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
University of Waterloo

August 6, 2023

Motivating Application

Figure 1: Sockeye Salmon (Photo by Kevin Phillips)

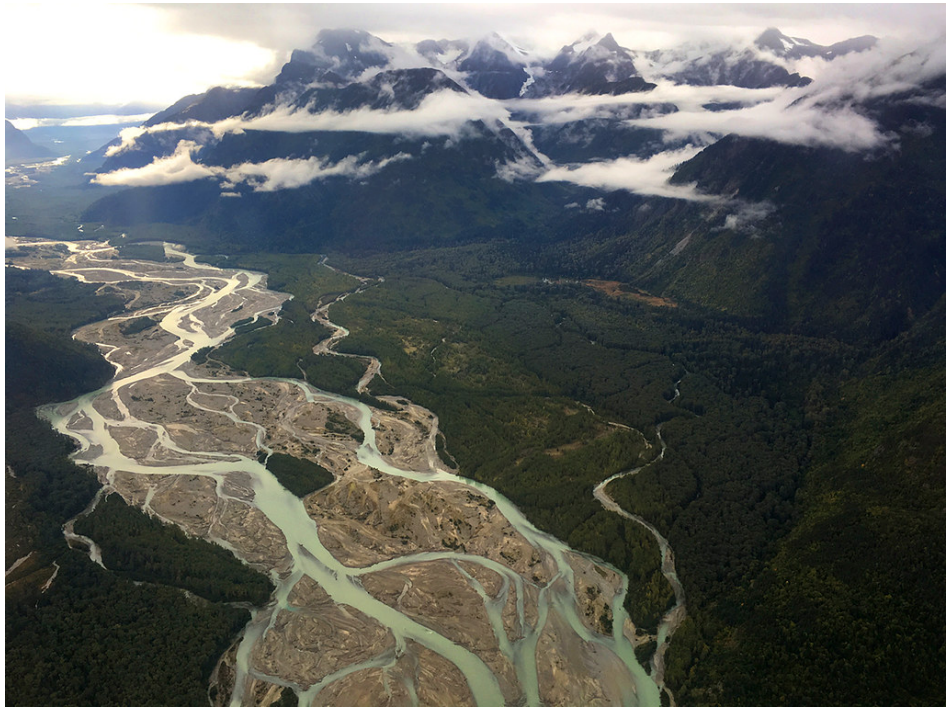


Figure 2: Taku River (Photo by Paul Vecsei)

Genetic Mark-Recapture (GMR) Method

Genetic Data:

Genetic Mark-Recapture (GMR) Method

Genetic Data:
Estimated Proportions for species of interest

Genetic Mark-Recapture (GMR) Method

Genetic Data:

Estimated Proportions for species of interest

Count Data:

