

Modeling and Bayesian Computations for Capture-Recapture Studies

by

Yiran Wang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2024

© Yiran Wang 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Dr. Saman Muthukumarana
Professor, Department of Statistics
University of Manitoba

Supervisors: Dr. Audrey Béliveau
Associate Professor, Department of Statistics and Actuarial Science
University of Waterloo

Dr. Martin Lysy
Associate Professor, Department of Statistics and Actuarial Science
University of Waterloo

Internal Members: Dr. Joel Dubin
Professor, Department of Statistics and Actuarial Science
University of Waterloo

Dr. Lucy Gao
Adjunct Professor, Department of Statistics and Actuarial Science
University of Waterloo

Internal-External Member: Dr. Zahid A. Butt
Assistant Professor, School of Public Health Sciences
University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Yiran Wang was the sole author for Chapter 1 which was written under the supervision of Drs. Audrey Béliveau and Martin Lysy and was not written for publication.

A version of Chapter 2 of this thesis has been submitted to a peer-reviewed journal as a research paper for publication. Chapters 3 and 4 have been prepared for submission to a peer-reviewed journal. All three papers are co-authored with my supervisors. As lead author of these three chapters, I was responsible for deriving main theorems, implementing the proposed methods through R programming, performing the simulations and real data analysis, and writing the initial drafts. My supervisors provided guidance during each step of the research and provided feedback and edits on draft manuscripts.

Abstract

Capture-recapture methods are often used for population size estimation, which plays a fundamental role in informing management decisions in ecology and epidemiology. In this thesis, we develop novel approaches to population size estimation that more comprehensively incorporate various sources of statistical uncertainty in the data which are often overlooked. By addressing these uncertainties, our methods provide more accurate and reliable estimates of the parameters of interest. Furthermore, we introduce various techniques to enhance computational efficiency, particularly in the context of Markov Chain Monte Carlo (MCMC) algorithms used for Bayesian inference.

In Chapter 2, we delve into the plant-capture method, which is a special case of classical capture-recapture techniques. In this method, decoys referred to as “plants” are introduced into the population to estimate the capture probability. The method has shown considerable success in estimating population sizes from limited samples in many epidemiological, ecological, and demographic studies. However, previous plant-recapture studies have not systematically accounted for uncertainty in the capture status of each individual plant. To address this issue, we propose a novel modeling framework to formally incorporate uncertainty into the plant-capture model arising from (i) the capture status of plants and (ii) the heterogeneity between multiple survey sites. We present two inference methods and compare their performance through simulation studies. We then apply these methods to estimate the homeless population size in five U.S. cities using the large-scale “S-night” study conducted by the U.S. Census Bureau.

In Chapter 3, we look into the uncertainty in compositional data. Understanding population composition is essential in many ecological, evolutionary, conservation, and management contexts. Modern methods like genetic stock identification (GSI) allow for estimating the proportions of individuals from different subpopulations using genetic data. These estimates are ideally obtained through mixture analysis, which can provide standard errors that reflect the uncertainty in population composition accurately. However, traditional methods that rely on historical data often only account for sample-level uncertainty, making them inadequate for estimating population-level uncertainties. To address this issue, we develop a reverse Dirichlet-multinomial model and multiple variance estimators to effectively propagate uncertainties from the sample-level composition to the population level. We extend this approach to genetic mark-recapture scenarios, validate it with simulation studies, and apply it to estimate the escapement of Sockeye Salmon (*Oncorhynchus nerka*) in the Taku River.

In Chapter 4, motivated by the long run times of some of the Bayesian computations in this thesis, we shift our focus to the development and evaluation of Bayesian credible

intervals. Markov chain Monte Carlo (MCMC) methods are crucial for sampling from posterior distributions in Bayesian analysis. However, slow convergence or mixing can hinder obtaining a large effective sample size due to limited computational resources. This issue is particularly significant when estimating credible interval quantiles, which require more MCMC iterations than posterior means, medians, or variances. Consequently, prematurely stopping MCMC chains can lead to inaccurate credible interval estimates. To mitigate this issue in cases where the posterior distribution is approximately normal, we make a case for the use of parametric quantile estimation for determining credible interval endpoints. This chapter investigates the asymptotic properties of the parametric quantile estimation and compares it with the empirical quantile method to illustrate performance as MCMC chains are prolonged. Furthermore, we apply these techniques to a real-world capture-recapture dataset on Leisler's bat to compare their performance in a practical scenario.

Overall, this thesis contributes to the field of population size estimation by developing innovative statistical methods that improve accuracy and computational efficiency. Our work addresses critical uncertainties and provides practical solutions for ecological and epidemiological applications, demonstrating the broad applicability and impact of advanced capture-recapture methodologies.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisors, Drs. Audrey Béliveau and Martin Lysy. Their unwavering academic and financial support throughout my PhD journey has been invaluable. They have provided me with numerous opportunities to attend conferences and build professional networks, and most importantly, they have taught me how to become an independent researcher. To borrow a phrase from one of my favorite novels, *The Three-Body Problem*, they have instilled in me the importance of “thinking long and hard first.”

I am also grateful to all the members of my proposal and defense committees, Dr. Joel Dubin, Dr. Lucy Gao, Dr. Samuel Wong, Dr. Saman Muthukumarana, and Dr. Zahid Butt for their time and valuable suggestions. A special thanks goes to Dr. Paul Marriott and Dr. Jock MacKay for their enlightening course on the history, development, and current state of statistics. The STAT 900 course they taught was one of the most rewarding and beneficial courses I took during my PhD. I also want to extend my heartfelt thanks for all the guidance, discussion, and support I received from faculty members in or outside Waterloo throughout my academic journey, including Dr. Yeying Zhu, Dr. Alexander Schied, Dr. Richard Cook, Dr. Weining Shen, Dr. Joslin Goh, and many others. I would also like to thank Ms. Mary Lou Dufton, Mr. Greg Preston, and Mr. Carlos Mendes for their great patience and help with my administrative issues.

I would like to extend my gratitude and best wishes to my friends who accompanied me during my PhD years: Lijia Wang, Jie Jian, Qiuqi Wang, Zhaoran Hou, Mingren Yin, Jingyue Huang, Zijin Liu, Kecheng Li, Xianwei Li, Dingding Hu, Meixi Chen, Shiyu He, Yuliang Shi, Jiayue Zhang, Cong Jiang, Qihuang Zhang, Chi-Kuang Yeh, Tianyi Pan, Yuling Chen, Trang Bui, Gracia Dong, Luke Hagar, Grace Topmpkins, Augustine Wigle, Mahsa Panahi, Alexandra Mossman, Minzee Kim, Mohan Wu, Yumeng Wang, Rui Wang, Mengxiao Wang, Ziqiao Lin, Ziqian Zhuang, Zhongyuan Zhang, Linke Li, Shiyao Ying, Ziang Zhang, Yuqing Liu, Yu Shi, and many others. The order of names is just random and does not reflect any particular ranking. Additionally, there are many other friends whose names are not included here, but please know that I am equally thankful to each and every one of you. Over the five years in the department, I witnessed many cohorts of graduates leave and welcomed new students each year. They have all left precious and joyful memories in my doctoral journey.

Finally, I wish to thank my parents Runshan Wang and Lanyun Xie for their constant concern and unwavering support. Above all, I want to express my deepest appreciation to my girlfriend, Mei Dong. We have shared an 11-year journey through our academic lives,

from undergraduate studies to our PhDs. She has always taken care of me meticulously, encouraging me every step of the way. I hope she will also successfully complete her PhD, and together, we can move on to the next chapter of our lives.

Dedication

*To my beloved parents and grandpa,
Runshan Wang, Lanyun Xie and Fuzhen Xie,
for their unconditional love and support.*

*To my dearest grandma,
Shulian Li.*

*To Mei Dong,
who has been by my side for 11 years,
tirelessly supporting and encouraging me every step of the way.*

Table of Contents

Examining Committee Membership	ii
Author’s Declaration	iii
Statement of Contributions	iv
Abstract	v
Acknowledgments	vii
Dedication	ix
List of Figures	xiii
List of Tables	xvi
1 Introduction	1
2 Plant-Capture Methods for Estimating Population Size from Uncertain Plant Captures	4
2.1 Introduction	4
2.2 Methodology	6
2.2.1 Basic Model for Uncertain Plant Captures (Model \mathcal{M}_{basic})	7
2.2.2 Incorporating Partial Identification Data (Model \mathcal{M}_{id})	9

2.2.3	Incorporating Heterogeneity Between Sites (Model \mathcal{M}_{class})	11
2.3	Inference Approaches	12
2.3.1	Frequentist Inference via Maximum Likelihood	12
2.3.2	Bayesian Inference via MCMC	14
2.4	Simulation Study	15
2.5	Application to the S-Night Street Enumeration Survey	19
2.6	Discussion	20
3	Rapid Scaling of Compositional Uncertainty from Sample to Population Levels	24
3.1	Introduction	24
3.2	Methodology	26
3.2.1	Approximate Reverse Dirichlet-Multinomial Model	26
3.2.2	Simple Analytical Formulas for Scaling Standard Errors	28
3.3	Methods for Genetic Mark-Recapture Studies	29
3.3.1	Bayesian Approaches	29
3.3.2	Frequentist Approaches	31
3.4	Simulation Study	32
3.4.1	Results	33
3.5	Application to the Taku River Salmon Run Estimation	36
3.6	Discussion	39
4	Parametric Quantile Estimation of Posterior Quantiles for Markov Chain Monte Carlo	41
4.1	Introduction	41
4.2	Methodology	43
4.2.1	Asymptotic Properties of the PQ Estimator	43
4.2.2	Comparative Evaluation under AR(1)	44
4.3	Simulation Study	47

4.3.1	Quantile Estimation under AR(1)	48
4.3.2	Quantile Estimation under Heavier Tails	48
4.3.3	Quantile Estimation for a Skewed Posterior	48
4.4	Real Data Example	51
4.5	Discussion	55
5	Conclusion	57
	References	59
	APPENDICES	68
A	Derivation of MLE for Model \mathcal{M}_{basic}	69
B	Alternative Computational Methods for Bayesian Inference	71
B.1	Bayesian Normal Approximation (BNA)	71
B.2	Uncertainty Propagation Method (UP)	72
C	Supplementary Tables for Chapter 2	74
D	Determining ψ in the AR(1) Prior Specification for $\pi_{k,t}$	77
E	Supplementary Figures for Chapter 3	79
F	CLT for the PQ Estimator	83
G	Estimating Leisler’s Bats Population Using Jolly-Seber and Data Augmentation	85
G.1	Notations	85
G.2	Model Specification	86
G.2.1	Prior	86
G.2.2	Likelihood	86
H	Supplementary Figures for Chapter 4	88

List of Figures

3.1	Distribution of the posterior mean for $\pi_{k,t}$ from the moment-matching Dirichlet model with AR(1) prior in the simulation study, for stocks $k = 1$ to 4 in weeks $t = 1$ to 12. Red horizontal lines indicate the true $\pi_{k,t}$ values.	35
3.2	Overview of Taku River system and study settings. R denotes river, Cr denotes creek and L denotes lake.	36
3.3	Two graphs displaying $\hat{p}_{k,t}$ and $s_{k,t}$, over weeks and regions, for the 2017 GSI dataset.	38
4.1	Monte Carlo standard error (MCSE) of PQ and EQ estimators under AR(1) with stationary distribution $N(0,1)$. “CLT” represents the MCSE computed from the theoretical CLT variance. “True” represents the true MCSE approximated via simulation. The shaded regions depict 95% HDI intervals for the estimated MCSE, obtained via simulation.	47
4.2	Expected estimates of the 0.975 quantile obtained from 100 AR(1) simulations with stationary distributions (rows) $N(0,1)$, t_{10} , or $Gamma(2,1)$ and autocorrelation parameter ρ (columns) of 0.1, 0.5, or 0.9. To reduce the influence of initial values, the first 100 iterations are omitted. The black dashed line indicates the true 0.975 quantile for each distribution, while the purple dashed line shows the 0.975 quantile of an approximate normal distribution with the same mean and variance as the corresponding distribution. The green line for $Gamma(2,1)$ represents the PQ estimates from Box-Cox transformed samples.	49

4.3	Root mean squared error (RMSE) of the 0.975 quantile obtained from 100 AR(1) simulations with stationary distributions (rows) $N(0,1)$, t_{10} , or $Gamma(2,1)$ and autocorrelation parameter ρ (columns) of 0.1, 0.5, or 0.9. To reduce the influence of initial values, the first 100 iterations are omitted. The RMSE of the PQ estimates for $Gamma(2,1)$ after Box-Cox transformation is included as the green line. Both x and y axes are \log_{10} transformed.	50
4.4	Estimates of the 0.025 and 0.975 quantile of $\bar{\phi}$. The left column shows the 0.025 quantile estimates, and the right column shows the 0.975 quantile estimates. The black vertical line indicates the 5,000th iteration, corresponding to the chain length used by Kéry and Schaub (2011) . The x-axis, on a \log_{10} scale, shows iterations from 2,500 to 100,000 (prior to thinning).	53
4.5	Estimates of the 0.025 and 0.975 quantile of N_s . The left column shows the 0.025 quantile estimates, and the right column shows the 0.975 quantile estimates. The black vertical line indicates the 5,000th iteration, corresponding to the chain length used by Kéry and Schaub (2011) . The x-axis, on a \log_{10} scale, shows iterations from 2,500 to 100,000 (prior to thinning).	54
D.1	Density plot for $\pi_{1,1}$ with different choices of ψ in the time series prior. . .	78
E.1	Distribution of the posterior mean for $\pi_{k,t}$ from the approximate reverse Dirichlet-multinomial model with Dirichlet prior in the simulation study, for stocks $k = 1$ to 4 in weeks $t = 1$ to 12. Red horizontal lines indicate the true $\pi_{k,t}$ values.	80
E.2	Distribution of the posterior mean for $\pi_{k,t}$ from the approximate reverse Dirichlet-multinomial model with AR(1) prior in the simulation study, for stocks $k = 1$ to 4 in weeks $t = 1$ to 12. Red horizontal lines indicate the true $\pi_{k,t}$ values.	81
E.3	Distribution of the posterior mean for $\pi_{k,t}$ from the moment-matching Dirichlet model with Dirichlet prior in the simulation study, for stocks $k = 1$ to 4 in weeks $t = 1$ to 12. Red horizontal lines indicate the true $\pi_{k,t}$ values. . .	82
H.1	Q-Q plots for N_s and $\bar{\phi}$ under different thinning and transformation scenarios. “High thinning” refers to running 100,000 iterations per chain and retaining a total of 2,500 samples. “No Thinning” refers to running 100,000 iterations per chain and retaining all the samples.	88

H.2	Squared error in the 0.025 and 0.975 quantile estimates for $\bar{\phi}$ compared to the EQ estimate with full set of samples. The red and dark red lines represent EQ estimates for the 0.025 and 0.975 quantiles, respectively. The blue and dark blue lines represent PQ estimates for the 0.025 and 0.975 quantiles, respectively. The green and dark green lines represent PQ estimates for the 0.025 and 0.975 quantiles after Box-Cox transformation. The x-axis, on a \log_{10} scale, shows iterations from 2,500 to 100,000. The black vertical line indicates the 5,000th iteration, corresponding to the chain length used by Kéry and Schaub (2011)	89
H.3	Squared error in the 0.025 and 0.975 quantile estimates for N_s compared to the EQ estimate with full set of samples. The red and dark red lines represent EQ estimates for the 0.025 and 0.975 quantiles, respectively. The blue and dark blue lines represent PQ estimates for the 0.025 and 0.975 quantiles, respectively. The green and dark green lines represent PQ estimates for the 0.025 and 0.975 quantiles after Box-Cox transformation. The x-axis, on a \log_{10} scale, shows iterations from 2,500 to 100,000. The black vertical line indicates the 5,000th iteration, corresponding to the chain length used by Kéry and Schaub (2011)	90

List of Tables

2.1	Results of the simulation studies with \mathcal{M}_{basic} for the MLEs and the Bayesian estimators. All the values are rounded to integers or 2 decimal points. . . .	16
2.2	Results of the simulation studies with \mathcal{M}_{id} for the MLEs and the Bayesian estimators. All the values are rounded to integers or 2 decimal points. . . .	17
2.3	Results of the simulation studies with \mathcal{M}_{class} for the MLEs and the Bayesian estimators. All the values are rounded to integers or 2 decimal points. . . .	18
2.4	The 1990 S-Night data reconstructed from the literature.	19
2.5	Results of the application to the S-Night data using \mathcal{M}_{id} without Equation (2.9) separately in each city. All the values are rounded to integers or 2 decimal points.	21
3.1	Comparison of estimation methods using results from the simulation study. “ARDM” denotes the approximate reverse Dirichlet-multinomial model, “MMD” represents the moment-matching Dirichlet model, “MoM” refers to the method-of-moments estimator with $(\hat{\sigma}_t^{lake})^2 = \tilde{\beta}_t \hat{p}_t^{lake} (1 - \hat{p}_t^{lake})$, “MoM(Alt)” refers to the method-of-moments estimator using $(\hat{\sigma}_t^{lake, alt})^2$, and “MoM(Naive)” refers to the method-of-moments estimator using $(s_t^{lake})^2$ instead of $(\hat{\sigma}_t^{lake})^2$.	34
3.2	2017 GSI data: weekly weights, sample sizes and variance inflation factor .	37
3.3	Results for the Taku River Sockeye Salmon application. “Estimate” for the Bayesian model refers to the posterior mean of N . “SD” denotes the posterior standard deviation for the Bayesian methods, and the estimated standard deviation for frequentist methods with different variance estimators. The 95% CI shows the 95% quantile-based credible intervals for the Bayesian approaches and $\hat{N} \pm 1.96 \cdot \sqrt{\widehat{\text{Var}}(\hat{N} \mathbf{w})}$ for the frequentist methods. “Time” represents the computing time for each setting.	39

4.1	Comparison of PQ and EQ CLT variances under AR(1) for different confidence levels and autocorrelation parameters. The ratio δ_{PQ}/δ_{EQ} indicates the relative efficiency of PQ compared to EQ.	46
C.1	Results of the simulation studies for \mathcal{M}_{basic} using BNA. All the values are rounded to integers or 2 decimal points.	74
C.2	Results of the simulation studies for \mathcal{M}_{id} using BNA and UP method. All the values are rounded to integers or 2 decimal points.	75
C.3	Results of the simulation studies for \mathcal{M}_{class} using BNA and UP method. All the values are rounded to integers or 2 decimal points.	76
C.4	Estimation of the homeless population size H using the Chapman-Bailey estimator. All the values are rounded to integers.	76

Chapter 1

Introduction

Population size assessment is a cornerstone of wildlife conservation, fisheries management, public health planning, and policy formulation (Williams et al., 2002; Barbraud et al., 2009; Coumans et al., 2017). Accurate estimates of population sizes are crucial for understanding population dynamics, assessing species' conservation status, managing natural resources sustainably, and planning effective public health interventions. Traditional methods, such as capture-recapture or mark-recapture, have been fundamental in this field. These methods involve capturing individuals from a population, marking them, releasing them, and then recapturing individuals to assess the proportion of marked individuals in subsequent samples (Seber, 1982; Nichols and MacKenzie, 2004). This proportion is used to estimate capture probabilities, which in turn allow researchers to extrapolate the total population size.

In recent decades, advancements in the field have led to the development of various novel capture-recapture methods. One innovative approach is plant-capture methods (Laska and Meisner, 1993), which only involve a single capture occasion on a population where marked individuals, referred to as “plants”, have already been introduced. A key metric derived from this approach is the proportion of planted individuals that were successfully captured. This proportion provides insight into the capture probability, a fundamental factor required for estimating the size of the target population from an incomplete count. Another advancement is genetic mark-recapture methods (Hamazaki and DeCovich, 2014), which utilize genetic information from sources as hair, fur, or other tissues (Taberlet et al., 1999; Waits and Paetkau, 2005) to identify individuals. Since individuals within a group, such as a specific subspecies or salmon from the same birthplace, often share the same genotype at multiple DNA loci, these shared genetic characteristics can function as molecular markers (Pella and Masuda, 2001), facilitating the differentiation of groups within a population.

By deriving group proportions in the genetic samples, we can estimate the total population size across all the genetics groups when the size of one or more genetic groups is known. These methods have expanded the applicability and accuracy of population size estimation across different ecological and epidemiological contexts.

Concurrent with these methodological advancements, the rapid development of computational power has revolutionized data analysis in capture-recapture studies. Bayesian methods, in particular, have become increasingly prevalent due to their ability to incorporate information from diverse sources and provide a full probabilistic framework for uncertainty quantification (Clark and Gelfand, 2006; McCarthy, 2007; Kéry, 2010; Kéry and Schaub, 2011; Schaub and Kéry, 2021). These methods are especially valuable in complex models where traditional frequentist approaches may fall short. However, Bayesian methods often require intensive computational resources, particularly when using Markov Chain Monte Carlo (MCMC) techniques for inference.

