Bayesian Genetic Mark-Recapture Methods For Estimating Seasonal River Run Size Of Stock Populations

Yiran Wang, Martin Lysy, Audrey Béliveau

Department of Statistics and Actuarial Science, Faculty of Mathematics, University of Waterloo
y3577wan@uwaterloo.ca, mlysy@uwaterloo.ca, audrey.beliveau@uwaterloo.ca

Introduction

Genetic mark-recapture (GMR) is a statistical technique used in estimating population size in ecology. By combining estimated relative abundance for all the species of interest from genetic data with counts for some species, GMR provides information about the total population size and the contributions of each species.

In this study, we propose a novel Bayesian GMR framework which presents several advantages over current frequentist techniques. First, the Bayesian framework can explicitly incorporate the sampling error from the genetic sample within the model specification to accommodate the fact that the observed relative proportions in the genetic sample differ from those in the population. Second, uncertainty in the estimates is easily obtained from the posterior samples, without the need for approximations or bootstrapping. Third, the Bayesian framework lends itself nicely to eventually incorporating additional sources of data into a single model.

The effectiveness of the new method is investigated via simulation studies and is shown to perform better than the frequentist estimator in a first simple simulation. However, in a second simulation, where some of the stocks have very small relative abundances, the method is very sensitive to the choice of prior. We discuss this matter in the Discussion Section.

Motivating application

We want to estimate total (N) and stock-specific (N_k) abundances of 17 stocks of Sockeye Salmon (of which 4 are lake-type and 13 are river-type) in the Taku River (BC, Canada.) Let stock $k = 1$ denote an aggregate of the lake-type stocks and stocks $k = 2, \ldots, K$ represent the river-type stocks, with $K = 14$ stocks.

Genetic Stock Identification (GSI) Data

Genetic tissues are collected weekly (over T weeks) on a sample of size n_t, $t = 1, ..., T$, at the Canadian commercial fishery. Samples are analyzed using a genetic MCMC algorithm [1]. It uses genetic profiles of baseline populations to produce in-sample posterior stock proportion estimates $(\mu_{k,t})$ along with posterior standard deviations $(\sigma_{k,t})$, in week t for stock k.

Weir Count Data

For each lake in the Taku River system, a counting weir is built at the entrance and monitored for the escapement of the Sockeye Salmon. Taku River lake-type Sockeye Salmon stocks are enumerated at the counting weirs (assuming perfect detection). The total aggregate count for lake-type stocks is denoted N_1. There are no such counts for river-type stocks.

Run Weight Data

CPUE at the Canyon Island fish wheels (close to the border between U.S. and Canada) is used to calculate the Sockeye Salmon run weight (relative abundance),w_t, in week t, with $\sum_{t=1}^{T} w_t = 1$.

Methods

Method of Moments (MM)

In 2001, W. J. Gazey developed a method of moments estimator to estimate total abundance of specific Salmon stocks in the Alsek River system[2]. This method has been used to estimate the abundance of Sockeye Salmon in the Taku River system by Pestal et al.[1]:

$$\hat{N} = \frac{N_1}{\sum_{t=1}^{T} w_t \mu_{1,t}}, \tag{1}$$

with

$$\widehat{\mathrm{Var}}(\hat{N}) = \sum_{t=1}^{T} (w_t \sigma_{1,t})^2 \left[\frac{\hat{N}}{\sum_{i=1}^{T} w_i \mu_{1,i}} \right]^2. \tag{2}$$

Note that equations (1) and (2) do not use data from river-type stocks. This method provides an asymptotically unbiased estimator. However, we note that the stock proportions in the GSI samples were assumed the same as in the population when deriving the variance estimator. Since the variance estimator does not account for the uncertainty in the sampling process, we believe the variance may be significantly underestimated.

Proposed Model

We propose a hierarchical model to model 1) the random selection of samples to undergo genetic identification and 2) the observed GSI data given the genetic samples selected.

