising given that relatively uninformative prior distributions for the parameters were used, and the likelihood functions of the two models are the same.

The Bayesian M_h estimate is somewhat above, but relatively close to the true N = 135, as are both M_h estimates from Program MARK. The Bayesian posterior density of N has a higher 97.5th percentile than the upper bound for the confidence interval given for Model M_h in Program MARK. The Model M_h posterior density for N is heavily right-skewed, and the posterior interval length could be significantly shortened by choosing an interval other than an equal-tailed interval, or by lowering the confidence level.

Ultimately, the M_{tb} point estimate of N, via the posterior median of N, is comparably accurate with the M_{tb} estimator from Program MARK. Also note the Program MARK M_{th} estimator is quite accurate for N. The Bayesian M_{tb} model has a narrower confidence interval than that from Program MARK, and the Bayesian interval contains the true N = 135.

Because this is only one data set, general conclusions cannot be made.

Analysis of Real Data Sets: Mead's Milkweed Flower Data

Alexander, Slade, and Kettle (1997) used mark-recapture methods to estimate the number of Mead's Milkweed plants on a 4.5-ha tract of land in Kansas. The capture periods consisted of an annual search of the land area over a span of four years. Observed plants were marked with a flag so that previous captures were detectable in subsequent years. Censuses were considered impossible because these plants are perennial and do not flower every year. Presence of flowering stems makes the plants easier to observe. The authors considered the population closed over the four-year span because the plant has a long lifespan, a high survival rate, and births and deaths were considered negligible during the study. Ultimately, a total of 129 flowers were observed in the study. Twenty-two plants were observed during one capture period, fifty-six were observed during two capture periods, twenty-five were observed during three capture periods, and twenty-six were observed during all four capture periods.

Model M_{tbh} was chosen as an ideal model for the data because time effects occur due to annual variation in flowering, behavioral effects occur because the visible flags make recapture easier in subsequent years, and heterogeneity effects occur because some plants have larger underground root systems which make them more likely to flower in a given year. Alexander, et al. used Program CAPTURE for the analysis (see http://www.mbr-pwrc.usgs.gov/software. html for details), Model M_{tbh} was unavailable in Program CAPTURE at that time, and they ultimately found some reasonable, but non-ideal, options for simplifying the data to allow the other seven closed population models to be fit.

Both Program MARK and WinBUGS v. 1.4 were used to analyze the data set and to choose the proper model from using the AIC statistic in each case. Results are listed below in Tables 6 and 7. Note that in Program MARK, M_{tb} and M_{tbh} models were fit with behavioral

effects additive across time periods and mixture groups.

Table 6: Program MARK Results for Mead's Milkweed Dataset

Estimator	$\hat{N}$	LCL	UCL	AIC
M_0	132	130	139	227.9
M_h (2 mixture groups)	135	132	144	218.7
M_b	222	162	393	115.8
M_t	129	130	135	80.2
M_{bh} (2 mixture groups)	945	230	6,769	109.7
M_{th} (2 mixture groups)	130	130	137	51.3
M_{tb}	167	137	326	48.1
M_{tbh}	1,228	233	11,769	38.4

Table 7: Bayesian Model Results for Mead's Milkweed Dataset

Model	$\hat{N}$	2.5%	97.5%	AIC Posterior Mean
M_0	133	130	138	263.4
M_h	136	131	145	259.3
M_b	270	172	2823	152.4
M_t	130	130	133	119
M_{tb}	632	336	2445	79.8
M_{bh}	365	182	2823	149
M_{th}	131	130	136	82.9
M_{tbh}	131	130	632	85

Comparing the results, it is observed that in Program MARK, the AIC statistic favors Model M_{tbh}. This model has a very large upper bound for the confidence interval and thus a wide confidence interval for N. Model M_{tb} in Program MARK has an AIC statistic that is fairly close to that of M_{tbh} and the confidence interval for N is much narrower than that of M_{tbh}.

