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

Introduction

Genetic mark-recapture (GMR) is a statistical technique used in estimating population size in ecology. By combining genetic data on the relative abundance of species from a sample with population counts obtained for a subset of the species, GMR allows the estimation of the total population size and the contributions of each species. However, the current method can suffer from a significantly underestimated variance, especially when the relative proportions in the genetic sample differ from those in the population.

In this work, we propose a novel Bayesian GMR framework to address this issue. The Bayesian framework can explicitly incorporate the sampling error in the genetic sample and readily lends itself to combining additional sources of data into a single model, such as capture-recapture data or telemetry data, which are also frequently used to estimate population size. The effectiveness of the new method is investigated via simulation studies and used to estimate the abundance of Sockeye Salmon in the Taku River.

Methods

- **Data Description**

Suppose we have K stocks where Stock 1 to L are lake-type stocks and (L+1) to K are river-type stocks. We have three types of datasets in this study.

- 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

- Weir Count Data

  E_w : Total aggregate count for lake-type stocks

- Run Weight Data

  w_t : The Sockeye Salmon run weight (relative abundance)

- **Current Method (MoM)**

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

with

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

The variance may be significantly underestimated!

- **Simulation Model**

$$\mu_t | \boldsymbol{X}_t \sim \text{Dirichlet}(\lambda_t \frac{\boldsymbol{X}_t}{n_t})$$

$$\boldsymbol{X}_t \sim \text{Multinom}(n_t, \boldsymbol{p}_t)$$

where

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

- **Inference Model**

$$\boldsymbol{\mu}_t \sim \text{Dirichlet}(\tilde{\lambda}_t \boldsymbol{p}_t)$$

where

$$\tilde{\lambda}_t = \frac{(n_t - 1)\lambda_t}{n_t + \lambda_t}$$

- **Choices of Prior**

▪ Dirichlet Prior

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

▪ Time Series Prior

$$p_{t,k} = \frac{\exp(Z_{t,k})}{\sum_{k=1}^{K} \exp(Z_{t,k})}$$

$$Z_{1,k} \overset{iid}{\sim} N(0, \psi^2)$$

$$Z_{t,k} = \phi Z_{t-1,k} + \epsilon_{t,k} \text{ for } t = 2, \dots, T$$

$$\epsilon_{t,k} \overset{iid}{\sim} N(0, (1-\phi^2)\psi^2)$$

$$\phi \sim \text{Unif}(-1,1)$$

We choose $\psi = 2$.

Photo by Kevin Phillips

Simulation Study

- 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 Pacific Salmon Commission report.
- The values of n_t, w_t, and $\sigma_{t,k}$ were set as the observed values from the Taku River dataset.
- 2,000 datasets are generated and analyzed using R and JAGS, with a reproducible seed.

Results

Table 1: Results of the simulation studies

Model	Prior	RBias	RRMSE	CP
New	Dir	0.0162	0.0335	0.933
New	AR(1)	0.0064	0.0314	0.95
MoM		0.0009	0.0306	0.898

Table 2: Results of Taku River application

Model	Prior	$\widehat{N}$	SD	95% CI
New	Dir	61,303	1,836	(57,924, 65,105)
New	AR(1)	60,596	1,858	(57,148, 64,424)
MoM		60,000	1,528	(57,006, 62,994)

Discussions

- The MoM method provides an unbiased estimate, but it fails to explain all the uncertainties and presents an underestimated variance of the estimate.
- We proposed a Bayesian method in this study, which incorporates the sampling error in the genetic sample.
- There are some limitations in our method, especially when we have small stock proportions and small sample size. However, we can use pooling technique to improve the performance.

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.

UNIVERSITY OF WATERLOO

Yiran Wang, Martin Lysy, Audrey Béliveau
y3577wan@uwaterloo.ca