1. Let $X_{k,t}$ denote the number of fish of stock k in the GSI sample in week t, then:

$$(X_{1,t}, X_{2,t}, ..., X_{K,t}) \sim \mathrm{Multinom}(n_t; \pi_{1,t}, \pi_{2,t}, ..., \pi_{K,t}), \tag{3}$$

where $\pi_{k,t}$ is the true proportion of stock k in the population in week t.

2. The observed GSI proportions $\mu_{k,t}$ are described as

$$\mu_{k,t} \sim TN\left(\frac{X_{k,t}}{n_t}, \sigma^2_{k,t}\right) \tag{4}$$

where TN represents a truncated normal distribution with a range of $[0, \infty)$. The choice of the normal distribution is motivated by the Bernstein-von Mises theorem.

We can then derive the total number of Sockeye Salmon as

$$N = \frac{N_1}{\sum_{t=1}^{T} w_t \pi_{1,t}} \tag{5}$$

and stock-specific abundances as $N_k = N \sum_{t=1}^{T} w_t \pi_{k,t}$, for $k = 2, ..., K$.

Choice of Prior (Proposed Model)

A prior needs to be specified on $(\pi_{1,t}, ..., \pi_{K,t})$.

Conjugate Dirichlet Prior (CDP)

The most intuitive prior is the Dirichlet prior, which is conjugate to the multinomial model:

$$(\pi_{1,t}, ..., \pi_{K,t}) \sim \mathrm{Dirichlet}(\theta_1, ..., \theta_K) \tag{6}$$

When $\theta_k = 1$, it is a uniform prior; when $\theta_k = \frac{1}{2}$, it is a Jeffreys prior[3]. As mentioned in the introduction, this prior did not perform well in our second simulation study, thus we considered alternate prior formulations given below.

Transformed Normal Prior (TNP)

The second type is a transformed normal prior proposed by Gelman[4]. Compared to the Dirichlet prior, this prior has more parameters which can provide more flexibility:

$$\pi_{k,t} = \frac{\exp(\tau_{k,t})}{\sum_{i=1}^{K} \exp(\tau_{i,t})} \tag{7}$$

$$\tau_{k,t} \overset{ind}{\sim} N(0, \xi^2). \tag{8}$$

We consider various values for ξ: 0.5, 1, 2, ..., 5 as well as a uniform hyperprior $\xi \sim U(0, 10)$.

Time Series Prior (TSP)

The transformed normal prior can be modified to incorporate autocorrelation between weeks. Our hypothesis was that this sharing of information across weeks could improve the precision of estimates. Starting from equation (7), we define an AR(1) process:

$$\tau_{k,t} = \rho \tau_{k,t-1} + \epsilon_{k,t} \tag{9}$$

$$\tau_{k,1} \sim N\left(0, \frac{s^2}{1-\rho^2}\right) \tag{10}$$

$$\epsilon_{k,t} \sim N(0, s^2) \tag{11}$$

$$\rho \sim \mathrm{Unif}(-1, 1) \tag{12}$$

For s, we consider the values 0.5, 1, 2, ..., 5 as well as a uniform hyperprior $s \sim U(0, 5)$. We also consider the case where s and ρ vary per stocks and weeks with independent priors.

Simulation Results

The simulations were implemented by JAGS, a software using gibbs sampling algorithm. Three chains were run until the second half of the chain converged based on Gelman-Rubin convergence diagnostic $\hat{R} < 1.05$ and provided an effective sample size of at least 1,000 after thinning to 10,000 iterations. Thinning was used to reduce storage space.