This thesis focuses on the development of novel modeling frameworks for various capture-recapture methods and introduces techniques to enhance computational efficiency, particularly in the context of Bayesian methods. By addressing the computational challenges associated with Bayesian inference, this work aims to make advanced capture-recapture methods more accessible and practical for researchers and practitioners.

In Chapter 2, we explore the plant-capture method, a specialized application of classical capture-recapture techniques used to estimate population sizes by introducing decoys, or “plants,” into a population. This approach has proven effective in estimating population sizes from limited samples in various fields, including epidemiology, ecology, and demography. However, previous research has often neglected to systematically account for uncertainties related to the capture status of each plant and the heterogeneity between different survey sites. In this chapter, we address these gaps by proposing novel approaches to formally incorporate these uncertainties into the plant-capture model. We develop and evaluate two inference methods through simulation studies to enhance the accuracy and reliability of population estimates. Furthermore, we demonstrate the practical application of these methods by estimating the homeless population size in several U.S. cities, using data from the large-scale “S-night” study conducted by the U.S. Census Bureau. This application underscores the potential of our refined plant-capture model to improve population size assessments in real-world scenarios.

In Chapter 3, we tackle the critical issue of accurately estimating population composition, which is essential for ecological, evolutionary, conservation, and management purposes. Modern techniques, such as genetic stock identification (GSI), have greatly enhanced our ability to estimate the proportion of individuals from different subpopulations based

on genetic data. Ideally, these estimates are derived through mixture analysis, which can provide standard errors that accurately reflect the uncertainty in population composition. However, historical data often rely on individual assignments that only include sample-level uncertainty for composition estimates, making traditional methods inadequate for correctly obtaining population-level uncertainties. To overcome this challenge, we introduce a novel method that propagates uncertainties from the sample level to the population level, using a reverse Dirichlet-multinomial model. This approach provides more robust and reliable estimates of population composition. We validate our method through extensive simulation studies and extend it to genetic mark-recapture contexts. The chapter culminates in the application of our method to estimate the escapement of Sockeye Salmon (*Oncorhynchus nerka*) in the Taku River, demonstrating its practical utility and effectiveness in managing and conserving mixed-stock fisheries.

In Chapter 4, we focus on the application of Markov Chain Monte Carlo (MCMC) methods in Bayesian analysis, which have gained popularity in the capture-recapture studies in recent decades. MCMC methods are widely used to sample from posterior distributions, allowing for posterior inference of parameters of interest. However, MCMC methods can suffer from slow convergence and mixing, limiting the effective sample size within given computational constraints. This issue is particularly pronounced when estimating credible interval quantiles, which typically require more MCMC iterations compared to other summary statistics like means or medians. To tackle this problem, we explore the use of parametric quantile estimation for determining credible interval endpoints, especially when the posterior distribution is approximately normal. We investigate the asymptotic properties of this method and compare it with standard empirical methods to evaluate its performance as MCMC chains are prolonged. By applying these techniques to a real-world capture-recapture dataset on Leisler's bat, we provide a practical comparison of their efficacy. This application highlights the potential of parametric quantile estimation to improve the accuracy of credible interval estimates in practical Bayesian analyses, especially under computational constraints.

Finally, it is important to note that Chapters 2, 3 and 4 are self-contained, each defining their own set of notations. Any notation in common between chapters is coincidental.

Chapter 2

Plant-Capture Methods for Estimating Population Size from Uncertain Plant Captures

2.1 Introduction

Plant-capture (Laska and Meisner, 1993) is a variation on Petersen capture-recapture experiments (Petersen, 1896) employed to estimate the size of a target population. Unlike traditional capture-recapture experiments which utilize two distinct sampling occasions, the plant-capture method operates with just one capture occasion. This capture occasion is carried out on a population where marked individuals, referred to as “plants”, have already been introduced. Much like other capture-recapture sampling schemes, this method operates under the assumption that planted individuals cannot be distinguished from the rest of the individuals in the target population during sampling. A critical piece of information derived from this approach is the proportion of planted individuals that were successfully captured. This proportion provides insight into the capture probability, a fundamental factor required for estimating the size of the target population from an incomplete count. Under the plant-capture setting, Laska and Meisner (1993) proposed a maximum likelihood estimator (equivalent to the Petersen estimator), a Chapman-Bailey estimator (Chapman, 1951; Bailey, 1952), and interval estimation for the target population size. Several papers have further discussed and extended this method (Martin et al., 1997; Goudie et al., 1998; Goudie and Ashbridge, 2000; Goudie et al., 2007; Ashbridge and Goudie, 2008).

Plant-capture methodology has been applied in different areas of survey research (Ashbridge and Goudie, 2008), such as ecology (Skalski and Robson, 1982; Yip and Fong, 1993), software reliability (Duran and Wiorkowski, 1981; Yip et al., 1999) and public health (Martin, 1992; McCandless et al., 2016). A common application of the plant-capture method is in estimating the size of homeless populations from point-in-time street surveys. As noted by Berry (2007), homeless surveys that omitted the inclusion of plants to adjust homeless counts have faced criticism for overlooking a substantial 40-70% of out-of-sight homeless individuals, which shows the importance of utilizing plants. Notably, the United States Department of Housing and Urban Development (2008) and the National Law Center on Homelessness & Poverty (2017) have endorsed the plant-capture method as an innovative approach for estimating the size of a homeless population. The method has been applied to several surveys (e.g., Hopper et al., 2008; OrgCode Consulting Inc., 2012; OrgCode Consulting, Inc., 2013; City of Red Deer, 2016; McCandless et al., 2016), including the large-scale Shelter and Street Night Enumeration (“S-night”) survey of the homeless conducted in five major US cities in 1990 by the United States Census Bureau (Hopper, 1991; Martin, 1992; Laska and Meisner, 1993; Martin et al., 1997).

Capture-recapture methods (including plant-recapture methods) typically operate under the assumption that once an individual is captured by an enumerator, the latter can tell with certainty if the individual is marked or not by the presence or absence of a mark. However, in street surveys of the homeless, survey enumerators may be counting individuals from a distance, thus unable to determine whether these individuals are plants or not since planted individuals do not have distinguishable marks. For convenience, we name this setting “capture without identification”. In such situations, determining which plants were captured relies on the plants’ self-assessment of whether they were counted or not. While some plants may be able to assess this with certainty, other plants’ assessments may be uncertain, as exemplified in the S-night survey, where plants had to answer “yes”, “maybe” or “no” to having been captured, via questionnaire.

To the best of our knowledge, the only study addressing uncertainty in plant assessment is Martin et al. (1997), which simply considers two extreme scenarios; one in which uncertain assessments (“maybes”) are all presumed captured and another in which they are all presumed not captured. While these two extremes can be used for sensitivity analysis, they cannot be readily synthesized to provide a single population estimate, nor its commonly accompanying measures of statistical uncertainty (e.g., standard errors, confidence intervals). That being said, the issue of uncertainty in assessment in the traditional capture-recapture setting has been extensively studied, particularly in animal ecology applications with physical markings at times replaced with natural marks such as DNA samples or photographs. Various methods have been proposed to address identification uncertainty in these con-

texts. For instance, [Wright et al. \(2009\)](#) proposed a method to estimate error rates for genetic misidentification, specifically targeting allelic dropout. [Lukacs and Burnham \(2005\)](#) introduced model $\mathbf{M}_{t,\alpha}$, an extension of the well-known model \mathbf{M}_t ([Otis et al., 1978](#)) incorporating parameter α to describe uncertainty in genotype identification. This approach has been subsequently refined in e.g., [Yoshizaki \(2007\)](#); [Link et al. \(2010\)](#); [Yoshizaki et al. \(2011\)](#); [Vale et al. \(2014\)](#); [Schofield and Bonner \(2015\)](#); [Bonner et al. \(2016\)](#). However, it is important to note that the genotype uncertainty paradigm above is limited to recognized (“yes”) and unrecognized (“maybe”) marks for which capture status is unknown, whereas the plant-capture identification assessment involves a third category: unrecognized and not captured (“no”). Moreover, estimation of parameter α in the genotype uncertainty paradigm requires more than one recapture occasion, which is infeasible in the settings for which plant-capture is required.

In this paper, we address uncertainty in plant captures by introducing a rigorous modeling framework that explicitly incorporates capture status uncertainty. A crucial assumption we introduce within this framework is a “Missing at random” (MAR) assumption ([Rubin, 1976](#)), which allows a formal account of the uncertainty inherent in the capture process. Our approach offers a flexible computational solution providing both frequentist and Bayesian population size estimators. Several simulation studies are conducted to validate the accuracy of our approach, establishing its ability to yield population estimates with the desired coverage probability.

Section 2.2 presents three plant-capture models tailored to different scenarios. In Section 2.3 we detail two approaches to inference, covering both frequentist and Bayesian paradigms. The performance of the proposed models and the differences between the two inference approaches are assessed through simulation studies in Section 2.4. In Section 2.5, we apply our methodology to the S-Night plant-capture survey ([Hopper, 1991](#)). A discussion of our findings and directions for further work are presented in Section 2.6.

2.2 Methodology

Plant-capture surveys involve two types of individuals: plants and individuals from the target population. The size of the target population H is unknown and is the target of inference, while the total number of plants, denoted by M , is known by design. In the context of capture without identification, the total number of plants can be partitioned into three observed quantities: $M = M^{yes} + M^{mb} + M^{no}$, where M^{yes} and M^{no} represent, respectively, the number of plants who are certain about having been captured/not cap-

tured, while M^{mb} represents the number of plants who are uncertain (i.e. plants could not determine if they were captured or not).

Other available data include the total number of captured individuals Y , collected by the enumerator during the survey. To describe Y , we introduce Assumption 1.

Assumption 1 *Plants who self-assessed as “yes” were captured and plants who self-assessed as “no” were not captured.*

Assumption 1 describes the certainty in the self-assessment, i.e. the plants’ self-assessments accurately represent their capture status. Under this assumption, we can express Y as the sum of three components: $Y = M^{yes} + M^{mb,c} + H^c$, where $M^{mb,c}$ is the number of plants with uncertain assessment that were captured and H^c is the number of captured individuals from the target population. It is noteworthy that $M^{mb,c}$ and H^c are latent variables. Though we know the sum of $M^{mb,c}$ and H^c via $Y - M^{yes}$, we cannot acquire them individually because we cannot tell whether a captured individual is a plant or not solely by observation.

In order to leverage the information on capture probability acquired from plants to estimate the target population size, we introduce Assumption 2:

Assumption 2 *The capture probability of plants and that of target individuals are equal and homogeneous across individuals.*

Assumption 2 is a fundamental assumption for the plant-capture method, making it possible to apply the estimated capture probability for plants to the target population and derive a target population size estimate.

With the above framework in place, we now propose a basic model to relate the observed and latent variables to the quantity of interest H , along with two extensions to this model to account for uncertainty in capture status.

2.2.1 Basic Model for Uncertain Plant Captures (Model \mathcal{M}_{basic})

Our basic model for estimating the population size H is

$$(M^{yes}, M^{mb}, M^{no}) \mid M \sim \text{Multinom}(M; \boldsymbol{\theta}_1) \quad (2.1)$$

$$M^{mb,c} \mid M^{mb} \sim \text{Binom}(M^{mb}, p^{c|mb}) \quad (2.2)$$

$$H^c \mid H \sim \text{Binom}(H, p^c) \quad (2.3)$$

$$Y = M^{yes} + M^{mb,c} + H^c. \quad (2.4)$$

Here θ_{I} is a vector of probabilities for the plants' self-assessments as “yes”, “maybe” and “no” respectively; p^c represents the capture probability; $p^{c|mb}$ denotes the probability that a plant was captured given that this plant is uncertain about having been captured.

Since M^{yes} , M^{mb} and M^{no} are known, θ_{I} can be estimated. However, p^c , $p^{c|mb}$ and H are not identifiable unless additional assumptions are made. Thus, we introduce a crucial assumption:

Assumption 3.I *For plants, being captured is independent of self-assessing as “maybe”, i.e., $p^{c|mb} = p^c$.*

Assumption 3.I indicates that plants' capture status are missing at random (MAR), which means that the missing capture status of plants can be fully accounted for by their self-assessment, and that there are no other unmeasured variables affecting their capture status (Rubin, 1976). This assumption enables us to use the information from the plants who are certain about their capture status (“Yeses” and “Nos”) to estimate the capture probability of the plants who are uncertain (“Maybes”). Otherwise, p^c and $p^{c|mb}$ are not identifiable. Although inherently untestable, our consideration of this assumption stems from its application as a simplifying assumption in other contexts, particularly in HIV surveillance studies (Vansteelandt et al., 2000; Gustafson, 2023). In this specific scenario, the uncertainty surrounding specific HIV test results bears resemblance to the uncertainty inherent in plant capture. Further discussion of the MAR assumption is provided in Section 2.6.

As a consequence of the introduced assumptions, we can express

$$\theta_{\text{I}} = [p^c(1 - p^{mb}), p^{mb}, (1 - p^c)(1 - p^{mb})]', \quad (2.5)$$

where p^{mb} represents the probability that a plant self-assessed as “maybe”, and is independent of being captured based on Assumption 3.I. Now p^c , which is a critical parameter of the model, can be informed by the ratio of M^{yes} to $M^{yes} + M^{no}$. Further, we can combine Equations (2.2), (2.3), and (2.4) into

$$Y - M^{yes} \mid M^{mb} \sim \text{Binom}(H + M^{mb}, p^c). \quad (2.6)$$

Therefore, we can easily derive the posterior inference of H from Equations (2.1) and (2.6), with θ_{I} specified as in Equation (2.5). We henceforth denote this basic model as $\mathcal{M}_{\text{basic}}$.

2.2.2 Incorporating Partial Identification Data (Model \mathcal{M}_{id})

In Section 2.1, the identity of captured individuals was unknown during the survey. However, in some plant-capture designs, it may be possible to acquire the identity of some captured individuals, as they may be identified via direct contact (e.g. interview) instead of simply being counted. In contrast to the “capture without identification” design considered earlier, we name this design “capture with (partial) identification”. An ideal scenario would be that the identities of all captured individuals are known, where we can apply traditional plant-recapture methods to directly estimate the capture probabilities and subsequently derive population size estimates. But obtaining complete identification data in practice is often challenging. For example, in the S-Night survey described in Section 2.5, enumerators were instructed to interview all individuals encountered in the survey sites who were not in uniform, engaged in money-making activities, sleeping or covered by sleeping bags or blankets between the hours of 2:00 and 4:00 a.m. (Martin et al., 1997). Despite these instructions, only 15% to 70% of plants sighted were interviewed, depending on the city. Therefore, identification data may only be available for part of the augmented population. Still, incorporating this partial identification data can provide valuable insights into capture probabilities and enhance the accuracy of our population size estimates.

In the capture with partial identification design, we introduce a new parameter specifically for captured individuals, which is the probability of being identified, denoted by p^{ic} . Furthermore, to facilitate the exchange of information between plants and the target population, we also introduce a new assumption:

Assumption 4 *All captured individuals have the same probability of being identified.*

In this study design, there is certainty regarding the captured status of identified individuals, rendering self-assessment unnecessary for this subgroup. Consequently, the total number of plants can be decomposed into four observed quantities: $M = M^i + M^{yes} + M^{mb} + M^{no}$, where M^i is the number of identified plants while the definitions of M^{yes} , M^{mb} and M^{no} are similar to those in model \mathcal{M}_{basic} but pertaining to non-identified plants only. The captured count Y can be partitioned into four components: $Y = M^i + M^{yes} + M^{mb,c} + H^c$ under Assumption 1. In addition, we observe H^i , the number of identified target individuals. A model for the “capture with (partial) identification” framework is thus:

$$(M^i, M^{yes}, M^{mb}, M^{no}) | M \sim \text{Multinom}(M; \boldsymbol{\theta}_{\text{II}}) \quad (2.7)$$

$$M^{mb,c} | M^{mb} \sim \text{Binom}(M^{mb}, p^{c|mb,ni})$$

$$H^c | H \sim \text{Binom}(H, p^c) \quad (2.8)$$

$$H^i | H^c \sim \text{Binom}(H^c, p^{i|c}) \quad (2.9)$$

$$Y = M^i + M^{yes} + M^{mb,c} + H^c, \quad (2.10)$$

where $p^{c|mb,ni}$ is the probability that a plant was captured given that it was not identified and that it self-assessed as “maybe”.

With the identification data included in the model, we adapt Assumption 3.I as follows:

Assumption 3.II *Among the non-identified plants, capture by an enumerator is independent of self-assessing as “maybe”, i.e.,*

$$p^{c|mb,ni} = \frac{p^c(1 - p^{i|c})}{p^c(1 - p^{i|c}) + (1 - p^c)}, \quad (2.11)$$

where the right-hand side is the capture probability among the non-identified plants.

Assumption 3.II limits the independence to the non-identified individuals, which means that the MAR mechanism is only assumed among the non-identified individuals since it applies to the plants’ self-assessments.

As a consequence of the assumptions, we have

$$\boldsymbol{\theta}_{\text{II}} = \begin{bmatrix} p^c p^{i|c} \\ p^c(1 - p^{i|c})(1 - p^{mb|ni}) \\ p^c(1 - p^{i|c})p^{mb|ni} + (1 - p^c)p^{mb|ni} \\ (1 - p^c)(1 - p^{mb|ni}) \end{bmatrix}, \quad (2.12)$$

where $p^{mb|ni}$ represents the probability that a plant self-assessed as “maybe” given not identified. We can then estimate H , the size of the target population, by fitting the model described in Equations (2.7) to (2.10) but parameterized by H , p^c , $p^{i|c}$ and $p^{mb|ni}$ via Equations (2.11) and (2.12). We henceforth refer to this model as \mathcal{M}_{id} .

2.2.3 Incorporating Heterogeneity Between Sites (Model \mathcal{M}_{class})

An important assumption for the plant-capture method is that capture probability is constant and equal for plants and individuals from the target population (Laska and Meisner, 1993). However, in practice, this assumption may be violated for various reasons – especially when there are multiple sites enumerated in a survey. For example, in point-in-time street surveys of the homeless, there can be variations in the capture probability across different sites due to several factors such as visual barriers, drug activities, time constraints, or enumerator behavior. Suppose we classify the sites as “hard” if more than 50 percent of plants at one site mentioned any of these problems, and “easy” otherwise. It is reasonable to expect that the probability of being captured by an enumerator would be larger in “easy” sites compared to “hard” sites.

To account for such heterogeneity between sites, it is possible to introduce classes in our model. We assume that there are K classes, and that capture probability varies across these classes, with p_k^c denoting the capture probability of each individual in class $k \in \{1, \dots, K\}$. To formulate the new model, we define the following within each class $k \in \{1, \dots, K\}$:

$$\begin{aligned}
 (M_k^i, M_k^{yes}, M_k^{mb}, M_k^{no}) \mid M_k &\sim \text{Multinom}(M_k; \boldsymbol{\theta}_{\text{III},k}) \\
 M_k^{mb,c} \mid M_k^{mb} &\sim \text{Binom}(M_k^{mb}, p_k^{c|mb,ni}) \\
 H_k^c &\sim \text{Binom}(H_k, p_k^c) \\
 H_k^i &\sim \text{Binom}(H_k, p^{i|c}) \\
 Y_k &= M_k^i + M_k^{yes} + M_k^{mb,c} + H_k^c.
 \end{aligned} \tag{2.13}$$

Here, all notations have similar interpretations as in \mathcal{M}_{id} but are within class k . Note that $p_k^{c|mb,ni}$ is indexed by k because the capture probability varies by class. While here we assume that $p^{i|c}$ and $p^{mb|ni}$ are homogeneous for the sake of demonstration, the model can also be generalized further to accommodate variations in these parameters across different classes by introducing additional parameters specific to each class.

For this class-based model, the assumptions defined for \mathcal{M}_{id} are applied within each class, notably the MAR structure is applied within each class (Assumption 3.III):

Assumption 3.III *Among the non-identified plants, capture by an enumerator is independent of self-assessing as “maybe” within each class, i.e., $p_k^{c|mb,ni} = \frac{p_k^c(1-p^{i|c})}{p_k^c(1-p^{i|c})+(1-p_k^c)}$, for $k = 1, \dots, K$.*

As a result,

$$\boldsymbol{\theta}_{\text{III},k} = \begin{bmatrix} p_k^c p^{i|c} \\ p_k^c (1 - p^{i|c}) (1 - p^{mb|ni}) \\ p_k^c (1 - p^{i|c}) p^{mb|ni} + (1 - p_k^c) p^{mb|ni} \\ (1 - p_k^c) (1 - p^{mb|ni}) \end{bmatrix}.$$

Hence, we can estimate the size of the homeless population H_k within each class, and sum them up to obtain an estimate of the total size of the homeless population $H = \sum_{k=1}^K H_k$. We denote the class-based model described in this section by \mathcal{M}_{class} .