Genetic Mark-Recapture (GMR) Method

Genetic Data:
Estimated Proportions for species of interest

Count Data:
Counts for some species

Genetic Mark-Recapture (GMR) Method

Genetic Data:
Estimated Proportions for species of interest

$$(\mu_1, ..., \mu_K)$$
$$(\sigma_1, ..., \sigma_K)$$

Count Data:
Counts for some species

Genetic Mark-Recapture (GMR) Method

Genetic Data:
Estimated Proportions for species of interest

$$(\mu_1, ..., \mu_K)$$
$$(\sigma_1, ..., \sigma_K)$$

Count Data:
Counts for some species

$$N_1$$

Genetic Mark-Recapture (GMR) Method

Genetic Data:
Estimated Proportions for species of interest

$$(\mu_1, ..., \mu_K)$$
$$(\sigma_1, ..., \sigma_K)$$

Count Data:
Counts for some species

$$N_1$$

$$\hat{N} = \frac{N_1}{\mu_1}$$

Available Datasets

Genetic Stock Identification (GSI) Data

n_t: Sample size of genetic samples

$\mu_{t,k}$: In-sample posterior stock proportion estimate

$\sigma_{t,k}$: In-sample posterior SD

* Stock 1 to L are lake-type stocks and $(L+1)$ to K are river-type stocks

Available Datasets

Genetic Stock Identification (GSI) Data

n_t: Sample size of genetic samples

$\mu_{t,k}$: In-sample posterior stock proportion estimate

$\sigma_{t,k}$: In-sample posterior SD

* Stock 1 to L are lake-type stocks and $(L+1)$ to K are river-type stocks

Weir Count Data

E_w: Total aggregate count for lake-type stocks

Available Datasets

Genetic Stock Identification (GSI) Data

n_t: Sample size of genetic samples

$\mu_{t,k}$: In-sample posterior stock proportion estimate

$\sigma_{t,k}$: In-sample posterior SD

* Stock 1 to L are lake-type stocks and $(L+1)$ to K are river-type stocks

Weir Count Data

E_w: Total aggregate count for lake-type stocks

Run Weight Data

w_t: The Sockeye Salmon run weight (relative abundance) in week t, with $\sum_{t=1}^{T} w_t = 1$

Study Setting

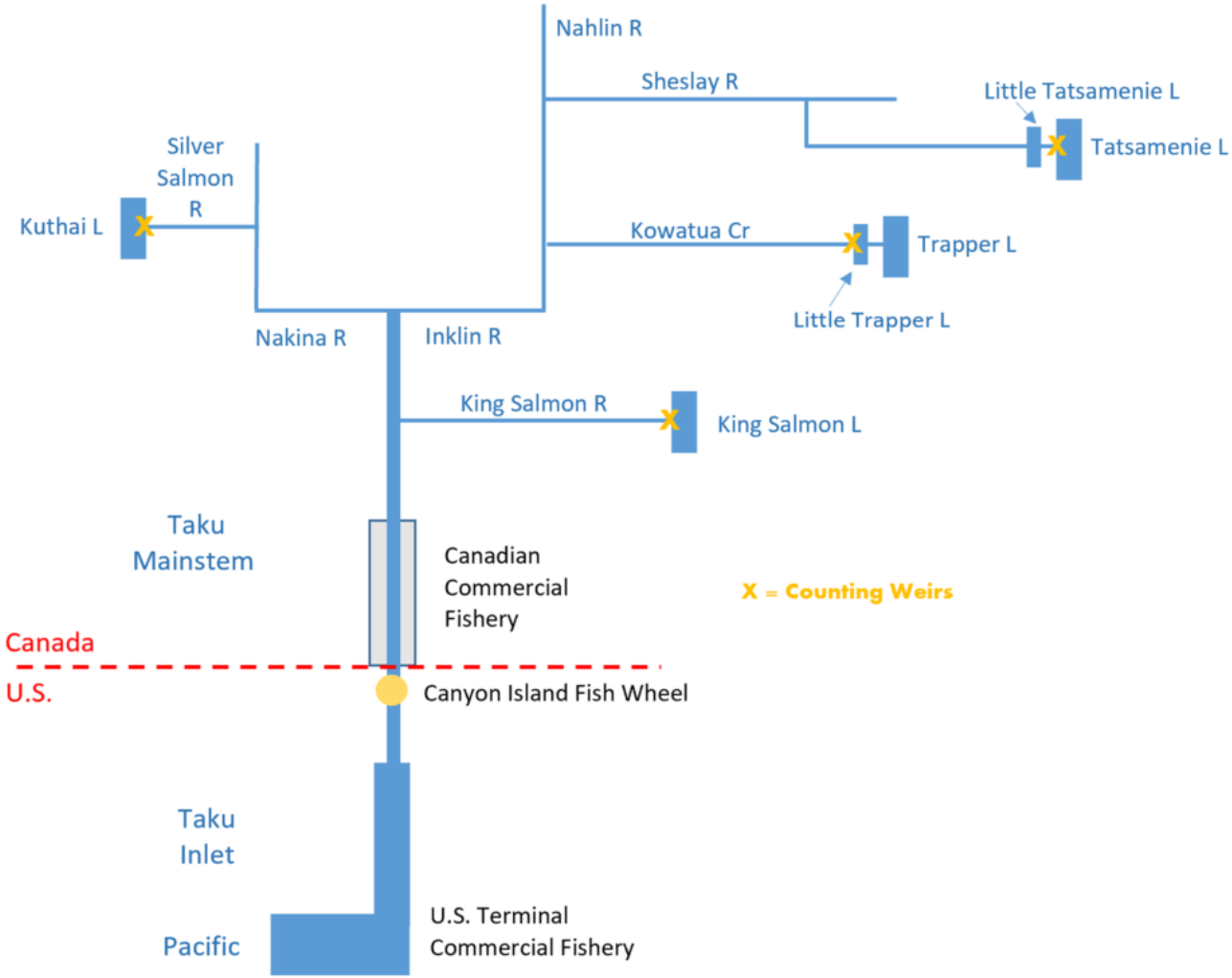


Figure 3: Overview of Taku River Sockeye Salmon stock structure and stock assessment [1]. R denotes river, Cr denotes creek and L denotes lake.