Simple Simulation

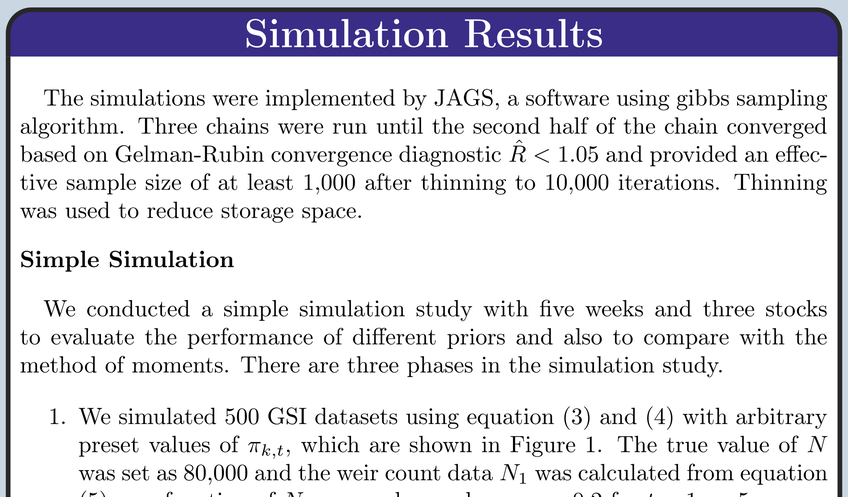
We conducted a simple simulation study with five weeks and three stocks to evaluate the performance of different priors and also to compare with the method of moments. There are three phases in the simulation study.

1. We simulated 500 GSI datasets using equation (3) and (4) with arbitrary preset values of $\pi_{k,t}$, which are shown in Figure 1. The true value of N was set as 80,000 and the weir count data N_1 was calculated from equation (5) as a function of N, $\pi_{k,t}$ and w_t, where $w_t = 0.2$ for $t = 1, ..., 5$.

2. We analysed each dataset with methods MM and the proposed method with all the various configurations of priors CDP, TNP and TSP. We calculated the estimate of N from each analysis.

3. We evaluated the performance of each method by calculating the Monte Carlo relative bias (RB), relative root mean square error (RRMSE), estimated standard error or posterior standard deviation (SD) and coverage probability of 95% confidence/credible intervals (CP).

Figure 1: Parameter values $(\pi_{k,t})$ of the simple simulation.



The results for all the priors and the MM method are close to the true value, with absolute RBs less than 0.01 and RRMSEs less than 0.07. However, the CP of the MM method is 34.2% while for our model they range from 94.6% to 96%, and the SD of $\hat{N}$ from our method is 4 times that of the MM method, which shows that the MM estimator underestimates the amount of uncertainty in the estimate. The failure in estimating the variance estimator was expected. As mentioned previously, the MM estimator does not account for the fact that $\mu_{k,t}$ is the stock proportion in the GSI samples, instead of in the population.

Simulation Results

Data-Based Simulation

We also performed another simulation study, which aimed to reproduce the reality at Taku river. We used 12 weeks and 17 stocks, of which four lake-type stocks are aggregated as stock 1. We used the same three phases as for the simple simulation study but with different parameter values. Total abundance $N = 60,000$ is based on the MM estimator in the PSC report[1]. The values of w_t and $\sigma_{k,t}$ were set to the observed values from the Taku River dataset. The values of w_t are (0.017, 0.014, 0.038, 0.027, 0.178, 0.179, 0.111, 0.109, 0.086, 0.156, 0.044, 0.042). Figure 2 shows the values of $\sigma_{k,t}$. The values of $\pi_{k,t}$ were set to the observed $\mu_{k,t}$ in the GSI data, shown in Figure 3.

Figure 2: Standard deviation values $(\sigma_{k,t})$ of the data-based simulation.