The Bayesian AIC statistic favors Model M_{tb}, and Models M_{th} and M_{tbh} are other possible choices. The Bayesian M_{tb} model has a large interval width and a large 97.5[th] percentile of the posterior distribution of N. The competing Bayesian models M_{th} and M_{tbh} have equal point estimates of N, but M_{tbh} has a much larger 97.5[th] percentile of the posterior distribution of N, leading to a wider posterior 95% interval.

Conclusion

In summary, useful findings for closed population capture-recapture models have been established. Eight Bayesian capture-recapture models accounting for the known sources of variability in the capture probabilities of closed animal populations were developed. Using the WinBUGS v.1.4 software, these models were easy to fit to capture-recapture data sets, and MCMC estimates of the posterior density of N are easily obtained from the output. Additionally, the modified version of AIC works well as a model selection tool for capture-recapture data sets, thus AIC is useful as a preliminary method of reducing the set of candidate models from eight down to a smaller subset worthy of further exploration to determine the best fitting model. The DIC criterion did not perform as well as AIC for capture-recapture data sets and the use of AIC over DIC is recommended.

Further areas of exploration include examining whether informative priors improve estimation of N when capture probabilities are small. Negative bias in estimating N is common for populations with heterogeneity, particularly when a significant fraction of the population has small capture probabilities. The performance of the heterogeneity models (M_h, M_{th}, M_{bh}, M_{tbh}) when the finite mixture distribution F has r > 2 mass points should also be examined.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, B. N. Petrov & F. Csaki (*Eds.*), 267-281. Akademiai Kiado, Budapest.

Alexander, H. M., Slade, N. A., & Kettle, W. D. (1997). Application of mark-recapture models to estimation of the population size of plants. *Ecology*, *78*(*4*), 1230-1237.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference, A practical information-theoretic approach* (*2nd Ed.*). NY: Springer-Verlag.

Castledine, B. (1981). Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, *67*, 197-210.

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *1*(*4*), 681 – 684.

Chao, A. (1988). Estimating animal abundance with capture frequency data. *Journal of Wildlife Management*, *52*, 295-300.

Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, *45*, 427-438.

Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, *6*(*2*), 158-175.

Coull, B.A., & Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, *55*(*1*), 294-301.

Durban, J. W., & Elston, D. A. (2005). Mark-recapture with occasion and individual effects: abundance estimation through bayesian model selection in a fixed dimensional parameter space. *Journal of Agricultural, Biological, and Environmental Statistics*, *10*(*3*), 291-305.

Edwards, W. R., & Eberhardt, L. (1967). Estimating cottontail abundance from live trapping data. *Journal of Wildlife Management*, *31*(*1*), 87-96.

Gelfand, A. E., & Ghosh, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, *85*, 1-11.

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

Ghosh, S. K., & Norris, J. L. (2005). Bayesian Capture-Recapture Analysis and Model Selection Allowing for Heterogeneity and Behavioral Effects. *Journal of Agricultural Biological and Environmental Statistics*, *10*, 35-49.

Gosky, R. M., & Ghosh, S. K. (2011), A Comparative Study of Bayes Estimators of Closed Population Size from Capture-Recapture Data. *Journal of Statistical Theory and Practice,* to appear.

King, R., & Brooks, S. P. (2008). Bayesian Estimation of a Closed Population Size in the Presence of Heterogeneity and Model Uncertainty. *Biometrics*, *64*, 816-824.

Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, *59*, 1123-1130.

Norris, J. L., & Pollock, K. H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, *52*, 639-649.

Otis, D. L., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). Statistical inference from capture data on closed animal population. *Wildlife Monographs*, *62*, 135 pp.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models under mixtures. *Biometrics*, *56*, 434-442.

Schwartz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, *6*, 461-464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit, (with discussion and rejoinder). *Journal of the Royal Statistical Society, Series B*, *64*, 583-639.