Method of Moments (MoM)

Gazey (2010) developed a method of moments estimator to estimate the total abundance of Chinook Salmon[2]:

$$\hat{N} = \frac{E_w}{\sum_{t=1}^{T} w_t \mu_{t,\text{Lake}}},$$

with

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

Method of Moments (MoM)

Gazey (2010) developed a method of moments estimator to estimate the total abundance of Chinook Salmon[2]:

$$\hat{N} = \frac{E_w}{\sum_{t=1}^{T} w_t \mu_{t,\text{Lake}}},$$

with

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

- **The variance may be significantly underestimated.**

Data-generating Process

Sampling from the population

$$(X_{t,1}, X_{t,2}, ..., X_{t,K}) \sim \text{Multinom}(n_t; p_{t,1}, p_{t,2}, ..., p_{t,K})$$

Modeling the samples

Assume $E(\mu_{t,k}) = \hat{p}_{t,k}$ and $Var(\mu_{t,k}) = \sigma^2_{t,k}$, where

$$(\hat{p}_{t,1}, ..., \hat{p}_{t,K}) = \left(\frac{X_{t,1}}{n_t}, ..., \frac{X_{t,K}}{n_t} \right)$$

- $\sum_{k=1}^{K} \mu_{t,k} = 1$ for $t = 1, ..., T$
- $\mu_{t,k} \in [0, 1]$ for $t = 1, ..., T$ and $k = 1, ..., K$.

Approximate Dirichlet Model

Approximate Dirichlet Model (ADM)

$$\boldsymbol{\mu}_t \mid \boldsymbol{X}_t \sim \text{Dirichlet}(\lambda_t \hat{\boldsymbol{p}}_t)$$
$$\boldsymbol{X}_t \sim \text{Multinom}(n_t, \boldsymbol{p}_t)$$

where

$$\lambda_t = \arg\min_c \sum_{k=1}^{K} \left[\frac{\mu_{t,k}^{\text{data}}(1 - \mu_{t,k}^{\text{data}})}{c+1} - {\sigma_{t,k}^{\text{data}}}^2 \right]^2$$

Choices of Prior

Dirichlet Prior

$$(p_{t,1}, ..., p_{t,K}) \overset{iid}{\sim} \text{Dirichlet}(1, ..., 1)$$

Time Series Prior

$$\begin{aligned}
p_{t,k} &= \frac{\exp(Z_{t,k})}{\sum_{k=1}^{K} \exp(Z_{t,k})} \\
Z_{1,k} &\overset{iid}{\sim} N(0, \psi^2) \\
Z_{t,k} &= \phi Z_{t-1,k} + \epsilon_{t,k} \text{ for } t = 2, ..., T \\
\epsilon_{t,k} &\overset{iid}{\sim} N(0, (1-\phi^2)\psi^2) \\
\phi &\sim \text{Unif}(-1, 1)
\end{aligned}$$

where we choose $\psi = 2$.

Proposed Estimator

Estimated Total Number of Sockeye Salmon

The mean of the posterior distribution of

$$N = \frac{E_w}{\sum_{t=1}^{T} w_t \sum_{k=1}^{L} p_{t,k}}$$

Estimated Stock-Specific Abundance

The mean of the posterior distribution of

$$N_k = N \sum_{t=1}^{T} w_t p_{t,k}$$

for $k = 1, ..., K$.

Moment-Matching Method

Including the latent variable $\boldsymbol{X}_t$ could slow down the computation process. So we try to remove $\boldsymbol{X}_t$ using moment matching method.