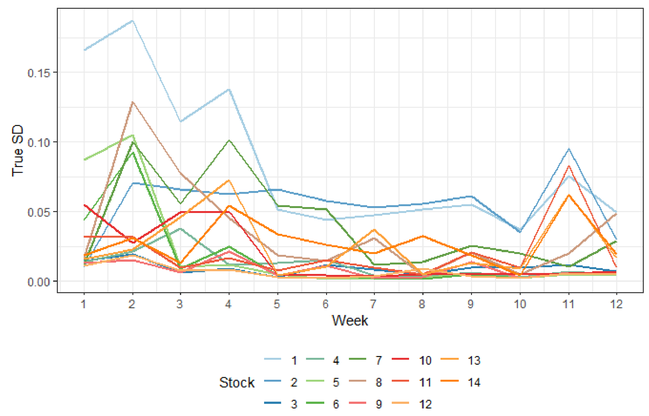
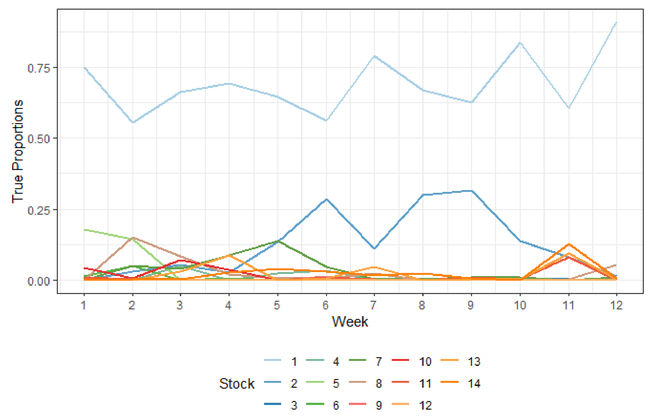


Figure 3: Parameter values $(\pi_{k,t})$ of the data-based simulation.



The results are shown in Table 1. Interestingly, in this specific simulation, the CP of the MM estimator is not as problematic as in the simple simulation. The proposed method with the TNP prior and $\xi = 2$ is the best performing model with the smallest RRMSE, and CP closest to the 0.95 target. However, all of the other estimators performed worse than the MM estimator.

Table 1: Results of the data-based simulation.

Setting	Relative Bias	RRMSE	SD	CP
MM	**0.00**	**0.032**	1,570	**0.88**
DCP $\theta_k = 1$	0.18	0.1868	2,334	0.00
DCP $\theta_k = 0.5$	0.09	0.092	2,042	0.18
TNP $\xi = 0.5$	0.25	0.253	2,621	0.00
TNP $\xi = 1$	0.06	0.064	1,983	0.58
TNP $\xi = 2$	**-0.01**	**0.030**	1,760	**0.93**
TNP $\xi = 3$	**-0.03**	**0.038**	1,694	**0.85**
TNP $\xi = 4$	**-0.03**	0.045	1,658	0.78
TNP $\xi = 5$	**-0.04**	0.049	1,634	0.69
TNP $\xi \sim \mathrm{U}(0, 10)$	**-0.05**	0.054	1,633	0.63
TSP $s \sim U(0, 5)$	**-0.04**	0.052	1,640	0.64

Discussion

In the first simple simulation study, our proposed Bayesian method was clearly superior to the MM estimator. However, in the data-based simulation study, it appears that the proposed method is particularly sensitive to the choice of prior. We believe that this is due to some of the $\pi_{k,t}$'s being very close to zero. One possible explanation could be that when simulating the data, we did not constrain $\sum_{k=1}^{K} \mu_{k,t} = 1$ for $t = 1, .., T$. We are currently investigating this matter. In addition, we are working on proposing priors that would not be so sensitive. In the meantime, the TNP model may still be reasonable to use if various values of ξ are considered and a model selection criteria such as the DIC or cross-validation is applied.

References

[1] Pestal, G., Schwarz, C.J. and Clark, R.A.(2020): *Taku River Sockeye Salmon Stock Assessment Review and Updated 1984-2018 Abundance Estimates*, Pacific Salmon Commission Technical Report No. 43.

[2] Gazey, W. J.(2010): *GSI Sample Size Requirements for In-river Run Reconstruction of Alsek Chinook and Sockeye Stocks*, Pacific Salmon Commission, Vancouver, British Columbia.

[3] Tuyl, F.(2017): *A Note on Priors for the Multinomial Model*, The American Statistician, 71(4), 298-301.

[4] Gelman, A. (1995): *Method of moments using Monte Carlo simulation*, Journal of Computational and Graphical Statistics, 4(1), 36-54.

Acknowledgements

We thank Carl J. Schwarz and the CANSSI team on Addressing Spatial and Computational Issues in Integrated Analysis of Modern Ecological Data for making this project possible.