2.3 Inference Approaches

2.3.1 Frequentist Inference via Maximum Likelihood

Frequentist inference – specifically maximum likelihood (ML) estimation – is more straightforward to conduct for model \mathcal{M}_{basic} . Based on Equations (2.1) and (2.6), the joint likelihood of the parameters of interest $\boldsymbol{\gamma} = (H, p^c, p^{mb})$ is

$$\begin{aligned} L(\boldsymbol{\gamma}; y, m^{yes}, m^{mb}, m^{no}) &= P_{\boldsymbol{\gamma}}(Y = y, M^{yes} = m^{yes}, M^{mb} = m^{mb}, M^{no} = m^{no}) \\ &= \frac{(m^{yes} + m^{mb} + m^{no})!}{m^{yes}! m^{mb}! m^{no}!} \{p^c(1 - p^{mb})\}^{m^{yes}} (p^{mb})^{m^{mb}} \{(1 - p^c)(1 - p^{mb})\}^{m^{no}} \times \\ &\quad \binom{H + m^{mb}}{y - m^{yes}} (p^c)^{y - m^{yes}} (1 - p^c)^{H + m^{mb} - y + m^{yes}}. \end{aligned}$$

Maximizing this joint likelihood with respect to $\boldsymbol{\gamma}$ yields the following ML estimators (MLEs): $\widehat{p}^c = M^{yes} / (M^{yes} + M^{no})$, $\widehat{p}^{mb} = M^{mb} / M$ and $\widehat{H} = \lfloor Y / \widehat{p}^c - M \rfloor$, where $\lfloor x \rfloor$ denotes the floor function. Derivations can be found in Appendix A.

In contrast, models \mathcal{M}_{id} and \mathcal{M}_{class} involve latent variables, such that the MLE of H cannot be expressed in closed form. To address this challenge, we adopt a practical strategy of marginalizing out the latent variables numerically, and identifying the mode of the joint likelihood using numerical optimization techniques to obtain the MLE of each parameter. Additionally, we approximate the variance-covariance matrix using the inverse of the negative Hessian matrix of the log-likelihood evaluated at the ML estimates. By leveraging the asymptotic normality property of MLEs, we further provide confidence intervals for the estimated parameters.

The probability functions for model \mathcal{M}_{id} and model \mathcal{M}_{class} can be expressed in a general form $L(\boldsymbol{\gamma}; \boldsymbol{x}) = P_{\boldsymbol{\gamma}}(\mathbf{X} = \boldsymbol{x}) = \sum_{\boldsymbol{z} \in \Omega} P_{\boldsymbol{\gamma}}(\mathbf{X} = \boldsymbol{x}, \mathbf{Z} = \boldsymbol{z})$, where $\boldsymbol{\gamma}$ represents the model

parameters, \mathbf{X} denotes the observed data, \mathbf{Z} represents latent variables, and Ω is the set of possible values for \mathbf{Z} . In model \mathcal{M}_{id} , we have $\mathbf{Z} = M^{mb,c}$, because knowing $M^{mb,c}$ also provides information about the other latent variable H^c through Equation (2.10). In this context, $\boldsymbol{\gamma} = (\log(H), \text{logit}(p^c), \text{logit}(p^{i|c}), \text{logit}(p^{mb|ni}))$ and $\mathbf{X} = (Y, M^i, M^{yes}, M^{mb}, M^{no}, H^i)$. By summing over \mathbf{Z} , we derive a marginal likelihood of \mathbf{X} that excludes latent variables, thereby facilitating the direct application of the MLE method. Note that some parameters within the model represent counts, taking on positive integer values, while others represent probabilities that range between 0 and 1. Therefore, we apply a log transformation on the counts and a logit transformation on the probabilities, eliminating any constraints on their bounds to prevent computational issues arising from boundary constraints.

Determining Ω depends on the domain of $M^{mb,c}$ and the domain of H^c . For \mathcal{M}_{id} , we establish the bounds based on two constraints. The first constraint,

$$0 \leq M^{mb,c} \leq M^{mb}, \quad (2.14)$$

is simply due to the domain of $M^{mb,c}$. In addition, the relationship between $M^{mb,c}$ and H^c in Equation (2.10) implies $M^{mb,c} = Y - M^i - M^{yes} - H^c$, with H^c bounded within its domain $H^i \leq H^c \leq H$. The second constraint arises as a consequence:

$$Y - M^i - M^{yes} - H \leq M^{mb,c} \leq Y - M^i - M^{yes} - H^i. \quad (2.15)$$

Combining constraints (2.14) and (2.15), we have $\Omega = [a_{II}, b_{II}]$ with

$$\begin{aligned} a_{II} &= \max(0, Y - M^i - M^{yes} - H) \\ b_{II} &= \min(M^{mb}, Y - M^i - M^{yes} - H^i). \end{aligned}$$

Similarly, for \mathcal{M}_{class} we use $\mathbf{Z} = (M_1^{mb,c}, \dots, M_K^{mb,c})'$ and $\Omega = \Omega_1 \times \dots \times \Omega_K$ where $\Omega_k = [a_{k,III}, b_{k,III}]$ with

$$\begin{aligned} a_{k,III} &= \max(0, Y_k - M_k^i - M_k^{yes} - H_k) \\ b_{k,III} &= \min(M_k^{mb}, Y_k - M_k^i - M_k^{yes} - H_k^i). \end{aligned}$$

Consider the log-likelihood $\log\{L(\boldsymbol{\gamma}, \mathbf{x})\}$ where \mathbf{x} is the observed data from \mathbf{X} . The ML estimator of the vector of parameters $\boldsymbol{\gamma}$ is the mode of the log-likelihood which can be approximated numerically using the Nelder–Mead optimization method (Nelder and Mead, 1965). This method is easy to implement with the function `optim` in R (R Core Team, 2024). To estimate the variance-covariance matrix, we numerically approximate

the Hessian matrix using Richardson extrapolation, which can be carried out with the function `hessian` from the package `numDeriv` (Gilbert and Varadhan, 2019) in R. The same numerical method is also applied to Model \mathcal{M}_{basic} in this study for simplicity.

Notably, given that our models involve parameter transformations within γ , it is essential to revert the MLEs ($\hat{\gamma}$), variance estimators ($\widehat{\text{Var}}(\hat{\gamma})$), and confidence intervals (CIs) to their original scales. For MLEs, this is achieved by applying the corresponding inverse functions to the estimates. For the variance estimators, we use the delta method to transform them to the original scale. For the CIs, we first construct the 95% CI of each parameter on the transformed scale as $\hat{\gamma} \pm 1.96\sqrt{\widehat{\text{Var}}(\hat{\gamma})}$. Then we apply the corresponding inverse functions on these CIs to obtain the CIs on the original scale.

2.3.2 Bayesian Inference via MCMC

When dealing with complex models, marginalizing out all latent variables may prove inefficient or impractical. Given the hierarchical structure inherent in our proposed models, a Bayesian framework, coupled with Markov Chain Monte Carlo (MCMC) methods, emerges as an effective alternative for conducting inference. Notably, probabilistic programming languages for MCMC sampling have a broad appeal with applied scientists as they allow symbolic coding of a hierarchical model along with its priors. This allows users flexibility in extending or customizing models (e.g. sharing parameters across years or specifying exchangeable parameters via hierarchical priors). Additionally, these languages conveniently come equipped with built-in MCMC algorithms for posterior distribution sampling.

Among probabilistic programming languages for MCMC sampling, those based on BUGS (Gilks et al., 1994) stand out for their unique ability to handle discrete latent variables, which is a feature present in our models. These languages include, for example, JAGS (Plummer et al., 2003) and NIMBLE (de Valpine et al., 2017). When using BUGS-based languages, we need to be careful about their grammar rules. For example, in Equations (2.10) and (2.13), Y is represented as the sum of two observed variables and two unobserved variables, while we have the observed value for Y as data. In this case, one should use the `dsum` function in JAGS so that the sampler will update the unobserved variables together ensuring that the sum constraint is preserved. There are also alternative ways, such as defining custom functions and distributions in NIMBLE or employing the zeros-ones trick (Ntzoufras, 2011). The code to implement our models is provided in the Supplementary Materials.

Despite the simplicity of conducting Bayesian inference, one of the drawbacks of Bayesian inference is that it may be sensitive to the choice of the prior. Though we strive for dif-

fuse priors when there is no prior knowledge available, the results could still be affected by the diffuse priors we choose. Further discussion on the choices of prior can be found in Gelman et al. (2013), and specific choices of priors for our analysis will be presented in Section 2.4. Besides, when dealing with complex models, BUGS-based languages can be computationally expensive. To address this issue, refer to Appendix B for two alternative computational approaches that enhance efficiency, and Supplementary Tables in Appendix C for additional simulation study results.

2.4 Simulation Study

To evaluate the performance of our models, we conducted a simulation study for each of the models, leveraging information from the S-night survey to emulate real-world conditions in homeless population size estimation. In each study, we considered two distinct scenarios: small city and large city. In the context of the small city scenario, we set the true value of (M, H) to $(15, 150)$ for \mathcal{M}_{basic} and \mathcal{M}_{id} , and $(30, 300)$ for \mathcal{M}_{class} . For the large city scenario, the true value of (M, H) is set to $(100, 1,500)$ for all three models. This contrast in city sizes allows us to discern potential variations in the performance of our methods when applied to cities with varying homeless population sizes.

Under each of these six specified settings, we simulated 1,000 datasets. We set the true values for (p^c, p^{mb}) to $(0.7, 0.2)$ in \mathcal{M}_{basic} , and we set $(p^c, p^{mb|ni}, p^{ic})$ to $(0.7, 0.2, 0.8)$ in \mathcal{M}_{id} . In \mathcal{M}_{class} , we assumed that there are two classes for the sites: easy and hard. For this setup, the true capture probability for easy sites was set to $p_1^c = 0.9$ while the capture probability for hard sites was set to $p_2^c = 0.4$. We also assumed that 60% of the sites were easy while the remaining 40% were hard, leading to an overall capture probability of 0.7, consistent with our setting for $p^c = 0.7$ in the other studies. Furthermore, we distributed the plants into these site classes proportionally, with 60% of plants in easy sites and 40% in hard sites. The true values for $(p^{mb|ni}, p^{ic})$ were also set to $(0.2, 0.8)$.

The inference for each dataset was conducted both via Maximum Likelihood and Bayesian estimation, which we aim to compare. When performing Bayesian inference, we used independent Uniform(0,1) priors on all parameters representing probabilities. For models \mathcal{M}_{basic} and \mathcal{M}_{id} , a log-normal prior (rounded to the nearest integer) with mean 0 and variance 100 was specified on the population size H ; for model \mathcal{M}_{class} equivalent independent prior were defined for every H_k . The MCMC algorithm was implemented in JAGS, employing 3 chains, each comprising 30,000 iterations. We treated the first 15,000 iterations as burn-in to guarantee the stability and convergence of our results.

Table 2.1: Results of the simulation studies with \mathcal{M}_{basic} for the MLEs and the Bayesian estimators. All the values are rounded to integers or 2 decimal points.

Method	M	Parameter	True Value	Estimate	SD	RBias	RRMSE	CP	LCI
MLE	15	H	150	149	31	-0.01	0.24	0.85	126
		p^c	0.7	0.73	0.12	0.04	0.19	0.98	0.48
		p^{mb}	0.2	0.20	0.10	0.01	0.51	0.98	0.42
Bayesian	15	H	150	159	43	0.06	0.24	0.97	160
		p^c	0.7	0.68	0.12	-0.03	0.16	0.97	0.45
		p^{mb}	0.2	0.23	0.10	0.14	0.49	0.98	0.37
MLE	100	H	1,500	1,497	114	-0.00	0.08	0.93	449
		p^c	0.7	0.70	0.05	0.01	0.07	0.95	0.20
		p^{mb}	0.2	0.20	0.04	0.00	0.20	0.94	0.16
Bayesian	100	H	1,500	1,513	120	0.01	0.08	0.94	466
		p^c	0.7	0.70	0.05	-0.00	0.07	0.94	0.20
		p^{mb}	0.2	0.20	0.04	0.02	0.20	0.94	0.15

In each of our simulation studies, we employed a comprehensive set of evaluation metrics to assess the performance of Bayesian and MLE approaches. These metrics include: average estimates, average standard deviations (SD), relative Monte Carlo biases (RBias), relative root mean squared errors (RRMSE), coverage probabilities (CP) and average lengths of the 95% confidence interval or credible interval (LCI). Note that for the Bayesian method, we used the expected posterior medians as estimates.

We begin with the simulation and analysis setting for model \mathcal{M}_{basic} , the results of which are shown in Table 2.1. In the small city scenario, the MLE of H has a negligible bias, whereas the Bayesian estimator has a relative bias of 6%. However, the coverage probability of the MLE is too low at 85%. In the large city scenario, both estimators perform similarly with negligible biases and coverage probabilities close to 95%. Additionally, we observe that although the two estimators have the same relative root mean squared errors, the Bayesian estimator yields a larger LCI compared to the MLE.

Moving on to model \mathcal{M}_{id} , we present the results of this analysis in Table 2.2, which demonstrate a pattern akin to that observed in model \mathcal{M}_{basic} . The Bayesian estimator of H yields a larger bias, while the MLE has a coverage probability farther from 95%. And the gap diminishes when transitioning to the large city scenario. Besides, the RRMSE and LCI follow a similar pattern as observed in model \mathcal{M}_{basic} .

Finally, in the context of model \mathcal{M}_{class} , the analysis results presented in Table 2.3 offer insights into the performance of this model. Our findings in \mathcal{M}_{class} echo the trends

Table 2.2: Results of the simulation studies with \mathcal{M}_{id} for the MLEs and the Bayesian estimators. All the values are rounded to integers or 2 decimal points.

Method	M	Parameter	True Value	Estimate	SD	RBias	RRMSE	CP	LCI
MLE	15	H	150	150	29	0.00	0.22	0.88	118
		p^c	0.7	0.72	0.12	0.03	0.18	0.98	0.46
		$p^{mb ni}$	0.2	0.21	0.13	0.04	0.83	0.96	0.70
		$p^{i c}$	0.8	0.80	0.04	-0.00	0.05	0.95	0.15
Bayesian	15	H	150	159	38	0.06	0.22	0.97	142
		p^c	0.7	0.68	0.11	-0.03	0.16	0.97	0.43
		$p^{mb ni}$	0.2	0.26	0.14	0.30	0.73	0.96	0.52
		$p^{i c}$	0.8	0.80	0.04	-0.01	0.05	0.95	0.14
MLE	100	H	1,500	1,498	107	-0.00	0.07	0.93	420
		p^c	0.7	0.70	0.05	0.01	0.07	0.94	0.19
		$p^{mb ni}$	0.2	0.20	0.06	-0.00	0.30	0.97	0.24
		$p^{i c}$	0.8	0.80	0.01	0.00	0.02	0.96	0.05
Bayesian	100	H	1,500	1,512	111	0.01	0.07	0.94	433
		p^c	0.7	0.70	0.05	-0.00	0.07	0.94	0.18
		$p^{mb ni}$	0.2	0.21	0.06	0.04	0.29	0.96	0.23
		$p^{i c}$	0.8	0.80	0.01	0.00	0.02	0.96	0.05

Table 2.3: Results of the simulation studies with \mathcal{M}_{class} for the MLEs and the Bayesian estimators. All the values are rounded to integers or 2 decimal points.

Method	M	Parameter	True Value	Estimate	SD	RBias	RRMSE	CP	LCI
MLE	30	H	300	313	65	0.04	0.25	0.97	358
		p_1^c (easy)	0.9	0.90	0.07	0.00	0.08	0.96	0.37
		p_2^c (hard)	0.4	0.42	0.14	0.06	0.38	0.97	0.52
		$p^{mb ni}$	0.2	0.19	0.10	-0.07	0.54	0.98	0.46
		$p^{i c}$	0.8	0.80	0.03	0.00	0.03	0.95	0.10
Bayesian	30	H	300	326	87	0.09	0.20	0.97	314
		p_1^c (easy)	0.9	0.84	0.08	-0.06	0.09	0.94	0.32
		p_2^c (hard)	0.4	0.42	0.13	0.04	0.32	0.96	0.49
		$p^{mb ni}$	0.2	0.22	0.10	0.09	0.49	0.97	0.38
		$p^{i c}$	0.8	0.80	0.03	-0.00	0.03	0.94	0.10
MLE	100	H	1,500	1,510	142	0.01	0.10	0.97	702
		p_1^c (easy)	0.9	0.91	0.04	0.01	0.05	0.93	0.16
		p_2^c (hard)	0.4	0.40	0.08	0.01	0.20	0.97	0.30
		$p^{mb ni}$	0.2	0.20	0.06	-0.01	0.30	0.96	0.23
		$p^{i c}$	0.8	0.80	0.01	-0.00	0.02	0.95	0.05
Bayesian	100	H	1,500	1,535	155	0.02	0.10	0.96	601
		p_1^c (easy)	0.9	0.89	0.04	-0.01	0.05	0.96	0.16
		p_2^c (hard)	0.4	0.40	0.08	0.01	0.19	0.96	0.30
		$p^{mb ni}$	0.2	0.21	0.06	0.04	0.29	0.95	0.23
		$p^{i c}$	0.8	0.80	0.01	-0.00	0.02	0.95	0.05

identified in previous simulation studies, particularly for the large city scenario. However, in the small city scenario, we observe some deviations from the trends observed earlier. Specifically, while the coverage probability of the MLE of H is the same as that of the Bayesian estimator, the MLE has a smaller relative bias (4% compared to 9%) with a larger RRMSE and a larger LCI. The difference could be attributed to the small sample size (M), resulting in less information available in each class, especially within the classes with small capture probabilities. Further discussion will be provided in Section 2.6.

Overall, both MLE and Bayesian estimators provide an accurate estimate of the target population size. The findings from the simulation studies shed light on the trade-off between bias and coverage probability in our models. The choice between MLE and Bayesian methods should be made based on the specific characteristics of the models and the desired objectives of the analysis.

Table 2.4: The 1990 S-Night data reconstructed from the literature.

	Chicago	New Orleans	Phoenix	New York	Los Angeles
Plants (M)	13	58	26	94	25
Interviewed (M^i)	2	41	18	40	16
Yes (M^y)	0	6	3	19	1
Maybe (M^m)	5	5	1	13	2
No (M^n)	6	6	4	22	6
Census (Y)	11	109	104	1,240	217

2.5 Application to the S-Night Street Enumeration Survey

In this section, we apply our method to re-analyze the plant-capture data from the iconic 1990 ‘‘S-Night’’ survey. On the night of March 20-21, 1990, the United States Census Bureau carried out the Shelter and Street Night Enumeration, also known as S-Night (Barrett et al., 1992). The survey was conducted in five major cities: New Orleans, New York, Phoenix, Los Angeles and Chicago. Prior to the enumeration, a known number of plants, trained to dress and act like homeless people, were deployed at designated sites. The plants were instructed to stay in an open area to allow the enumerators to see and enumerate them during street enumeration between 2 to 4 a.m., and enumerators were asked to interview all individuals encountered in the pre-assigned sites. After the enumeration, the plants were requested to fill out questionnaires to report whether an enumerator interviewed them and whether they believed they were counted by an enumerator if not interviewed. For more details refer to Martin (1992). Note that this survey is designed to estimate the homeless population size present in the areas targeted during the survey; it is not meant to provide an exhaustive count of the homeless population in the entire city.

The data we use for our demonstration are shown in Table 2.4. Given the time elapsed since the original study, the original data could not be retrieved, therefore we reconstructed Table 2.4 approximately from data summaries published in Martin (1992) and Martin et al. (1997). Notably, while the number of plants interviewed could be reconstructed from the literature, the number of homeless interviewed and the exact counts in easy/hard sites were not available. Thus our methodology is demonstrated using a modified version of model \mathcal{M}_{id} . In this variant, Equation (2.9) is omitted due to the unavailability of H^i .

We conduct the data analysis separately for each city. The results for both MLEs

and Bayesian estimators (posterior medians) are presented in Table 2.5. For the Bayesian inference, we applied the prior specifications described in Section 2.4. For our target, the homeless population size H , the Bayesian method provided larger estimates and standard deviations compared to MLE, except for Chicago. This could be due to a sensitivity to the prior, which tends to inflate estimates slightly. However, Chicago stands out with significantly distinct results between Bayesian estimators and MLEs in comparison to other cities. This discrepancy arises from the absence of plants self-assessing as “yes” in Chicago. As a result, the MLE of $p^{i|c}$ is 1 (which is an overestimate) with an estimated variance of zero. In contrast, Bayesian estimates are influenced by the prior setting, introducing variability and leading to divergent results between the two methods.

Although we do not have the true value for these parameters, the consistency in estimation outcomes suggests that both Bayesian and MLE approaches offer valuable insights. The choice between the methods should be guided by the specific analytical needs and the availability of prior information, ensuring researchers can harness the most suitable technique for their particular context.