Suppose that $\boldsymbol{X}_t \sim \text{Multinom}(n_t, \boldsymbol{p}_t)$ and $\boldsymbol{\mu}_t \mid \boldsymbol{X}_t \sim \text{Dirichlet}(\lambda_t \boldsymbol{X}_t / n_t)$. Then we have

$$\text{E}[\boldsymbol{\mu}_t] = \text{E}[\text{E}[\boldsymbol{\mu}_t \mid \boldsymbol{X}_t]] = \text{E}[\boldsymbol{X}_t / n_t] = \boldsymbol{p}_t$$

and

$$\text{Var}(\boldsymbol{\mu}_t) = \left[\frac{1}{n_t} + \frac{1 - 1/n_t}{\lambda_t + 1}\right] \left\{\text{diag}(\boldsymbol{p}_t) - \boldsymbol{p}_t \boldsymbol{p}_t'\right\}$$

These are exactly the mean and variance of $\text{Dirichlet}(\tilde{\lambda}_t \boldsymbol{p}_t)$, where

$$1/(\tilde{\lambda}_t + 1) = \left[\frac{1}{n_t} + \frac{1 - 1/n_t}{\lambda_t + 1}\right].$$

Moment-Matching Method

Moment-Matching Dirichlet Model (MDM)

$$\boldsymbol{\mu_t} \sim \text{Dirichlet}(\tilde{\lambda}_t \boldsymbol{p_t})$$

where

$$\begin{aligned}\tilde{\lambda}_t &= 1/\left[\frac{1}{n_t} + \frac{1 - 1/n_t}{\lambda_t + 1}\right] - 1 \\ &= \frac{(n_t - 1)\lambda_t}{n_t + \lambda_t}\end{aligned}$$

Modified Variance Estimator for MoM Estimator

MoM Estimator

$$\hat{N} = \frac{E_w}{\sum_{t=1}^{T} w_t \mu_{t,\text{Lake}}}$$

Previous Variance Estimator

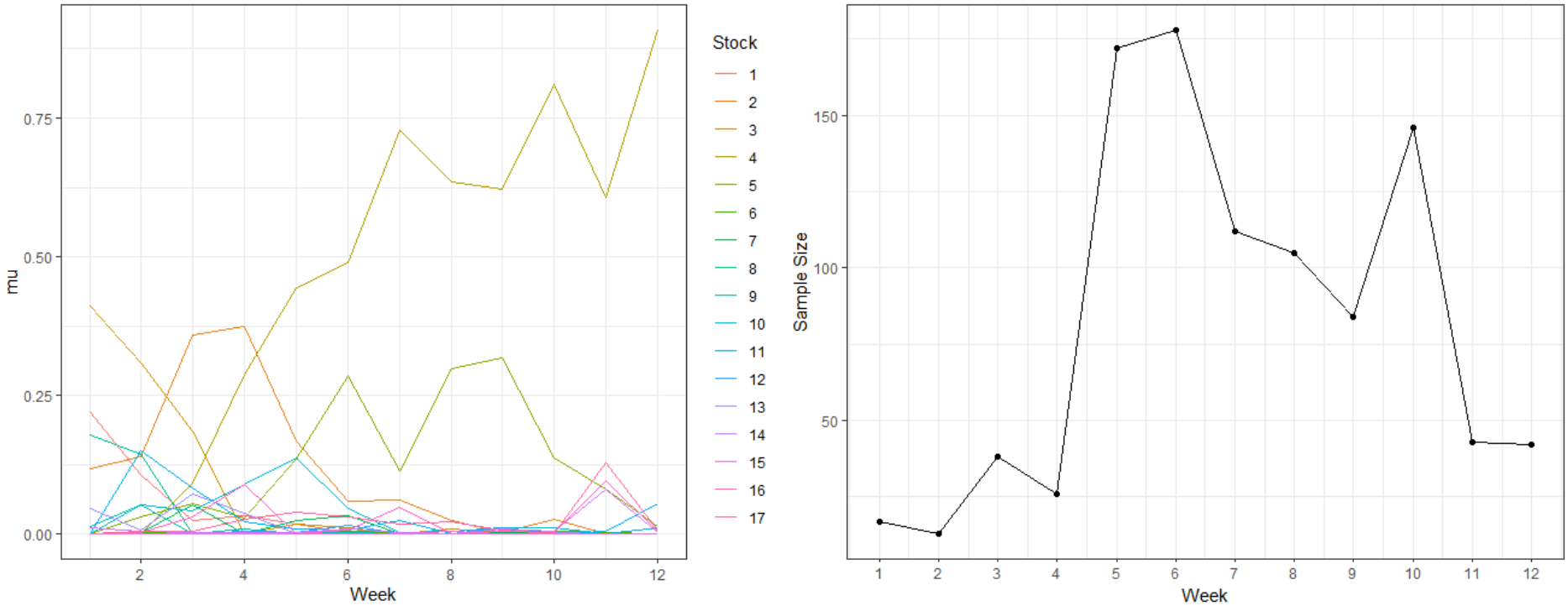
$$\widehat{\text{Var}}(\hat{N}) = \left[\frac{\hat{N}}{\sum_{t=1}^{T} w_t \mu_{t,\text{Lake}}}\right]^2 \sum_{t=1}^{T} (w_t^2 \sigma_{t,\text{Lake}}^2)$$