As a point of comparison, Supplementary Table C4 presents the estimates and the 95% confidence intervals obtained using the hybrid Chapman-Bailey estimator, as described in [Laska and Meisner \(1993\)](#). This estimator is calculated under two extreme scenarios considered in [Martin et al. \(1997\)](#), representing contrasting treatments of “maybe” responses. In one scenario all the plants who self-assessed as “maybe” are considered as “yes”, while in the other, they are treated as “no”, since the Chapman-Bailey estimator cannot handle uncertain assessments. Our estimates are all included in the 95% CI for both scenarios, except for Chicago, where only the 95% CI under the second scenario includes our estimates.

2.6 Discussion

In this work, we have introduced a novel plant-capture modeling framework that incorporates uncertain assessment of capture and can allow for partial identification of plants as well as heterogeneity across survey sites. Within this framework, we have proposed two distinct inference approaches: frequentist maximum likelihood estimation and a Bayesian methodology. Our simulation studies have demonstrated the Bayesian approach’s ability to achieve coverage probabilities close to the desired 95% while exhibiting a slightly larger bias compared to the MLEs. In contrast, MLEs tend to have a smaller bias but the coverage probabilities can occasionally deviate from the ideal 95% in small population settings. Importantly, our simulations have revealed that as the population size increases, the

Table 2.5: Results of the application to the S-Night data using \mathcal{M}_{id} without Equation (2.9) separately in each city. All the values are rounded to integers or 2 decimal points.

Parameter	Bayesian			MLE		
	Estimate	SD	95% CrI	Estimate	SD	95% CI
Chicago						
H	37	40	(11, 156)	54	38	(13, 217)
p^c	0.22	0.12	(0.06, 0.51)	0.16	0.10	(0.04, 0.46)
$p^{mb ni}$	0.46	0.13	(0.21, 0.72)	0.45	0.15	(0.20, 0.73)
$p^{i c}$	0.71	0.22	(0.21, 0.99)	1.00	0.00	(1.00, 1.00)
New Orleans						
H	70	7	(61, 87)	69	6	(58, 82)
p^c	0.84	0.05	(0.73, 0.93)	0.86	0.05	(0.73, 0.94)
$p^{mb ni}$	0.31	0.10	(0.13, 0.54)	0.29	0.11	(0.13, 0.54)
$p^{i c}$	0.82	0.06	(0.69, 0.91)	0.83	0.06	(0.68, 0.91)
Phoenix						
H	102	12	(87, 135)	98	10	(80, 120)
p^c	0.81	0.08	(0.64, 0.93)	0.84	0.08	(0.64, 0.94)
$p^{mb ni}$	0.18	0.12	(0.03, 0.49)	0.12	0.12	(0.02, 0.54)
$p^{i c}$	0.82	0.08	(0.63, 0.94)	0.84	0.08	(0.61, 0.94)
New York						
H	1,709	142	(1,494, 2,056)	1,688	131	(1,450, 1,964)
p^c	0.69	0.05	(0.57, 0.78)	0.70	0.05	(0.59, 0.79)
$p^{mb ni}$	0.25	0.06	(0.15, 0.37)	0.24	0.06	(0.14, 0.37)
$p^{i c}$	0.61	0.06	(0.49, 0.73)	0.61	0.06	(0.48, 0.73)
Los Angeles						
H	290	47	(233, 415)	282	40	(215, 372)
p^c	0.69	0.09	(0.49, 0.84)	0.71	0.09	(0.50, 0.86)
$p^{mb ni}$	0.26	0.13	(0.07, 0.56)	0.22	0.14	(0.06, 0.58)
$p^{i c}$	0.89	0.08	(0.69, 0.98)	0.92	0.07	(0.63, 0.99)

discrepancy between the two methods diminishes, and their performance improves simultaneously. However, an exception arises under \mathcal{M}_{class} : it may have suboptimal performance due to insufficient information about different sites, especially when the population size in each site is relatively small. Furthermore, we have applied these models to estimate the homeless population size using the 1990 S-Night data, shedding light on their real-world applicability.

The insights gained from this research have the potential to significantly contribute to public health planning and policy formulation, especially with regard to addressing the needs of vulnerable populations. Estimating the vulnerable population size within a society holds immense importance due to the multifaceted impact on their quality of life. Access to housing play vital roles in education and labor market participation. Moreover, quantifying vulnerable populations, such as the homeless population, is instrumental in shaping governmental policies, particularly those related to housing provisions, and enables the evaluation of intervention effectiveness (Coumans et al., 2017).

To explore the applicability of our models in real-world scenarios, further investigations are warranted, with a particular focus on assessing the validity of the independence assumptions. The independence (MAR) assumptions introduced in Section 2.2 play a crucial role in our models, allowing estimation of the size of the target population despite uncertainty arising from the plants self-assessed as “maybe”. However, it is important to recognize that these assumptions are essentially untestable and may not hold under certain circumstances, particularly when there is a preference for “yes” or “no” responses among the “maybe” category. An example in the homeless survey could be, if a significant portion of “maybe” plants were sleeping in locations that were difficult for enumerators to observe, these plants would be more likely to go unnoticed, thus violating the assumption of independence, unless sleeping locations are used to define classes within model \mathcal{M}_{class} . However, there is a limit to the number of classes that should be used as the model can fail to provide accurate estimators in the cases when either no plants self-assess as “yes”, or no plants self-assess as “no”. Under these situations, all the plants certain about their status belong to a single group, leaving no information available about the other group. As a result, all the plants self-assessing as “maybe” will be categorized into the same group, which is an extreme ratio for the “maybe” plants and potentially leads to an imprecise estimator, especially when the number of “maybe” plants is relatively large. For this reason, we recommend that \mathcal{M}_{class} presented in Section 2.2.3 should only be applied when there are sufficient plants in each class to avoid this issue. Or, one could pool information about unknown parameters across sites with the help of a random-effects model. Another potential avenue for improvement is to consider a more flexible modeling approach. Instead of setting equality in Assumptions 3.I, 3.II and 3.III, it may be possible to relax the

MAR assumptions by treating these probabilities as exchangeable via a hierarchical prior that controls the degree to which the MAR assumption is relaxed. This would propagate additional uncertainty into the final estimates.

Finally, there may be room for improvement in addressing variations between different survey sites through model refinements. For instance, a more sophisticated approach might involve modeling site-specific probabilities using logistic regression by incorporating covariates such as site characteristics and GPS location data. Notably, the Counting Us Mobile App ([Simtech Solutions, Inc., 2023](#)) has already demonstrated its capability to track the locations of enumerators and plants during surveys, facilitating point-in-time counts in 50 regions across the United States. Utilizing data on the distances between plants and enumerators could potentially lead to a more accurate estimation of capture probabilities. This could enable adjustments in cases where enumerators were delayed, positioned inaccurately, or absent altogether. Additionally, such detailed spatial data could provide valuable insights into the validity of our MAR assumptions, allowing for the modeling of deviations from this assumption and enhancing the robustness of population size estimates.

Chapter 3

Rapid Scaling of Compositional Uncertainty from Sample to Population Levels

3.1 Introduction

Determining the proportion of various subgroups within a population is crucial in various ecological, evolutionary, conservation, and management applications (e.g., [Allendorf et al., 2012](#)).

Consider a mixed population of animals from different geographic and genetic origins forming $K > 2$ distinct subpopulations differentiable through genetic means. The aim is to estimate the population composition, denoted by the vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)'$ on the simplex, from the analysis of genetic data collected from samples. Such analysis may rely on a probabilistic method named genetic stock identification (GSI) ([Grant et al., 1980](#); [Pella and Masuda, 2001](#); [Hess et al., 2014](#)), which compares the genotypes of the sampled individuals with those of a baseline population. Ideally, estimates are obtained through mixture analysis (MA), a type of GSI analysis accompanied by standard errors that reflect uncertainty in the population composition (e.g., via resampling or Bayesian methods; see [Koljonen et al. \(2005\)](#), [Kuismin et al. \(2020\)](#)). However, published estimates are sometimes obtained via individual assignments (IA) ([Cornuet et al., 1999](#); [Luikart and England, 1999](#)) rather than MA, and in this case only accompanied by standard error estimates of the sample composition.

Our goal is to propose rapid strategies to rescale these standard error estimates from the sample- to population-level. Our approach has the advantage of not requiring understanding nor modeling of how IA estimates were obtained (e.g. Bayesian methods of [Pella and Masuda \(2001\)](#)), treating these methods as a black box. However, we make a key simplifying assumption that the only source of correlation between composition estimates is their constraint to sum to one, as data on correlations between groups are rarely published alongside means and standard errors. For GSI data, this assumption may require that proportions are reported at the level of “reporting units” (e.g. regions) instead of stocks, to prevent strong correlations between stocks that are difficult to differentiate genetically ([Millar, 1991](#)). Based on this premise, we propose a reverse Dirichlet-multinomial model for propagating uncertainties and derive analytical formulas for population-level variance estimators.

We extend our methods to the context of genetic mark-recapture (GMR), specifically targeting the estimation of escapement in mixed-stock salmon fisheries ([Hamazaki and DeCovich, 2014](#)). Escapement refers to the number of adult salmon that migrate from the ocean back to their natal freshwater habitats for reproduction, while escaping mortality along the way. In fisheries science, determining population proportions in mixed-stock fisheries is vital for enabling managers to target specific stocks in accordance with their abundance levels ([Östergren et al., 2020](#)). GMR is particularly useful when there is a nearly complete escapement count for some stocks (e.g., those born in lakes) but not for others (e.g., those born in rivers). In mark-recapture terminology, lake-type stocks are regarded as “marked”, while those identified in the GSI sample are regarded as “recaptured”. The overall escapement is estimated by dividing the escapement count of lake-type stocks by their estimated proportion in the GSI sample, using a Petersen-type estimator. We address the challenge of estimating escapement and its associated standard error in GMR when standard errors on composition are available only at the sample level. Our solution involves a Bayesian approach, introducing a novel prior that leverages autocorrelation between weeks to mitigate the impact of the prior on the inference. Additionally, we develop analytical formulas to estimate the variance of the escapement estimator.

In [Section 3.2](#), we introduce a new approach for estimating population composition using sample proportions, accompanied by two distinct variance estimators. [Section 3.3](#) extends our methodology to the domain of GMR, offering both Bayesian and frequentist techniques to estimate total escapement in mixed-stock fisheries along with its variance. In [Section 3.4](#), we conduct a simulation study to evaluate the performance of the proposed methods. [Section 3.5](#) demonstrates the practical application of our methodologies on Sockeye Salmon data collected in the Taku River in 2017, providing a comparative analysis across different settings. A discussion of our methods and directions for further

work are presented in Section 3.6.

3.2 Methodology

Our goal is to estimate a vector of population proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)'$ (satisfying $\sum_{k=1}^K \pi_k = 1$) based on compositional data obtained from a sample of size n , which provides a summary vector of estimated proportions $\hat{\boldsymbol{p}} = (\hat{p}_1, \dots, \hat{p}_K)'$, along with a corresponding vector of standard errors $\boldsymbol{s} = (s_1, \dots, s_K)$. In this section, we present a comprehensive methodology, followed by an elaboration on its applicability within the realm of GMR in Sections 3.3 and 3.4.

Let $\boldsymbol{X} \sim \text{Multinomial}(n, \boldsymbol{\pi})$ be a vector of latent sample sizes in groups 1 to K . The proportion of the sample belonging to each group is $\boldsymbol{\rho} = (\rho_1, \dots, \rho_K)' = \boldsymbol{X}/n$. We aim to relate the data $\hat{\boldsymbol{p}}$ to $\boldsymbol{\rho}$, while accounting for the standard errors, \boldsymbol{s} . To link the sample and the population, we assume that the vector of estimated sample proportions $\hat{\boldsymbol{p}}$ is a random variable within the $(K - 1)$ -dimensional simplex with a mean $E(\hat{\boldsymbol{p}} \mid \boldsymbol{X}) = \boldsymbol{\rho}$ and variances $\text{Var}(\hat{p}_k \mid X_k) = s_k^2$. Although s_k^2 is an estimate for the true variance of \hat{p}_k , we simplify by assuming that the standard errors are known standard deviations, a practice similar to that used historically in meta-analysis of aggregate data (Borenstein et al., 2021).

3.2.1 Approximate Reverse Dirichlet-Multinomial Model

The Dirichlet distribution is the most well-known distribution on the simplex, so we propose an approximate generative model for the compositional data as follows:

$$\hat{\boldsymbol{p}} \mid \boldsymbol{X} \sim \text{Dirichlet}(\lambda \boldsymbol{\rho}), \tag{3.1}$$

$$\boldsymbol{X} \sim \text{Multinomial}(n, \boldsymbol{\pi}), \tag{3.2}$$

where a scaling factor λ is introduced to control the variance of the Dirichlet distribution. We term this the approximate reverse Dirichlet-multinomial model, with “reverse” indicating that the data follow a Dirichlet distribution rather than the parameters of interest, distinguishing it from the conventional Dirichlet-multinomial model. It is important to note that by employing a Dirichlet distribution, we are assuming that the correlation between \hat{p}_k and $\hat{p}_{k'}$ ($k \neq k'$) given ρ_k , arises solely from the constraint of the simplex and not from any other factors (in contrast to e.g. a multivariate logistic normal distribution). The Dirichlet distribution, thus, represents the “ultimate in independence hypotheses” (Aitchison, 1986).

It is typically unfeasible to configure λ to match exactly all predetermined standard errors s_1, \dots, s_k (Aitchison, 1986; Gelman, 1995). Therefore, we propose selecting λ to minimize the discrepancy between the variance of the Dirichlet distribution and the observed variances in the data. This is likely to be a reasonable approximation due to the specific relationship between the variance and the mean of the Dirichlet distribution: $\text{Var}(\hat{p}_k | \mathbf{X}) \propto \text{E}(\hat{p}_k | \mathbf{X})(1 - \text{E}(\hat{p}_k | \mathbf{X}))$, with

$$\begin{aligned} \text{E}(\hat{p}_k | \mathbf{X}) &= \frac{\lambda \rho_k}{\sum_i \lambda \rho_i} = \frac{\lambda \rho_k}{\lambda} = \rho_k, \\ \text{Var}(\hat{p}_k | \mathbf{X}) &= \frac{\frac{\lambda \rho_k}{\sum_i \lambda \rho_i} (1 - \frac{\lambda \rho_k}{\sum_i \lambda \rho_i})}{\sum_i \lambda \rho_i + 1} = \beta \rho_k (1 - \rho_k), \end{aligned} \quad (3.3)$$

where $\beta = (\lambda + 1)^{-1}$. We expect the relationship in Equation (3.3) to approximately hold on the data as well, i.e., $s_k^2 \approx \beta \hat{p}_k (1 - \hat{p}_k)$, indicating that estimated variance s_k^2 have an approximate quadratic relationship to \hat{p}_k , becoming close to zero when \hat{p}_k is close to 0 or 1. This pattern can easily be assessed on the data by plotting s_k^2 against $\hat{p}_k(1 - \hat{p}_k)$. In our application in Section 3.5, this pattern is closely satisfied (Figure 3.3b). If the assumption is violated, the Dirichlet distribution may not be appropriate for modeling the data, and proceeding with this distribution may not be advisable in such cases.

Formally, we propose to determine β using Equation (3.4), which minimizes the squared differences between the observed variance in the GSI dataset and those predicted by the Dirichlet model, where $\hat{p}_{k,t}^{\text{obs}}$ and $s_{k,t}^{\text{obs}2}$ represent the observed stock proportions and corresponding variances:

$$\hat{\beta} = \arg \min_{\beta} \sum_{k=1}^K \left\{ s_k^{\text{obs}2} - \beta \hat{p}_k^{\text{obs}} (1 - \hat{p}_k^{\text{obs}}) \right\}^2. \quad (3.4)$$

The solution to Equation (3.4) is the least squares estimator of β in a simple linear regression model without intercept $s_k^{\text{obs}2} = \beta \hat{p}_k^{\text{obs}} (1 - \hat{p}_k^{\text{obs}}) + \epsilon_k$, $k = 1, \dots, K$, where $\hat{p}_k^{\text{obs}} (1 - \hat{p}_k^{\text{obs}})$ is the predictor, and ϵ_k is an error term. Hence, we define

$$\lambda = \hat{\beta}^{-1} - 1 = \frac{\sum_k \{ \hat{p}_k^{\text{obs}} (1 - \hat{p}_k^{\text{obs}}) \}^2}{\sum_k \hat{p}_k^{\text{obs}} (1 - \hat{p}_k^{\text{obs}}) (s_k^{\text{obs}})^2} - 1. \quad (3.5)$$

3.2.2 Simple Analytical Formulas for Scaling Standard Errors

For the reverse Dirichlet-multinomial model described in Equations (3.1) and (3.2), $\hat{\boldsymbol{\pi}}$ is an unbiased estimator for $\boldsymbol{\pi}$, since $E(\hat{\boldsymbol{\pi}}) = E\{E(\hat{\boldsymbol{\pi}} \mid \mathbf{X})\} = E(\mathbf{X}/n) = \boldsymbol{\pi}$. The variance of this estimator for estimating the population-level parameter $\boldsymbol{\pi}$ is $\text{Var}(\hat{\boldsymbol{\pi}}) = E\{\text{Var}(\hat{\boldsymbol{\pi}} \mid \mathbf{X})\} + \text{Var}\{E(\hat{\boldsymbol{\pi}} \mid \mathbf{X})\}$, according to the law of total variance. The first term is

$$\begin{aligned} E\{\text{Var}(\hat{\boldsymbol{\pi}} \mid \mathbf{X})\} &= E[\beta\{\text{diag}(\mathbf{X})/n - \mathbf{X}\mathbf{X}'/n^2\}] \\ &= \beta(1 - 1/n)\{\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\}, \end{aligned}$$

while the second term is $\text{Var}\{E(\hat{\boldsymbol{\pi}} \mid \mathbf{X})\} = \text{Var}(\mathbf{X}/n) = \frac{\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'}{n}$. Summing the two terms, we arrive at the expression:

$$\text{Var}(\hat{\boldsymbol{\pi}}) = \tilde{\beta}\{\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\}, \quad (3.6)$$

where

$$\tilde{\beta} = \frac{n-1}{n}\hat{\beta} + \frac{1}{n},$$

which is a decreasing function of n bounded between β and 1. It approaches 0 when n is large since β approaches 0 when \mathbf{s} approaches $\mathbf{0}$.

Since marginally $\text{Var}(\hat{p}_k) = \tilde{\beta}\pi_k(1 - \pi_k)$, we propose a straightforward estimator for $\sigma_k^2 \equiv \text{Var}(\hat{p}_k)$, that is $\hat{\sigma}_k^2 = \tilde{\beta}\hat{p}_k(1 - \hat{p}_k)$. An alternate estimator, which leverages the approximate proportional relationship between s_k^2 and $\hat{p}_k(1 - \hat{p}_k)$ described in Section 3.2.1, is $\hat{\sigma}_k^{2,\text{alt}} = \frac{\tilde{\beta}}{\beta}s_k^2 = (1 + \frac{\lambda}{n})s_k^2$. While it has the desirable property $\hat{\sigma}_k^{2,\text{alt}} \geq s_k^2$, this property is not guaranteed to hold for $\hat{\sigma}_k^2$. Hence, $\hat{\sigma}_k^{2,\text{alt}}$ may be more robust against deviations from the proportional relationship, as it directly incorporates s_k^2 and is less sensitive to inaccuracies in \hat{p}_k . Another significant characteristic of $\hat{\sigma}_k^{2,\text{alt}}$ is that for large sample sizes n , $\hat{\sigma}_k^{2,\text{alt}} \approx s_k^2$. Finally, it is crucial to emphasize that both variance estimators rely on the assumption that the simplex constraint is the sole source of covariance between groups when assuming a Dirichlet distribution.

To summarize this section, we have proposed a simple analytical approach to estimating π via \hat{p} , along with two variance estimators. This approach offers the advantage of not requiring fitting the reverse Dirichlet-multinomial model.

3.3 Methods for Genetic Mark-Recapture Studies

In GMR studies of mixed-stock fisheries, our goal is to estimate the total escapement N over the migration season. Given that these studies typically span multiple weeks, there is often an interest in estimating the escapement for a specific group k during week t , denoted as $Nw_t\pi_{k,t}$, where w_t represents the proportion of the run occurring in week t , known as the “run weight”, and $\pi_{k,t}$ denotes proportion of the run attributable to group k during week t . Aggregates for weeks or groups can be expressed by summing $Nw_t\pi_{k,t}$ over t and/or k . Since compositional GSI data is collected weekly, all previously introduced notations in Section 3.2 are now appended with an index, $t \in 1, \dots, T$, representing weeks. Let $M = \sum_{t=1}^T \sum_{k=1}^L Nw_t\pi_{k,t}$ denote the total count of lake-type stocks, where L represents the number of lake-type stocks. As is usual practice, we assume that M and w_t are accurately known without error, both obtained from auxiliary surveys (Gazey, 2010).

3.3.1 Bayesian Approaches

We develop a Bayesian methodology to estimate N using the reverse Dirichlet-multinomial model described in Section 3.2.1. The posterior distribution of the escapement for the season, based on the reverse Dirichlet-multinomial likelihood, is derived by sampling from

$$N = \frac{M}{\sum_{t=1}^T w_t \sum_{k=1}^L \pi_{k,t}} \quad (3.7)$$

using Markov Chain Monte Carlo. We take the posterior mean (or median) as our estimate. Escapements per region in each week can then be obtained as $N\pi_{k,t}$ and summed to derive region-specific week-specific escapements. Our approach relies on standard GMR assumptions such as closed population (no entries/deaths/migration of fish between the GSI sample and the lake counts), equal sampling probabilities in the GSI sample within each week, correct recognition of genotypes collected and forward movement (no overcounting fish in multiple weeks). Additionally, we assume that lake-type stocks are abundant enough so that the $\pi_{k,t}$ ’s from our model can be considered population proportions, and not solely super-population parameters.

Prior Specification for π

To complete the Bayesian treatment of our model, we need to assign a prior distribution to π . While a straightforward option is a Dirichlet($\mathbf{1}$) prior, it may not provide sufficient

uninformativeness when applied independently within each week, especially with small GSI sample sizes. Thus, we introduce a time series prior that shares information between weeks and leverages the correlation in $\pi_{k,t}$ across weeks. Our prior is inspired by [Gelman \(1995\)](#)'s transformed normal prior for parameters constrained to sum to 1, originally applied in physiological pharmacokinetic models in [Gelman et al. \(1996\)](#). For a vector (π_1, \dots, π_K) constrained to sum to 1, each π_k is defined as $\pi_k = \frac{e^{Y_k}}{\sum_{i=1}^K e^{Y_i}}$, where each Y_k follows an independent normal distribution, $Y_k \sim N(\theta_k, \psi_k^2)$. We propose a natural AR(1) extension of this formulation as follows:

$$\begin{aligned} \pi_{k,t} &= \frac{e^{Z_{k,t}}}{\sum_{i=1}^K e^{Z_{i,t}}}, \text{ for } t = 1, \dots, T, k = 1, \dots, K \\ Z_{k,t} &= \phi Z_{k,t-1} + \epsilon_{k,t}, \text{ for } t = 2, \dots, T, k = 1, \dots, K \\ \epsilon_{k,t} &\stackrel{iid}{\sim} N(0, (1 - \phi^2)\psi^2), \text{ for } t = 2, \dots, T, k = 1, \dots, K, \end{aligned}$$

with hyperpriors $Z_{k,1} \stackrel{iid}{\sim} N(0, \psi^2)$ and $\phi \sim \text{Unif}(-1, 1)$. Determining an appropriate value for ψ is crucial to ensure that $\pi_{k,t}$ can encompass values across the entire range of $[0, 1]$. We detail this process in [Appendix D](#).

Enhancing Computational Efficiency Using Moment Matching

While the reverse Dirichlet-multinomial model offers a useful approximate description of the data generation process, it can be computationally expensive to implement, as demonstrated in our simulation study in [Section 3.4](#). To enhance computational efficiency without sacrificing statistical performance, we streamline the model to a simple Dirichlet model, using moment-matching to avoid the use of latent \mathbf{X} 's in the model. The model, indexed by week, is:

$$\hat{\mathbf{p}}_t \sim \text{Dirichlet}(\tilde{\lambda}_t \boldsymbol{\pi}_t), \quad t = 1, \dots, T,$$

where $\tilde{\lambda}_t = \tilde{\beta}_t^{-1} - 1$. The choice of parameters for the Dirichlet model, $\tilde{\lambda}_t \boldsymbol{\pi}_t$, ensures that the moments $E(\hat{\mathbf{p}}_t)$ and $\text{Var}(\hat{\mathbf{p}}_t)$ of this Dirichlet model match the moments of the reverse Dirichlet-multinomial model, given as $\boldsymbol{\pi}_t$ and [Equation \(3.6\)](#).

3.3.2 Frequentist Approaches

GMR estimation of the mixed-stock population size may also be conducted via a method-of-moments estimator,

$$\hat{N} = \frac{M}{\sum_t w_t \hat{p}_t^{\text{lake}}},$$

where the denominator serves as an estimate of the proportion of lake-type stock within the mixed-stock population, with $\hat{p}_t^{\text{lake}} = \sum_{k=1}^L \hat{p}_{k,t}$ (Gazey, 2010). The conditional variance of \hat{N} can be approximated using a first-order Taylor series expansion as

$$\text{Var}(\hat{N}|\mathbf{w}) \approx \frac{M^2}{\{\sum_t w_t \text{E}(\hat{p}_t^{\text{lake}}|\mathbf{w})\}^4} \sum_t w_t^2 \text{Var}(\hat{p}_t^{\text{lake}}|\mathbf{w}).$$

It is important to point out that the population-level variance, $\text{Var}(\hat{p}_t^{\text{lake}}|\mathbf{w})$ is $\text{E}(\text{Var}(\hat{p}_t^{\text{lake}}|\mathbf{w}, \mathbf{g})|\mathbf{w}) + \text{Var}(\text{E}(\hat{p}_t^{\text{lake}}|\mathbf{w}, \mathbf{g})|\mathbf{w})$, which is larger than the sample-level variance, $\text{E}(\text{Var}(\hat{p}_t^{\text{lake}}|\mathbf{w}, \mathbf{g})|\mathbf{w})$, estimated with $(s_t^{\text{lake}})^2$, where \mathbf{g} represents all the genetic information from the genetic samples. Hence, we propose the variance estimator

$$\widehat{\text{Var}}(\hat{N}) = \left(\frac{\hat{N}}{\sum_t w_t \hat{p}_t^{\text{lake}}} \right)^2 \sum_t w_t^2 (\hat{\sigma}_t^{\text{lake}})^2,$$

obtained by estimating $\text{E}(\hat{p}_t^{\text{lake}}|\mathbf{w})$ with \hat{p}_t^{lake} and $\text{Var}(\hat{p}_t^{\text{lake}}|\mathbf{w})$ with $(\hat{\sigma}_t^{\text{lake}})^2 = \tilde{\beta}_t \hat{p}_t^{\text{lake}} (1 - \hat{p}_t^{\text{lake}})$. Here, \hat{p}_t^{lake} is an estimate for $\pi_t^{\text{lake}} = \sum_{k=1}^L \pi_{k,t}$.

This estimator represents a clear advancement compared to the one used in Pestal et al. (2020), where $\tilde{\beta}_t$ was defined as $\frac{1}{n_t^{\text{eff}}}$, with effective sample size $n_t^{\text{eff}} = n_t(a + b\hat{p}_t^{\text{lake}})$ for some estimated constants a and b . In their formulation, as the sample size n_t grows, $(\hat{\sigma}_t^{\text{lake}})^2$ decreases towards 0, which is not desirable since it should approach the population-level variance.

An alternative to $(\hat{\sigma}_t^{\text{lake}})^2$ would be to use $(\hat{\sigma}_t^{\text{lake, alt}})^2 = \frac{\tilde{\beta}_t}{\hat{\beta}_t} (s_t^{\text{lake}})^2 = \left(1 + \frac{\lambda_t}{n_t}\right) (s_t^{\text{lake}})^2$ in place of $(\hat{\sigma}_t^{\text{lake}})^2$, leveraging the relationship between means and variances exposed in Section 3.2.2. Here, s_t^{lake} denotes the sample-level standard error in estimating the lake-type proportion. Therefore, $1 + \lambda_t/n_t$ can be thought of as an inflation factor for the sample variance.

The standard error s_t^{lake} may not be directly available in published data, in which case we propose estimating it based on our earlier assumption that the correlation between

regions is solely due to the simplex. Hence, we calculate a pooled standard error for lakes as

$$s_t^{\text{lake}} = \max \left(\sqrt{\sum_{k=1}^L s_{k,t}^2 - 2 \sum_{k=1}^L \sum_{l=k+1}^L \tilde{\beta}_t \hat{\rho}_{k,t} \hat{\rho}_{l,t}}, \sqrt{\sum_{k=1}^L s_{k,t}^2 - 2 \sum_{k=1}^L \sum_{l=k+1}^L s_{k,t} s_{l,t}} \right),$$

where the second argument is the theoretical lower bound under a correlation of -1.

3.4 Simulation Study

To assess and compare the performance of our methods, we conducted a simulation study designed to closely mimic a real-world GMR scenario. We selected parameter values inspired by the Taku River Sockeye Salmon data discussed in Section 3.5. Specifically, we utilized $\hat{\rho}_{k,t}^{\text{obs}}$ and $s_{k,t}^{\text{obs}}$ from the Taku River GSI data as fixed constants to compute λ_t using Equation (3.5). Additionally, we employed $\hat{\rho}_{k,t}^{\text{obs}}$ from the Taku River GSI data as the actual values of $\pi_{k,t}$. To provide a realistic context, we set the true population size N at 60,000 (close to the estimate reported in Pestal et al. (2020)) and calculated $M = 41,326$ using Equation (3.7). We assumed a time span of $T = 12$ weeks and modeled $K = 4$ distinct stocks, among which $L = 2$ were lake-type and the remaining two were river-type, similar to the regional composition of the Taku River datasets. Applying the approximate reverse Dirichlet-multinomial model, we simulated $\hat{\rho}_{k,t}$ using Equation (3.1) and denoted their corresponding standard errors by $s_{k,t}$.

A challenge during simulation was the potential for simulated $\hat{\rho}_{k,t}$ values to reach extremes like 0 or 1, which could lead to errors in subsequent analyses. Additionally, if one regional stock dominates certain weeks in the GSI dataset, $X_{k,t}$ could be 0, causing parameters in the Dirichlet model to be 0 and resulting in errors. To mitigate these issues, we implemented two constraints during the simulation process using rejection sampling, retaining only the simulated datasets that met the conditions: $\hat{\rho}_{k,t} \in [10^{-10}, 1 - 10^{-7}]$ and $\rho_{k,t} > 10^{-10}$. These precautions were crucial for ensuring the validity of our simulated datasets and the reliability of subsequent analyses.

To analyze the data, we employed both the approximate reverse Dirichlet-multinomial model and the moment-matching Dirichlet model through Bayesian and frequentist methods. In Bayesian approaches, we investigated the impact of two types of priors: a diffuse prior, specifically a Dirichlet(1) prior on $\pi_{k,t}$, and the AR(1) prior described in Section 3.3.1, to compare their performance. For each simulated dataset, we recalculated λ_t to derive $\tilde{\lambda}_t$

based on the simulated $\hat{p}_{k,t}$ and $s_{k,t}$. Since our primary interest is in the posterior distribution of the escapement N , we obtained samples of N at each iteration of the MCMC process using Equation (3.7), based on the posterior samples of $\pi_{k,t}$ from fitting our models.

For each simulation study setting, we generated and analyzed 1,000 datasets using R (Version 4.3.1; R Core Team, 2024), ensuring reproducibility with a fixed seed. Bayesian methods were implemented using JAGS (Version 4.3.2; Plummer et al., 2003), where we ran three chains until the second half of the chains showed convergence, confirmed by a Gelman-Rubin convergence diagnostic $\hat{R} < 1.1$. Furthermore, the chains were thinned to 10,000 iterations to reduce storage space. For the frequentist methods, we applied three variance estimators: $(\hat{\sigma}_t^{\text{lake}})^2$ calculated as $\tilde{\beta}_t \hat{p}_t^{\text{lake}} (1 - \hat{p}_t^{\text{lake}})$, $(\hat{\sigma}_t^{\text{lake, alt}})^2$, and $(s_t^{\text{lake}})^2$. The last one is a naive variance estimator that does not propagate uncertainty from the sample level to the population level. To systematically evaluate our approach, we computed several metrics on N for each simulation study. These included the relative Monte Carlo bias of the estimates (RBias), the relative root mean squared error of the estimates (RRMSE), as well as the average length (LCI), and coverage probability (CP) of the 95% quantile-based credible interval for Bayesian methods or $\hat{N} \pm 1.96 \cdot \sqrt{\widehat{\text{Var}}(\hat{N})}$ for the method-of-moments in frequentist approaches.

3.4.1 Results

Table 3.1 presents a comprehensive overview of the estimated escapement across the simulation settings. The first two rows provide posterior summary statistics of N under different priors. Meanwhile, the bottom rows include results from the method-of-moments estimator with different variance estimators as discussed in Section 3.3.2, for comparison with the proposed models. Additionally, we show the average computing time for one dataset for each method in the table to compare their computational efficiency.

The results demonstrate that the choice of prior significantly influences the performance of the proposed models. Specifically, using the AR(1) prior yields smaller relative bias and relative RMSE, along with a coverage probability closer to 95%, compared to the Dirichlet prior. This can be attributed to the weekly regional proportions in the Taku dataset used in our simulation study, as shown in Figure 3.3a. The regional proportions vary significantly across weeks and exhibit temporal trends in at least three regions. In such scenarios, a diffuse Dirichlet prior may not adequately capture the underlying dynamics, hence affecting the model’s performance.

Additionally, the comparison between different Bayesian inference models indicates that moment-matching Dirichlet model with the AR(1) prior demonstrates superior per-

Table 3.1: Comparison of estimation methods using results from the simulation study. “ARDM” denotes the approximate reverse Dirichlet-multinomial model, “MMD” represents the moment-matching Dirichlet model, “MoM” refers to the method-of-moments estimator with $(\hat{\sigma}_t^{\text{lake}})^2 = \tilde{\beta}_t \hat{p}_t^{\text{lake}}(1 - \hat{p}_t^{\text{lake}})$, “MoM(Alt)” refers to the method-of-moments estimator using $(\hat{\sigma}_t^{\text{lake, alt}})^2$, and “MoM(Naive)” refers to the method-of-moments estimator using $(s_t^{\text{lake}})^2$ instead of $(\hat{\sigma}_t^{\text{lake}})^2$.

Model	Prior	RBias	RRMSE	CP	LCI	Time/Dataset
ARDM	Dir	0.026	0.039	0.88	7121	33.7 secs
ARDM	AR(1)	0.010	0.032	0.95	7182	44.6 secs
MMD	Dir	0.015	0.032	0.95	7056	0.3 secs
MMD	AR(1)	0.005	0.031	0.95	7172	2.9 secs
MoM	N/A	-0.001	0.030	0.94	6987	0.001 secs
MoM(Alt)	N/A	-0.001	0.030	0.96	8046	0.001 secs
MoM(Naive)	N/A	-0.001	0.030	0.76	4174	0.002 secs

formance in terms of RBias, RRMSE, and CP. Moreover, the moment-matching Dirichlet model significantly reduces computation time, achieving over a 15-fold reduction compared to the reverse Dirichlet-multinomial model with the AR(1) prior, and approximately a 100-fold reduction compared to the same model with the Dirichlet prior, while maintaining comparable statistical performance.

Furthermore, the estimates provided by the frequentist method exhibit the smallest relative bias, underscoring the unbiased nature of this approach. However, there is no significant difference in the relative RMSEs compared to the Bayesian methods. Most notably, the coverage probability of the method-of-moments estimator with the naive variance estimator (MoM(Naive)) is much lower than the nominal 95% than the other estimators, highlighting the inadequacy of directly using the sample-level variance as the population-level variance. This finding emphasizes the importance of correctly estimating variance to achieve reliable coverage probabilities, as discussed in Sections 3.1 and 3.3.

Interestingly, the average lengths of the 95% credible intervals from the Bayesian methods fall between those of the MoM and MoM(Alt) methods. This suggests that the Bayesian methods provide a balanced estimation of uncertainty compared to the frequentist estimators, although the differences are subtle and may be attributed to Monte Carlo error.

We further explore the ability of our Bayesian models to estimate the proportions $\pi_{k,t}$. Figures 3.1, and E.3, E.2, E.1 in Appendix E present results for MMD with AR(1) prior,

MMD with Dirichlet prior, ARDM with AR(1) prior, and ARDM with Dirichlet prior, respectively. Model MMD generally performs better than ARDM and the AR(1) prior outperforms the Dirichlet prior, which is consistent with the results above.

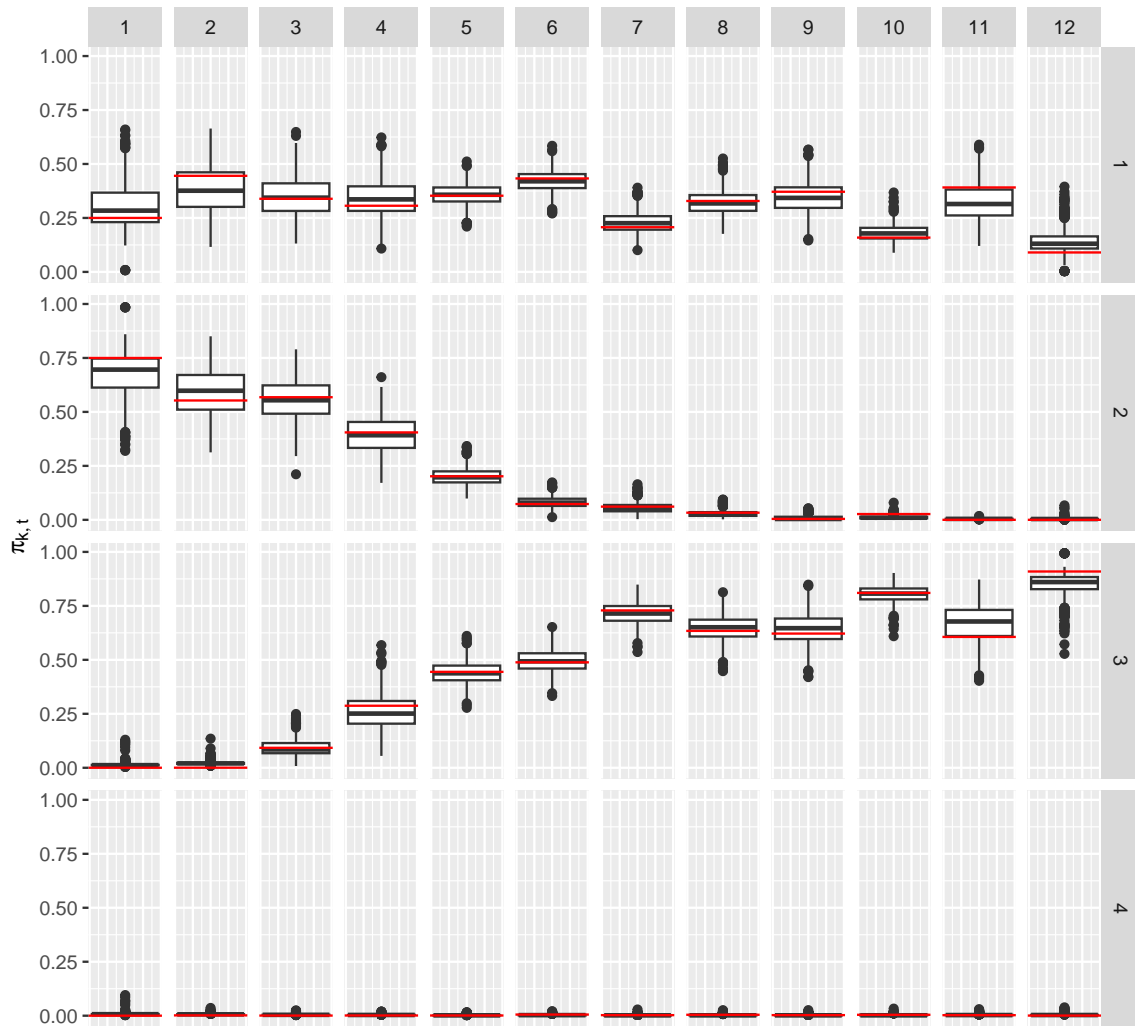


Figure 3.1: Distribution of the posterior mean for $\pi_{k,t}$ from the moment-matching Dirichlet model with AR(1) prior in the simulation study, for stocks $k = 1$ to 4 in weeks $t = 1$ to 12. Red horizontal lines indicate the true $\pi_{k,t}$ values.

3.5 Application to the Taku River Salmon Run Estimation

The Taku River is a river system originating from the Stikine Plateau in northwestern British Columbia, Canada, flowing into Juneau, Alaska. Renowned for its productivity, the Taku River hosts one of the largest runs of Sockeye Salmon in Southeast Alaska and northern British Columbia. Figure 3.2 provides an overview map of the Taku River system and the study settings.