Modified Variance Estimator

$$\widehat{\text{Var}}(\hat{N}) = \left[\frac{\hat{N}}{\sum_{t=1}^{T} w_t \mu_{t,\text{Lake}}}\right]^2 \sum_{t=1}^{T} \left[w_t^2 \frac{\mu_{t,\text{Lake}}(1-\mu_{t,\text{Lake}})}{\tilde{\lambda}_t + 1}\right]$$

Pooling Technique

- When we have a large number of stocks where some stocks have low proportions with a small sample size and a uninformative prior, the performance of the method could be poor and it may take a long time for chains to converge.



(a) Weekly stock proportions

(b) Weekly sample size

Figure 4: Weekly stock proportions and sample size in GSI data.

Pooling Technique

- Pooling technique could be applied to address this problem.

$$\mu_{t,\text{Lake}} = \sum_{k=1}^{L} \mu_{t,k}$$

$$\mu_{t,\text{River}} = \sum_{k=L+1}^{K} \mu_{t,k}$$

and

$$\sigma^2_{t,\text{Lake}} = \sum_{k=1}^{L} \sigma^2_{t,k}$$

$$\sigma^2_{t,\text{River}} = \sum_{k=L+1}^{K} \sigma^2_{t,k}$$

Data-Based Simulation

- Values:
  - 12 weeks, 4 regions (2 lake-types vs. 2 river-types).
  - Using observed $\mu_{t,k}$ in the GSI data as the values of $p_{t,k}$.
  - The true value of N was set as 60,000 based on the MoM estimator in the PSC report[1].
  - The values of n_t, w_t and $\sigma_{t,k}$ were set as the observed values from the Taku River dataset.
- MCMC Process:
  - 2,000 datasets are generated and analyzed using R (Version 4.1.2;[3]) and JAGS (Version 4.3.0;[4]), with a reproducible seed.
  - Three chains are run until the second half of the chain converged based on Gelman-Rubin convergence diagnostic $\hat{R} < 1.05$ and provide an effective sample size of at least 1,000 after thinning to 4,000 iterations. Thinning is used to reduce storage space.

Simulation Study Results

Table 1: Results of the simulation studies.

Simulation Model	Inference Model	Prior	RBias	RRMSE	CP
ADM	MDM	Dir	0.0164	0.0335	0.934
ADM	MDM	AR(1)	0.0067	0.0314	0.955
ADM	ADM	Dir	0.0266	0.0394	0.865
ADM	ADM	AR(1)	0.0108	0.0322	0.952
ADM	MoM	N/A	0.0010	0.0305	0.910
ADM	MoM(M)	N/A	0.0010	0.0305	0.953

Real-Data Application Results

Table 2: Posterior inference for the Taku River Sockeye Salmon application.

Model	Prior	Estimate	SD	MCSE	95% CI
MDM	Dirichlet	61,303	1,836	32	(57,924, 65,105)
MDM	AR(1)	60,596	1,858	11	(57,148, 64,424)
ADM	Dirichlet	61,581	1,825	46	(58,223, 65,375)
ADM	AR(1)	60,768	1,861	48	(57,320, 64,601)
MoM	N/A	60,000	1,528	N/A	(57,006, 62,994)
MoM(M)	N/A	60,000	1,801	N/A	(56,469, 63,531)

Reference I

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

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

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: https://www.R-project.org/.

[4] Martyn Plummer et al. "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling". In: *Proceedings of the 3rd international workshop on distributed statistical computing*. Vol. 124. 125.10. Vienna, Austria. 2003, pp. 1–10.

Thank You