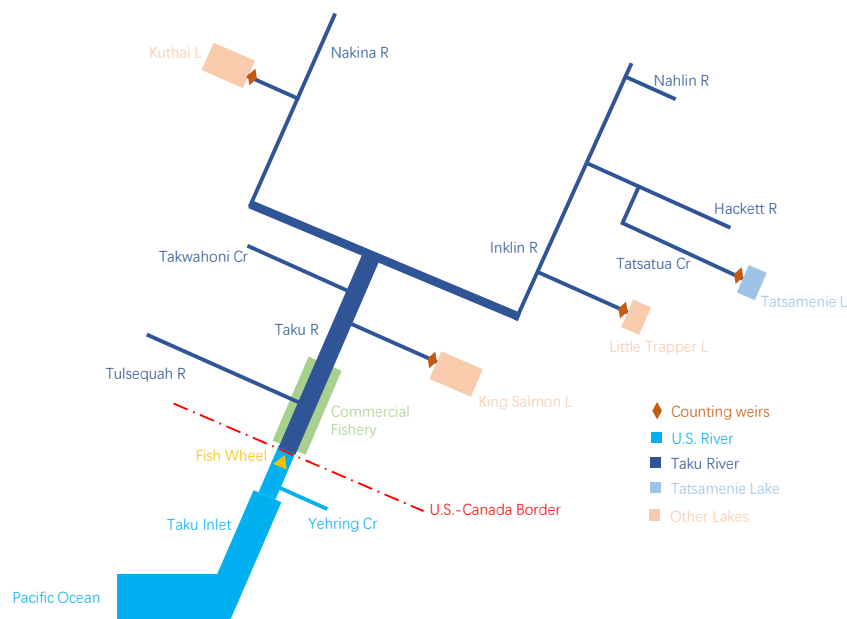


Figure 3.2: Overview of Taku River system and study settings. R denotes river, Cr denotes creek and L denotes lake.

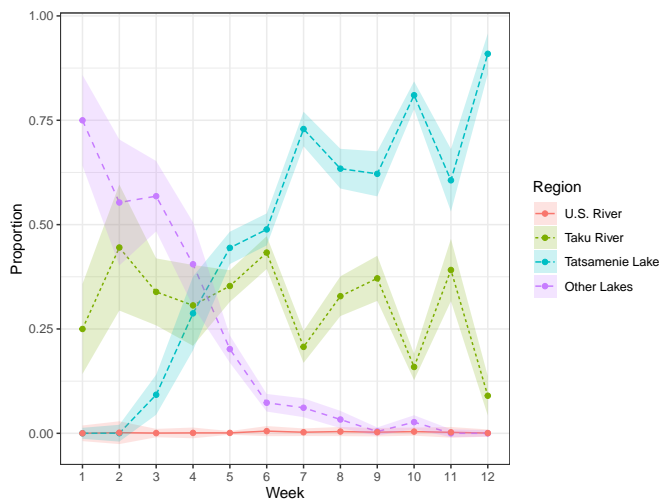
Genetic data on Sockeye Salmon in the Taku River has been collected since 2008 by the Canadian commercial fishery and annually analyzed by the Molecular Genetics Laboratory of Fisheries and Oceans Canada (DFO). Starting in 2012, the U.S. District 111 Commercial Fishery also began collecting genetic data annually, with analysis conducted by the Gene Conservation Laboratory of the Alaska Department of Fish and Game (ADF&G). These collaborative efforts have led to the development of a comprehensive genetic baseline for Taku River Sockeye Salmon, comprising 17 unique genetic groups, including 13 river-type and 4 lake-type stocks. Given that many of these stocks have very small proportions, often close to zero, there is a high likelihood that some stocks may not be included in a given sample. To mitigate this issue and ensure more reliable estimates, these stocks are further grouped into four regions/reporting units based on their type and geographic location: U.S. River, Taku River, Tatsamenie Lake, and Other Lakes, as illustrated in Figure 3.2.

We arbitrarily focus on analysing data from the year 2017. Table 3.2 provides GSI sample sizes n_t , while Figure 3.3a presents GSI summary data from weekly samples at the regional level. These summaries were derived using the Bayesian GSI algorithm proposed by Pella and Masuda (2001). The algorithm generates, for each fish in the sample, individual assignments (IA) to stocks via Markov Chain Monte Carlo. These IAs are then summarized by averaging over individuals to obtain composition estimates $\hat{p}_{k,t}$ along with their sample-level standard errors $s_{k,t}$. Figure 3.3b displays the relationship between $s_{k,t}^2$ and $\hat{p}_{k,t}(1 - \hat{p}_{k,t})$ in each week $t = 1, \dots, 12$. We observe that the weekly relationships exhibit a close proportionality, which is a crucial prerequisite for our Dirichlet assumption. Furthermore, we include the variance inflation factor $1 + \lambda_t/n_t$, discussed in Section 3.2 and Section 3.3.2, in Table 3.2. We can see that the effect of this variance inflation factor remains relatively constant across weeks, reaching a maximum when the GSI sample size is 112 in week 7. However, if n_t were very large, we would expect this inflation factor to diminish towards 1.

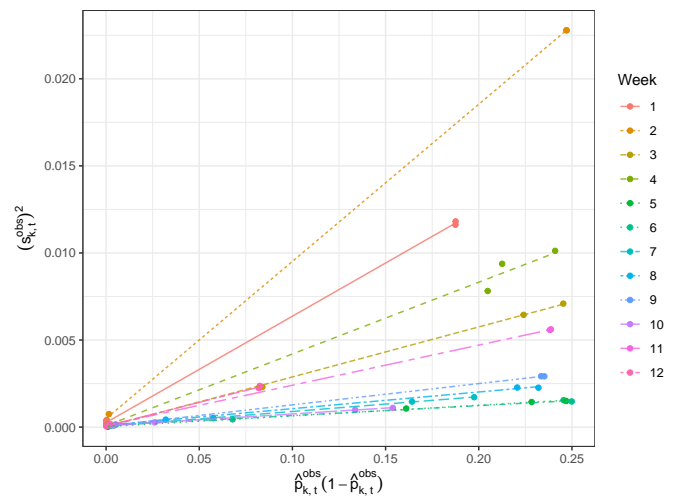
Table 3.2: 2017 GSI data: weekly weights, sample sizes and variance inflation factor

Week (t)	1	2	3	4	5	6	7	8	9	10	11	12
w_t	0.02	0.01	0.04	0.03	0.18	0.18	0.11	0.11	0.09	0.16	0.04	0.04
n_t	17	13	38	26	172	178	112	105	84	146	43	42
$1 + \frac{\lambda_t(\mathbf{s})}{n_t}$	1.88	1.76	1.89	1.89	1.92	1.91	2.01	1.94	1.94	1.93	1.97	1.83

In addition to genetic data collection, fish wheels are set up close to the U.S.-Canada border, providing weekly weights displayed in Table 3.2. Finally, counting weirs are constructed at the entrances of the lakes within the river system. Considering the salmon's predisposition to return to their native lakes and rivers, these counting weirs can nearly



(a) Weekly proportions ($\hat{p}_{k,t}^{\text{obs}}$) of genetic samples from each region k . Each dot in the scatter plot represents a mean proportion while shaded bands extend one standard error ($s_{k,t}^{\text{obs}}$) above and below the mean.



(b) Weekly variances ($s_{k,t}^{\text{obs}})^2$ as a function of $\hat{p}_{k,t}^{\text{obs}}(1 - \hat{p}_{k,t}^{\text{obs}})$, with regression lines for each week.

Figure 3.3: Two graphs displaying $\hat{p}_{k,t}$ and $s_{k,t}$, over weeks and regions, for the 2017 GSI dataset.

perfectly tally the fish entering the lakes, offering a comprehensive count $M = 34,351$ for lake-type stocks.

We analyzed the 2017 Taku River Sockeye Salmon data in multiple ways, comparing all the approaches implemented in the simulation study. The results on total escapement estimation are presented in Table 3.3. The results across different settings are close to each other. Nonetheless, several noteworthy observations can be made from the table. Firstly, the posterior mean estimates with the AR(1) prior are closer to the method-of-moments estimate compared to those with Dirichlet prior. Secondly, the proposed models using the Dirichlet prior yield a smaller posterior standard deviation compared to those with the AR(1) prior, while the frequentist method with the naive variance estimator has the smallest standard deviation. Moreover, compared to the naive variance estimator, both variance estimators proposed in Section 3.3.2 provide a standard deviation closer to the posterior standard deviations of the Bayesian approaches.

Table 3.3: Results for the Taku River Sockeye Salmon application. “Estimate” for the Bayesian model refers to the posterior mean of N . “SD” denotes the posterior standard deviation for the Bayesian methods, and the estimated standard deviation for frequentist methods with different variance estimators. The 95% CI shows the 95% quantile-based credible intervals for the Bayesian approaches and $\hat{N} \pm 1.96 \cdot \sqrt{\widehat{\text{Var}}(\hat{N}|\mathbf{w})}$ for the frequentist methods. “Time” represents the computing time for each setting.

Model	Prior on $\boldsymbol{\pi}$	Estimate	SD	95% CI	Time
ARDM	Dirichlet	51,164	1,480	(48,439, 54,265)	97 secs
ARDM	AR(1)	50,367	1,482	(47,655, 53,455)	427 secs
MMD	Dirichlet	50,956	1,526	(48,147, 54,116)	21 secs
MMD	AR(1)	50,354	1,551	(47,490, 53,536)	104 secs
MoM	N/A	49,873	1,498	(46,937, 52,809)	0.02 secs
MoM(Alt)	N/A	49,873	1,685	(46,570, 53,175)	0.03 secs
MoM(Naive)	N/A	49,873	876	(48,156, 51,589)	0.05 secs

3.6 Discussion

In this study, we developed a novel methodology to estimate population-level compositions from sample-level data. We introduced a reverse Dirichlet-multinomial model and applied moment-matching techniques to provide accurate estimates while addressing computational

challenges. We further extended our methods to estimate the escapement of mixed-stock fisheries in GMR settings. Through a comprehensive simulation study and a real-data analysis based on the Taku River Sockeye Salmon fishery data, we compared our Bayesian approaches with the frequentist method-of-moments estimator.

While our Bayesian models exhibited reliable performance, yielding accurate estimates of escapement and providing insights into the variability of stock proportions over time, they come with limitations. Firstly, the approximate reverse Dirichlet-multinomial model could demand substantial computational time, particularly for large datasets. Secondly, a key assumption of our method is the low dependency setting provided by the Dirichlet distribution. Therefore, our method may not be suitable for applications in scenarios where stocks are negatively correlated due to having genetic compositions that are challenging to distinguish.

Nonetheless, our proposed methodologies extend beyond salmon stocks as they could be applied to other contexts where compositional data are summarized by their central tendency and variability. Future research may explore methods that accommodate greater interdependence between groups, leveraging correlation data to enhance the accuracy and robustness of estimations in diverse contexts.

Chapter 4

Parametric Quantile Estimation of Posterior Quantiles for Markov Chain Monte Carlo

4.1 Introduction

Markov Chain Monte Carlo (MCMC) methods are fundamental tools in Bayesian inference, particularly for sampling from a posterior distribution. These techniques produce auto-correlated samples that are treated as draws from a posterior distribution, enabling inference about model parameters. Software such as JAGS (Plummer et al., 2003) and Stan (Stan Development Team, 2024) are widely used by practitioners in fields like ecology and public health due to their automated implementation of MCMC algorithms for a wide range of models. After collecting MCMC samples, the posterior distribution is usually summarized using marginal quantities such as the posterior mean, the posterior standard deviation, and quantiles of credible intervals. For instance, in R, these quantities are automatically reported by popular packages such as `coda` (Plummer et al., 2006), `MCMCvis` (Youngflesh, 2018) and `stansummary` (Gabry et al., 2023).

Gelman et al. (2013) noted that approximately 100 independent samples are often adequate for estimating marginal posterior means. However, when it comes to estimating Bayesian credible intervals, Kruschke (2014) argued that around 10,000 independent (effective) samples are required to accurately estimate quantiles of marginal posterior distributions, assuming that the quantile is estimated using the empirical quantile (EQ) of the MCMC draws (Casella and Berger, 2002; Gelman et al., 2013; Liu et al., 2016; Turkman

et al., 2019). EQs, which are standard practice, offer the advantage of being easy to compute (Chen and Shao, 1999; Rice and Ye, 2022) and are provided as standard outputs from the aforementioned R packages. However, EQs computed from a small effective sample size can exhibit high uncertainty, as they rely heavily on accurate information about the tails of the posterior distribution. While one typical strategy for practitioners is to extend the MCMC chains to obtain a larger effective sample size, this approach may require more time than practical considerations allow, particularly when the built-in MCMC samplers of standard software are inefficient or in the context of Bayesian simulation studies (Kelter, 2024).

An alternative method for estimating the quantiles of a credible interval with a confidence level of $1 - \alpha$ involves using a parametric quantile (PQ) estimator. This approach approximates the posterior distribution with a parametric distribution. By treating the samples as if they are independently and identically distributed, we can fit this parametric model to the data and then utilize its quantile function to determine the desired quantiles. In this work, we specifically focus on using the normal distribution to approximate posterior distributions for two primary reasons. First, the normal distribution provides a simple and efficient form for the PQ estimator. When the posterior distribution is assumed to be normal, the PQ estimator derived from MCMC chains of length n is $\bar{X}_n \pm \Phi^{-1}(1 - \alpha/2)S_n$, where \bar{X}_n is the MCMC sample mean, S_n is the MCMC sample standard deviation, and Φ^{-1} is the quantile function of a standard normal distribution. This simplicity in both calculation and interpretation makes the normal distribution an appealing choice for PQ estimation. Second, the Bernstein-von Mises Theorem (Vaart, 1998) provides a theoretical foundation for the use of normal approximations. According to this theorem, under certain conditions, the posterior distribution of a parameter tends to a normal distribution as the sample size increases. This asymptotic normality is observed when there is a large amount of data and the influence of prior information on the posterior distribution is minimal. Such conditions are frequently met in practical applications, such as regression models, where the estimated regression coefficients often follow a normal distribution. Consequently, the normal approximation is not only theoretically justified but also practically relevant.

Though the PQ estimator could be biased when the posterior distribution deviates from normality, we postulate that it should be more precise than the EQ estimator, as it relies less on information from the tails, which are harder to sample. In contrast, the posterior mean and standard deviation, which form the EQ estimator, require fewer independent samples to estimate. Therefore, when the assumption of normality holds or is approximately valid, the PQ estimator may be preferable for estimating parameters from the posterior distribution.

In this study, our primary goal is to evaluate the performance of the PQ estimator in estimating posterior quantiles using MCMC samples. We begin by deriving the asymptotic

properties of the PQ estimator of sample quantiles for a normal posterior distribution and compare it with the EQ estimator in Section 4.2. Subsequently, in Section 4.3, we conduct simulation studies to evaluate and compare the performance of the PQ estimator of quantiles and the EQ estimator under three different scenarios. Additionally, in Section 4.4, we demonstrate the practical application of the PQ estimator by replicating the Bayesian analysis of a Leisler’s bat capture-recapture dataset described in Kéry and Schaub (2011), employing a resource-intensive data augmentation approach. A discussion of our findings and directions for further work are presented in Section 4.5. Our findings shed light on the behavior of these estimators in the early stages of MCMC, offering practitioners an avenue to extract useful information when computational resources are constrained.

4.2 Methodology

In this section, we delve into the theoretical foundation of the PQ estimator for quantiles in the context of MCMC sampling. By understanding the asymptotic properties of the PQ estimator, we can better evaluate the performance and reliability of the PQ estimator compared to the EQ estimator.

4.2.1 Asymptotic Properties of the PQ Estimator

Suppose $\{X_n : n \geq 0\}$ is an irreducible, reversible Markov chain of length n with a stationary distribution π , representing the target posterior distribution in our context. We define the PQ estimator for the p^{th} sample quantile as

$$\hat{\Gamma}_{p,n}^{\text{PQ}} = \bar{X}_n + \Phi^{-1}(p)S_n,$$

where \bar{X}_n is the sample mean and S_n is the sample standard deviation.

Noting that $\hat{\Gamma}_{p,n}^{\text{PQ}}$ can be represented as a function of the first sample moment \bar{X}_n and the second sample moment $\overline{X_n^2}$, we establish the following multivariate Markov Chain Central Limit Theorem (CLT):

$$\sqrt{n} \left\{ \begin{pmatrix} \bar{X}_n \\ \overline{X_n^2} \end{pmatrix} - \begin{pmatrix} \mu_X \\ \mu_X^2 + \sigma_X^2 \end{pmatrix} \right\} \xrightarrow{d} N(\mathbf{0}, \Sigma_{\text{PQ}}),$$

where μ_X and σ_X^2 denote the true posterior mean and variance, respectively, and

$$\Sigma_{\text{PQ},11} = \text{Var}(X_i) + 2 \sum_{k=1}^{\infty} \text{Cov}(X_i, X_{i+k}) = \sigma_X^2 \left(1 + 2 \sum_{t=1}^{\infty} \rho_t \right), \quad (4.1)$$

$$\Sigma_{\text{PQ},22} = \text{Var}(X_i^2) + 2 \sum_{k=1}^{\infty} \text{Cov}(X_i^2, X_{i+k}^2), \quad (4.2)$$

$$\Sigma_{\text{PQ},12} = \Sigma_{\text{PQ},21} = \text{Cov}(X_i, X_i^2) + 2 \sum_{k=1}^{\infty} \text{Cov}(X_i, X_{i+k}^2), \quad (4.3)$$

with ρ_t denoting the lag- t autocorrelation of the Markov chain $\{X_n\}$.

By applying the multivariate delta method, we have the following CLT for the PQ estimator:

$$\sqrt{n} \left\{ \hat{\Gamma}_{p,n}^{\text{PQ}} - (\mu_X + \Phi^{-1}(p)\sigma_X) \right\} \xrightarrow{d} N(0, \sigma_{\text{PQ}}^2),$$

where

$$\sigma_{\text{PQ}}^2 = \left(1 - \frac{\Phi^{-1}(p)\mu_X}{\sigma_X} \right)^2 \Sigma_{\text{PQ},11} + \left(\frac{\Phi^{-1}(p)}{2\sigma_X} \right)^2 \Sigma_{\text{PQ},22} + \frac{\Phi^{-1}(p)}{\sigma_X} \left(1 - \frac{\Phi^{-1}(p)\mu_X}{\sigma_X} \right) \Sigma_{\text{PQ},12}. \quad (4.4)$$

A proof of this result is given in Appendix F. Additionally, based on the strong law of large numbers (SLLN) for Markov chains (Meyn and Tweedie, 2009), $\hat{\Gamma}_{p,n}^{\text{PQ}} \xrightarrow{a.s.} \mu_X + \Phi^{-1}(p)\sigma_X$ as $n \rightarrow \infty$. If the posterior distribution is normal, then $\mu_X + \Phi^{-1}(p)\sigma_X$ corresponds to the quantile of interest. However, it is important to note that if the posterior distribution is not normal, the difference between $\mu_X + \Phi^{-1}(p)\sigma_X$ and the true quantile value can create considerable bias. Therefore, the PQ method should only be used when the posterior distribution is approximately or exactly known.

4.2.2 Comparative Evaluation under AR(1)

As an analytically tractable example, consider an AR(1) process parameterized by ρ , with stationary distribution $N(0, \sigma^2)$. This process is defined as follows:

$$X_n = \rho X_{n-1} + \epsilon_n, \quad \epsilon_n \sim N(0, (1 - \rho^2)\sigma^2).$$

Under such an AR(1) process, we have

$$\begin{aligned}\Sigma_{\text{PQ},11} &= \sigma^2 \left(1 + 2 \sum_{t=1}^{\infty} \rho^t \right) = \sigma^2 \frac{1+\rho}{1-\rho}, \\ \Sigma_{\text{PQ},22} &= [\text{E}(X_1^4) - \text{E}(X_1^2)^2] \left(1 + 2 \sum_{t=1}^{\infty} \rho^{2t} \right) = 2\sigma^4 \frac{1+\rho^2}{1-\rho^2}, \\ \Sigma_{\text{PQ},12} &= \Sigma_{\text{PQ},21} = 0,\end{aligned}$$

and the value of σ_{PQ}^2 in the CLT becomes $\sigma_{\text{PQ}}^2 = \sigma^2 \delta_{\text{PQ}}(p, \rho)$ with $\delta_{\text{PQ}}(p, \rho) = \frac{1+\rho}{1-\rho} + \frac{[\Phi^{-1}(p)]^2}{2} \frac{1+\rho^2}{1-\rho^2}$.

To establish a comparison with the asymptotic behavior of the EQ estimator under AR(1), we consider the EQ Central Limit Theorem derived by [Doss et al. \(2014\)](#). The CLT for the EQ estimator of the p th quantile with n samples, $\hat{\Gamma}_{p,n}^{\text{EQ}}$, is

$$\sqrt{n}(\hat{\Gamma}_{p,n}^{\text{EQ}} - \gamma_p) \xrightarrow{d} N(0, \eta(\gamma_p)/[f_{\pi}(\gamma_p)]^2),$$

where γ_p is the true value of the p^{th} quantile, f_{π} is the probability density function of π , and

$$\eta(y) := \text{Var}\{I(X_0 \leq y)\} + 2 \sum_{k=1}^{\infty} \text{Cov}\{I(X_0 \leq y), I(X_k \leq y)\}.$$

Under AR(1), the CLT variance simplifies to $\sigma^2 \delta_{\text{EQ}}(p, \rho)$, with

$$\delta_{\text{EQ}}(p, \rho) = \frac{p(1-p) + 2 \sum_{k=1}^{\infty} [F_{0,k}(\gamma_p/\sigma, \gamma_p/\sigma) - p^2]}{[\phi(\gamma_p/\sigma)]^2},$$

where ϕ is the density of the standard normal distribution and $F_{0,k}$ is the CDF of a bivariate normal distribution with a compound symmetric variance-covariance structure with 1 on the diagonal and ρ^k on the off-diagonal. It is important to note that δ_{EQ} does not depend on σ because the distribution of X_n/σ does not depend on σ . Moreover, the infinite sum can be accurately approximated from a large finite number of terms.

In [Table 4.1](#) we compare the PQ CLT variance with the EQ CLT variance under AR(1) for various confidence levels. Our results indicate that the PQ estimator consistently has a lower variance compared to the EQ estimator, as the ratio $\delta_{\text{PQ}}/\delta_{\text{EQ}}$ is consistently below 1. Alternatively, our findings suggest that the PQ estimator can achieve comparable accuracy to EQ with fewer iterations, making it a more efficient choice for quantile estimation in

MCMC. Not surprisingly, the improvement of PQ over EQ is more noticeable when the confidence level is high, i.e. the quantile to estimate is further in the tails.

The improvement provided by the PQ estimator is particularly significant when autocorrelation is low. This finding is highly relevant in scenarios where autocorrelation is initially very high, which is often the case when using automated MCMC software. High autocorrelation in MCMC chains typically requires a greater number of iterations to achieve the desired accuracy. To address this, practitioners frequently thin the chains, thereby diminishing the autocorrelation. As a result, the lower autocorrelation observed after thinning actually represents some of the most challenging MCMC sampling problems. In these scenarios, the PQ estimator is especially beneficial, providing accurate quantile estimates with fewer iterations.

Table 4.1: Comparison of PQ and EQ CLT variances under AR(1) for different confidence levels and autocorrelation parameters. The ratio $\delta_{\text{PQ}}/\delta_{\text{EQ}}$ indicates the relative efficiency of PQ compared to EQ.

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\delta_{\text{PQ}}(0.95, \rho)$	2.35	2.60	2.97	3.48	4.20	5.25	6.87	9.62	15.16	31.89
$\delta_{\text{PQ}}(0.975, \rho)$	2.92	3.18	3.58	4.16	4.99	6.20	8.08	11.28	17.75	37.30
$\delta_{\text{PQ}}(0.995, \rho)$	4.32	4.61	5.09	5.83	6.91	8.53	11.05	15.36	24.11	50.60
$\frac{\delta_{\text{PQ}}(0.95, \rho)}{\delta_{\text{EQ}}(0.95, \rho)}$	0.53	0.55	0.58	0.62	0.66	0.70	0.74	0.77	0.81	0.83
$\frac{\delta_{\text{PQ}}(0.975, \rho)}{\delta_{\text{EQ}}(0.975, \rho)}$	0.41	0.43	0.46	0.49	0.53	0.57	0.61	0.65	0.69	0.72
$\frac{\delta_{\text{PQ}}(0.995, \rho)}{\delta_{\text{EQ}}(0.995, \rho)}$	0.18	0.19	0.21	0.23	0.25	0.28	0.32	0.35	0.39	0.42

Finally, it is desirable to also be able to estimate the efficiency of the PQ estimator. Given that we have established a CLT for PQ, it is possible to estimate the Monte Carlo standard error (MCSE) $\sqrt{\sigma_{\text{PQ}}^2/n}$ by plugging in an estimate of Equation (4.4). We estimate Σ_{PQ} using multivariate batch means methods (Vats et al., 2019), which are implemented in the R package `mcmcse` (Flegal et al., 2021), and we estimate μ_X and σ_X using the sample mean and sample standard deviation, respectively. To compare with the EQ estimator, we can estimate the MCSE of the EQ estimator using the batch means method for quantiles (Doss et al., 2014), also implemented in the `mcmcse` package. Figure 4.1 illustrates the simulated performance of the MCSE estimator for PQ and EQ under an AR(1) process with a $N(0,1)$ stationary distribution. Additionally, the 95% Highest Density Interval (HDI) constructed from 1,000 estimated MCSEs for the PQ estimator is narrower than that of the EQ estimator at any number of iterations, demonstrating that the PQ MCSE estimator is more precise and exhibits less variability.

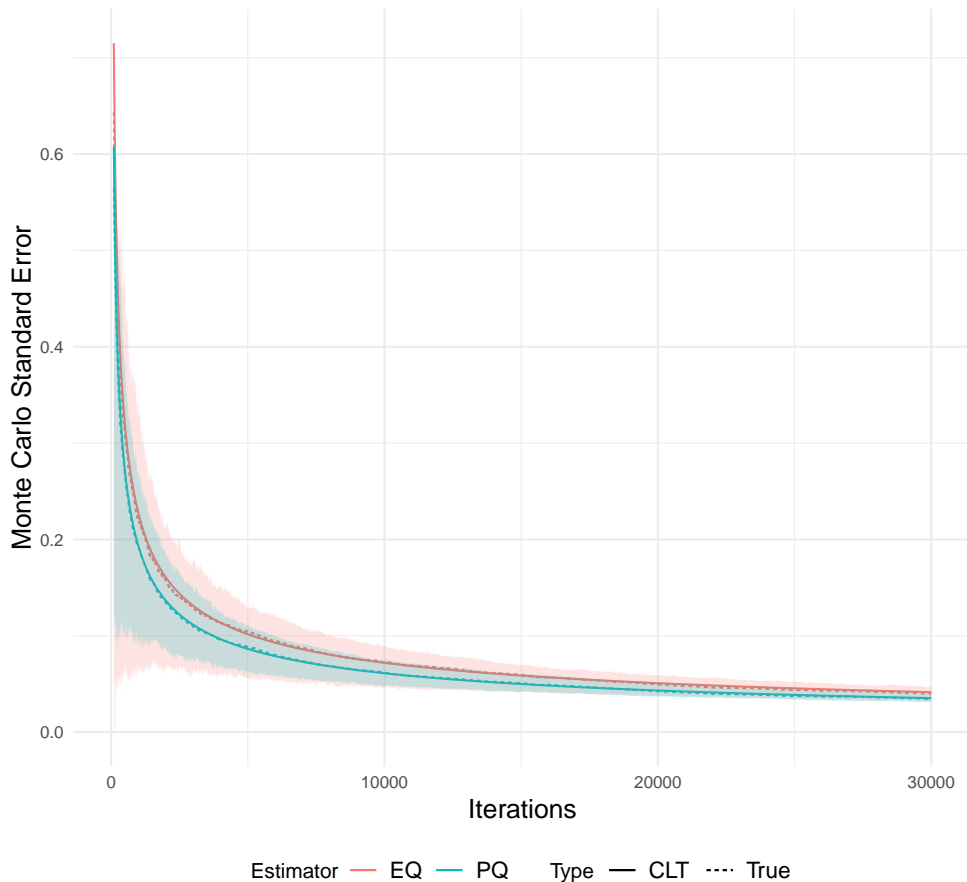


Figure 4.1: Monte Carlo standard error (MCSE) of PQ and EQ estimators under AR(1) with stationary distribution $N(0,1)$. “CLT” represents the MCSE computed from the theoretical CLT variance. “True” represents the true MCSE approximated via simulation. The shaded regions depict 95% HDI intervals for the estimated MCSE, obtained via simulation.

4.3 Simulation Study

We conducted a simulation study across three different scenarios to comparatively assess the performance of the PQ estimator and the EQ estimator. Suppose we are interested in estimating the 95% credible interval of a posterior distribution, or equivalently, the stationary distribution of an MCMC. For convenience, we present results only for the upper bound of the credible interval, which corresponds to the 0.975 quantile of the posterior distribution. To represent different scenarios, we consider three values of autocorrelation

ρ : 0.1, 0.5, and 0.9. For each value of ρ , we simulated 100 datasets from an AR(1), each with 5,000 iterations, to assess the performance of the two types of estimators.

4.3.1 Quantile Estimation under AR(1)

We begin by examining a basic AR(1) process with stationary distribution $N(0,1)$. The first row of Figure 4.2 presents the expected value of the 0.975 quantile estimate under the PQ and EQ methods, as a function of the number of iterations of the AR(1) and different levels of autocorrelation, ρ . When $\rho = 0.1$ and 0.5, the bias in the PQ estimator is negligible after about 500 iterations, while the EQ method requires over 3,000 iterations to achieve comparable accuracy. When $\rho = 0.9$, both methods demonstrate slower convergence.

To further complete the comparison of the estimators, the first row of Figure 4.3 shows the root mean squared error (RMSE) of the quantile estimators. The PQ method consistently exhibits a smaller RMSE compared to the EQ method, irrespective of the number of iterations. Notably, larger differences in RMSE are observed at lower autocorrelations.

4.3.2 Quantile Estimation under Heavier Tails

In practice, the stationary distribution may deviate from the normal distribution. To represent mild non-normality, we consider Student’s t distribution with 10 degrees of freedom for the stationary distribution. This distribution is known for its symmetric bell-shaped curve but with heavier tails compared to a normal distribution. Samples from the AR(1) setting obtained in Section 4.3.1 are transformed to the t_{10} distribution via quantile matching.

The second row of Figure 4.2 shows the expected estimates for the 0.975 quantile using both methods. We observe that the PQ estimates quickly converge to the 0.975 quantile of the approximate normal distribution, rather than the true quantile value, whereas the EQ estimator, though slower, converges toward the true value. Despite the bias in the PQ estimates, Figure 4.3 (second row) shows that the RMSEs for PQ are still smaller than those for EQ. This suggests that the variability of the EQ estimator is significantly greater than that of the PQ estimator, and that the PQ estimator is relevant despite its bias.

4.3.3 Quantile Estimation for a Skewed Posterior

As a final scenario, we consider the case of a $Gamma(2,1)$ stationary distribution, characterized by significant right-skewness and substantial non-normality. Using quantile matching, we transform the simulated samples from the AR(1) setting to conform to this gamma

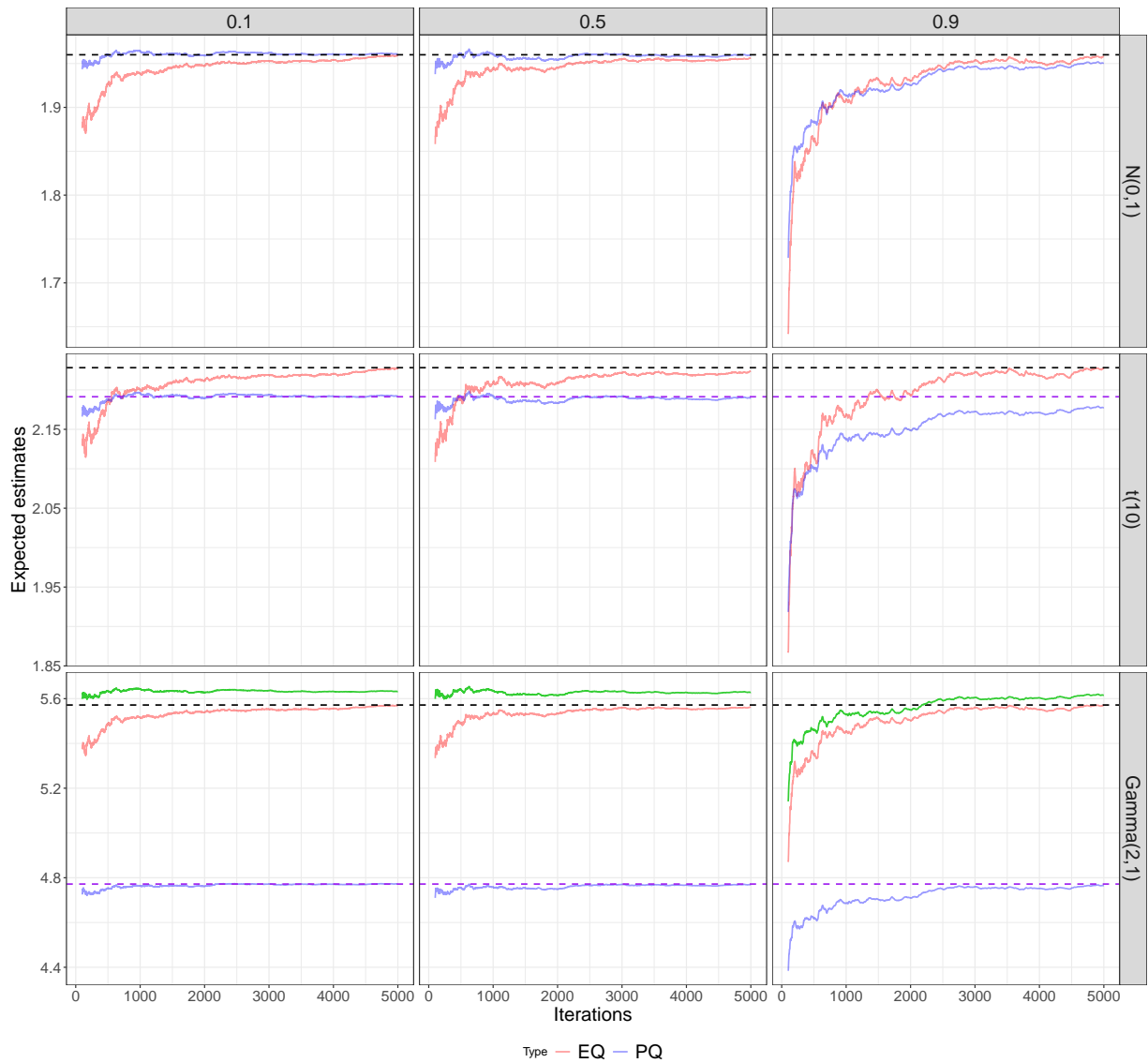


Figure 4.2: Expected estimates of the 0.975 quantile obtained from 100 AR(1) simulations with stationary distributions (rows) $N(0,1)$, t_{10} , or $\text{Gamma}(2,1)$ and autocorrelation parameter ρ (columns) of 0.1, 0.5, or 0.9. To reduce the influence of initial values, the first 100 iterations are omitted. The black dashed line indicates the true 0.975 quantile for each distribution, while the purple dashed line shows the 0.975 quantile of an approximate normal distribution with the same mean and variance as the corresponding distribution. The green line for $\text{Gamma}(2,1)$ represents the PQ estimates from Box-Cox transformed samples.

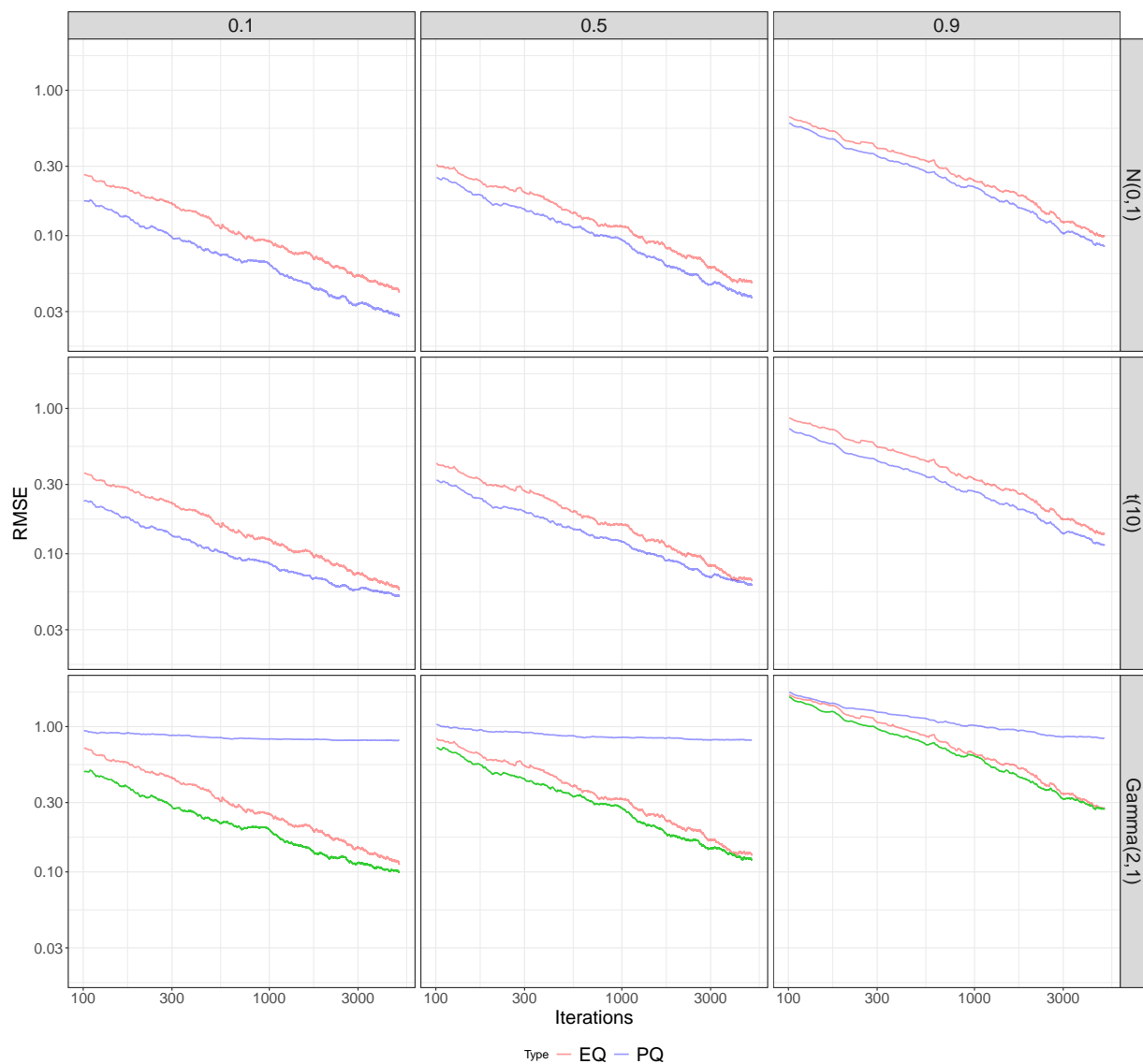


Figure 4.3: Root mean squared error (RMSE) of the 0.975 quantile obtained from 100 AR(1) simulations with stationary distributions (rows) $N(0,1)$, t_{10} , or $Gamma(2,1)$ and autocorrelation parameter ρ (columns) of 0.1, 0.5, or 0.9. To reduce the influence of initial values, the first 100 iterations are omitted. The RMSE of the PQ estimates for $Gamma(2,1)$ after Box-Cox transformation is included as the green line. Both x and y axes are \log_{10} transformed.

distribution. Additionally, given that the PQ estimator relies on a normality assumption, we consider a variant of PQ where we apply PQ to samples after a Box-Cox transformation (Box and Cox, 1964), with the estimates then back-transformed to the original scale. We expect this Box-Cox procedure to enhance the performance of the PQ estimates when the distribution is non-normal.

The third row of Figure 4.2 presents the PQ and EQ expected estimates. Due to the right-skewness of the distribution, the 0.975 quantile of the approximate normal distribution is smaller than that of a $Gamma(2,1)$ distribution. While EQ estimates gradually converge to the true quantiles, PQ estimates converge to the quantiles of the approximate normal distribution, resulting in biased estimates. However, PQ estimates derived from Box-Cox transformed samples show a notable improvement in reducing this bias.

Correspondingly, Figure 4.3 finds that the RMSEs of the PQ estimates with the application of the Box-Cox transformation are smaller than those of the EQ estimates. This improvement can be attributed to the considerable variability inherent in the EQ estimates, as discussed in the previous section. This indicates that despite their unbiased nature, EQ estimates might not surpass PQ estimates because of their greater variability.

4.4 Real Data Example

To demonstrate our findings in a practical context, we performed a Bayesian analysis on data from a capture-recapture study of adult female Leisler’s bats. This dataset, consisting of 181 individuals, was collected between 1989 and 2008 in a forest in Thuringia, Germany, where bat boxes were regularly monitored. Detailed information about this study is available in Schorcht et al. (2009). Kéry and Schaub (2011, Chapter 10) conducted a Bayesian analysis of this dataset (implemented in WinBUGS) using a Jolly-Seber model (Jolly, 1965; Seber, 1965) with data augmentation through superpopulation parameterization (Royle and Dorazio, 2008), where “superpopulation” refers to all individuals ever alive during the study. All priors were set as continuous uniform distributions to remain uninformative. The specific model used can be found in Appendix G. We replicated this analysis using JAGS, focusing on two key parameters of interest: the size of the superpopulation, N_s , and the mean survival probability throughout the study, $\bar{\phi}$. We specifically chose these parameters because they are informed by 19 years of data, enhancing the available information and the degree of normality in their marginal posterior distributions.

To evaluate the performance of the PQ and EQ quantile estimators under practical conditions, we considered the effect of thinning on the estimation results. While we do not

advocate for thinning, as it discards valuable information, it remains a common practice among practitioners to manage long chains. The original analysis by [Kéry and Schaub \(2011\)](#) used three chains, with the first 5,000 iterations discarded as burn-in followed with 5,000 iterations, thinned to 2,500 samples. To be consistent with this study, we also used three chains but to further reduce the influence of the early sampling phase, we extended the burn-in period to the first 10,000 iterations and then retained 100,000 iterations after burn-in. Subsequently, we designed three thinning scenarios: high thinning (2,500 samples), medium thinning (10,000 samples), and no thinning (100,000 samples). We focused on the 0.025 and 0.975 quantiles of the parameters as they represent the endpoints of a 95% credible interval. To better satisfy the normality assumption, we also applied a Box-Cox transformation to both parameters and then transformed the quantile estimates back to the original scale for inference. [Figure H.1](#) in [Appendix H](#) presents the Q-Q plots for both parameters with and without a Box-Cox transformation, under no thinning and high thinning scenarios. For N_s , the untransformed samples exhibit right skewness, while the Box-Cox transformation significantly improves normality. In contrast, for $\bar{\phi}$, the untransformed data presents slightly fat tails. Although the Box-Cox transformation provides some improvement, the change is not substantial.

[Figures 4.4](#) and [4.5](#) illustrate the 0.025 and 0.975 quantile of $\bar{\phi}$ and N_s , respectively, under different thinning scenarios. For $\bar{\phi}$, the PQ estimator generally provides higher estimates compared to the EQ estimator. However, after applying a Box-Cox transformation, the PQ estimates align more closely with the EQ estimates, especially for the 0.025 quantile. Additionally, the PQ estimates exhibit greater stability and less fluctuation than the EQ estimates, consistent with the results presented in [Section 4.2](#) and [Section 4.3](#). The PQ estimates around the 5,000th iteration are similar to later estimates, while the EQ estimates still show significant changes.

For N_s , the PQ estimates based on the original samples are slightly lower for both quantiles across iterations, while the other two types of estimates are similar. Since N_s is a discrete parameter, the EQ estimates are confined to integer values. This restriction can limit the precision of the EQ estimates. In contrast, the PQ estimates for Box-Cox transformed samples provide more flexibility and are not bound to integer values, leading to potentially more accurate estimates. In fact, if the PQ estimates were rounded to the nearest integer, they would almost always align closely with the target value, demonstrating their effectiveness.

To further investigate the uncertainty in the estimates, [Figures H.2](#) and [H.3](#) in [Appendix H](#) display the squared error of the quantile estimates compared to the EQ estimate with the full set of samples. The PQ estimator based on the original samples shows a notably higher error than the EQ estimator, while the PQ estimator based on Box-Cox

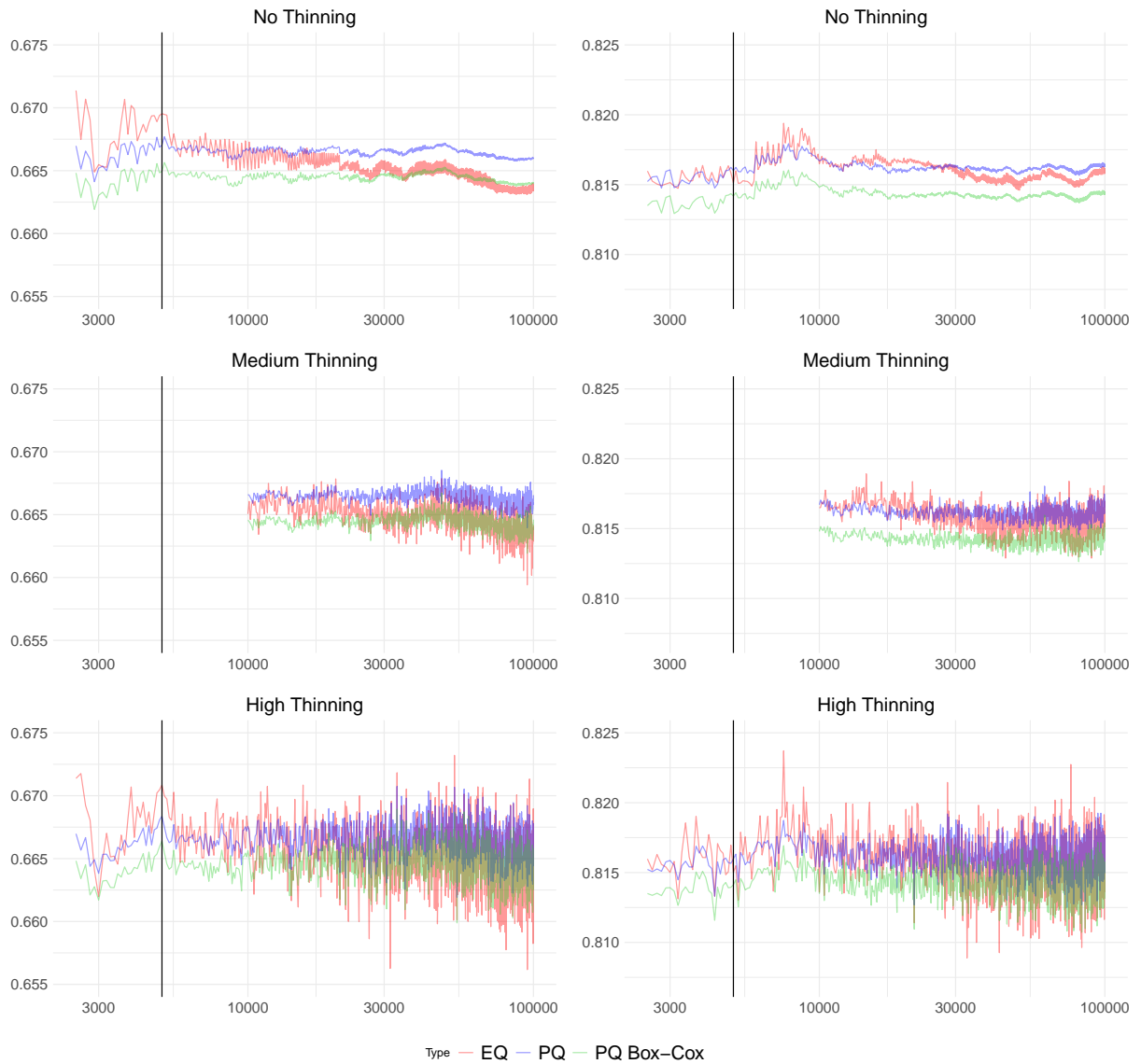


Figure 4.4: Estimates of the 0.025 and 0.975 quantile of $\bar{\phi}$. The left column shows the 0.025 quantile estimates, and the right column shows the 0.975 quantile estimates. The black vertical line indicates the 5,000th iteration, corresponding to the chain length used by [Kéry and Schaub \(2011\)](#). The x-axis, on a \log_{10} scale, shows iterations from 2,500 to 100,000 (prior to thinning).

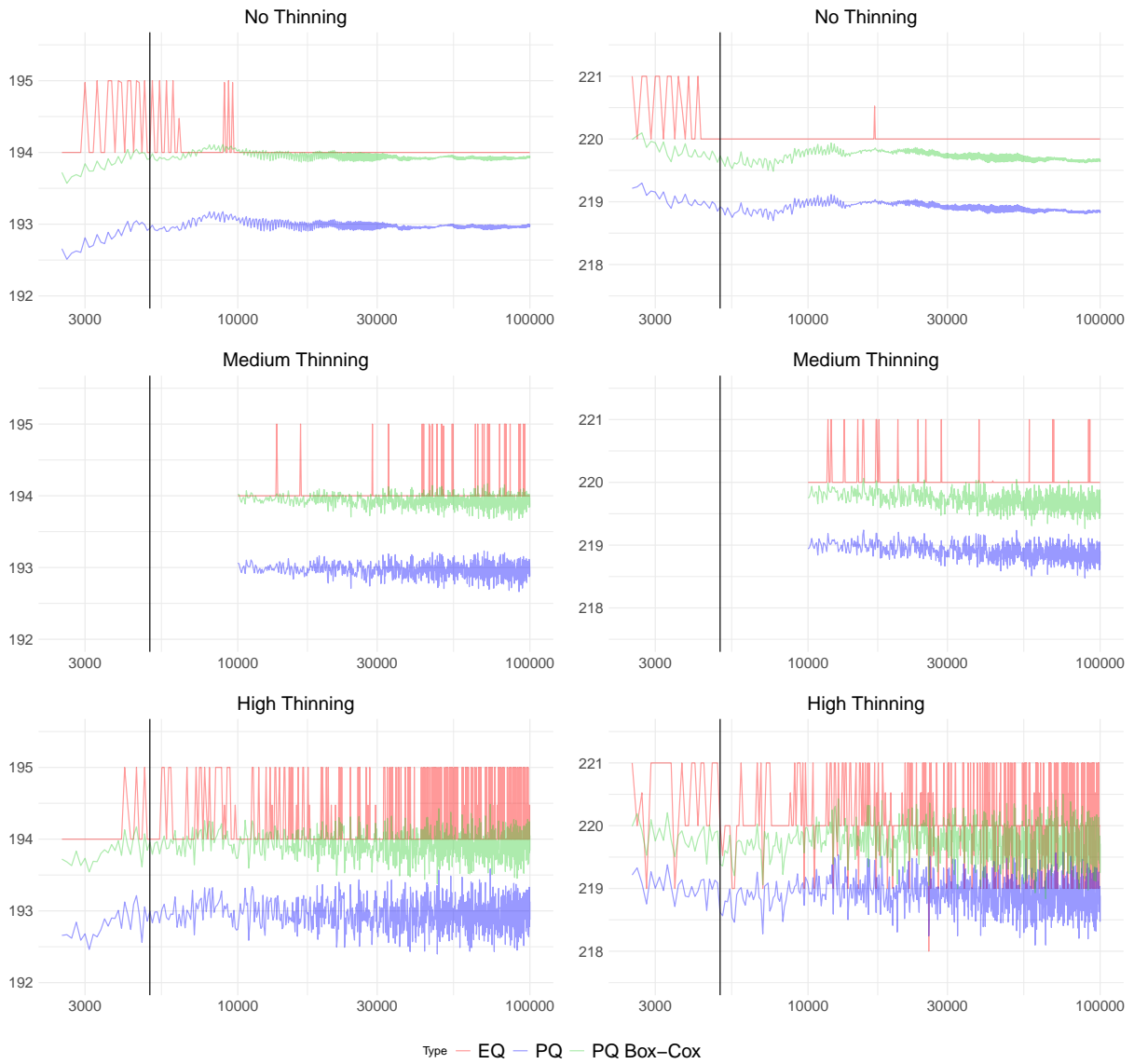


Figure 4.5: Estimates of the 0.025 and 0.975 quantile of N_s . The left column shows the 0.025 quantile estimates, and the right column shows the 0.975 quantile estimates. The black vertical line indicates the 5,000th iteration, corresponding to the chain length used by [Kéry and Schaub \(2011\)](#). The x-axis, on a \log_{10} scale, shows iterations from 2,500 to 100,000 (prior to thinning).

transformed samples exhibits the lowest squared error across all scenarios.

Additionally, we observed that increasing thinning levels led to greater variability in the quantile estimators. This underscores that as thinning reduces the effective sample size, the influence of variance becomes more significant compared to bias. Notably, despite this increased variability, the PQ estimates remained within the range of the EQ estimates, indicating their reliability. However, it is important to remember that thinning should only be used to manage computer memory limitations.

4.5 Discussion

In this chapter, we investigated the performance of parametric quantile estimation in MCMC processes, focusing on normal approximation. We first introduced a CLT for the PQ estimator, offering insights into its asymptotic behavior, and discussed methods for estimating the CLT variance. We then conducted simulation studies to demonstrate the effectiveness of the PQ estimator as an alternative to the EQ estimator under three scenarios: a normal scenario and two nonnormal scenarios. Finally, we applied the estimation methods to a female Leisler’s bats dataset and observed a similar trend to that seen in the empirical investigation.

Our study demonstrates that the PQ estimator offers notable advantages over the EQ estimator in MCMC sampling. The PQ estimator exhibits faster convergence and lower RMSE, particularly in the early stages of sampling. However, in cases where the posterior distribution deviates from normality, PQ estimates can be inconsistent and biased. In such scenarios, methods like the Box-Cox transformation can significantly enhance the performance of the PQ estimator. This was particularly evident in our real data analysis of Leisler’s bats, where the PQ estimates for Box-Cox transformed data provided more stable and accurate quantile estimates. These findings underscore the effectiveness of the PQ estimator in estimating posterior quantiles using MCMC samples, highlighting its potential to offer more reliable results under various conditions.

While our study provides valuable insights into the performance of two quantile estimation methods, there remain limitations and avenues for future research. Firstly, further investigation is needed to explore the performance of these estimation methods in broader scenarios, particularly in the context of MCMC sampling involving multiple dependent parameters. In such cases, the advantageous properties seen in simpler models like AR(1) processes may not hold, complicating the estimation process. Secondly, our results indicated that the PQ estimator significantly reduces the number of iterations required to

achieve the same accuracy as the EQ estimator under normality assumptions. This suggests the potential to develop practical stopping rules that can help practitioners determine when they have reached the desired accuracy of the quantile estimators, thereby optimizing computational efficiency. Lastly, in non-normality situations, we found that a Box-Cox transformation could reduce the bias of the PQ estimators, potentially resulting in a smaller RMSE compared to the unbiased EQ estimators due to the high variance of the EQ estimator. For this study, we focused on normal approximations; however, it is also worth considering other parametric distributions to approximate the posterior. Additionally, alternative transformations, such as the inverse normal transformation discussed by [Beasley et al. \(2009\)](#), could be explored as they may be more effective or broadly applicable than the Box-Cox transformation.

In conclusion, our work contributes to the understanding of uncertainty assessment in Bayesian inference by evaluating the performance of PQ and EQ quantile estimation methods in MCMC processes. By providing insights into the convergence behavior, robustness, and asymptotic properties of estimation methods, we aim to equip researchers with the knowledge needed to select the most appropriate estimation techniques for their Bayesian quantile estimation, potentially improving the accuracy and efficiency of their inferential results.

Chapter 5

Conclusion

This thesis developed novel methodologies for population size estimation using advanced capture-recapture methods. The primary focus has been on developing both Bayesian and frequentist methods to account for uncertainties that are often overlooked. This research was divided into two core topics: uncertain capture status in plant-capture and uncertainty propagation for genetic mark-recapture. In addition, a secondary focus was the improvement of computational techniques for Bayesian inference, notably for the estimation of the Bayesian credible interval.

In Chapter 2, we introduced a novel plant-capture modeling framework with missing at random assumptions. This framework addresses the uncertainty in the capture status of the plants and the heterogeneity between survey sites. We developed and compared two inference methods — one Bayesian and one frequentist — through simulation studies. The results showed that with small sample size, Bayesian methods provide coverage probabilities closer to the desired 95%, while frequentist methods exhibit smaller biases. The differences between methods diminish as the sample size increases. Our application of these methods to estimate the size of the homeless population in several U.S. cities highlighted the practical utility of the improved plant-capture model.

Chapter 3 focused on estimating population compositions using genetic stock identification (GSI) data. We proposed a reverse Dirichlet-Multinomial model to propagate uncertainties from the sample level to the population level and extended this approach to genetic mark-recapture contexts. To improve the computational efficiency, we developed moment-matching techniques, which provide estimates with similar accuracy while significantly reducing computational time. Additionally, an AR(1) prior was introduced to incorporate potential temporal trends in the study. Through comprehensive simulation

studies, we evaluate the performance of our proposed methods across various settings. The results demonstrated that our methods outperformed the current methods which underestimate the uncertainty of the parameters of interest. The real-data analysis on the Taku River Sockeye Salmon showcased the practical applicability of our techniques.

In Chapter 4, we delved into computational challenges in Bayesian inference, particularly in estimating credible intervals. Throughout this thesis, we encountered long computational times in several scenarios. To potentially reduce computation times, we investigated parametric quantile estimation methods in comparison to the conventional empirical quantile methods. We proposed a central limit theorem for the parametric quantile estimator of a normal posterior distribution and explored its asymptotic properties. Our simulation study found that when the posterior distribution is normal or approximately normal, the parametric quantile estimation provides more efficient estimates, reducing the computational time required to achieve similar accuracy to empirical quantile estimation. The practical benefits of our methods were further shown through a real-world application to a capture-recapture dataset on Leisler's bat.

Overall, this thesis has made significant contributions to capture-recapture methods, population size estimation and their Bayesian applications. The advancements presented in this thesis not only contribute to the theoretical landscape of statistical inference but also have tangible impacts on practical applications in ecology, epidemiology, and related fields.

References

- Aitchison, J. (1986). *The statistical analysis of compositional data*. Monographs on statistics and applied probability. Chapman and Hall, London.
- Allendorf, F. W., Luikart, G. H., and Aitken, S. N. (2012). *Conservation and the genetics of populations*. John Wiley & Sons, 2nd edition.
- Ashbridge, J. and Goudie, I. (2008). Conditionally unbiased estimation of population size under plant-capture. *Communications in Statistics—Theory and Methods* **38**, 1–12.
- Bailey, N. T. (1952). Improvements in the interpretation of recapture data. *The Journal of Animal Ecology* **21**, 120–127.
- Barbraud, C., Delord, K., Marteau, C., and Weimerskirch, H. (2009). Estimates of population size of white-chinned petrels and grey petrels at kerguelen islands and sensitivity to fisheries. *Animal Conservation* **12**, 258–265.
- Barrett, D. F., Anolik, I., and Abramson, F. H. (1992). The 1990 census shelter and street night enumeration. In *JSM Proceedings, Survey Research Methods Section*, pages 194–198, Alexandria, VA. American Statistical Association.
- Beasley, T. M., Erickson, S., and Allison, D. B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior genetics* **39**, 580–595.
- Berry, B. (2007). A repeated observation approach for estimating the street homeless population. *Evaluation review* **31**, 166–199.
- Bonner, S. J., Schofield, M. R., Noren, P., and Price, S. J. (2016). Extending the latent multinomial model with complex error processes and dynamic markov bases. *The Annals of Applied Statistics* **10**, 246–263.

- Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons, 2nd edition.
- Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **26**, 211–243.
- Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury Press, Pacific Grove, 2nd edition.
- Chan, K. S. and Geyer, C. J. (1994). Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics* **22**, 1747–1758.
- Chapman, D. G. (1951). Some properties of the hypergeometric distribution with applications to zoological censuses. *Univ. Calif. Stat.* **1**, 60–131.
- Chen, M.-H. and Shao, Q.-M. (1999). Monte carlo estimation of bayesian credible and hpd intervals. *Journal of computational and Graphical Statistics* **8**, 69–92.
- City of Red Deer (2016). *2016 Point in Time Homeless Count Report*. Red Deer, Canada.
- Clark, J. S. and Gelfand, A. E. (2006). A future for models and data in environmental science. *Trends in Ecology & evolution* **21**, 375–380.
- Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A., and Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989–2000.
- Coumans, A., Cruyff, M., Van der Heijden, P. G., Wolf, J., and Schmeets, H. (2017). Estimating homelessness in the netherlands using a capture-recapture approach. *Social Indicators Research* **130**, 189–212.
- de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* **26**, 403–417.
- Doss, C. R., Flegal, J. M., Jones, G. L., and Neath, R. C. (2014). Markov chain monte carlo estimation of quantiles. *Electronic Journal of Statistics* **8**, 2448–2478.
- Duran, J. W. and Wiorkowski, J. J. (1981). Capture-recapture sampling for estimating software error content. *IEEE Transactions on Software Engineering* **SE-7**, 147–148.

- Flegal, J. M., Hughes, J., Vats, D., Dai, N., Gupta, K., and Maji, U. (2021). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA, and Kanpur, India. R package version 1.5-0.
- Fournier, D., Beacham, T., Riddell, B., and Busack, C. (1984). Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characteristics. *Canadian Journal of Fisheries and Aquatic Sciences* **41**, 400–408.
- Gabry, J., Češnovar, R., and Johnson, A. (2023). *cmdstanr: R Interface to 'CmdStan'*. <https://mc-stan.org/cmdstanr/>, <https://discourse.mc-stan.org>.
- Gazey, W. J. (2010). Gsi sample size requirements for in-river run reconstruction of alsek chinook and sockeye stocks. Technical report, Pacific Salmon Commission, Vancouver, British Columbia.
- Gelman, A. (1995). Method of moments using monte carlo simulation. *Journal of Computational and Graphical Statistics* **4**, 36–54.
- Gelman, A., Bois, F., and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association* **91**, 1400–1412.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press, 3rd edition.
- Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical science* **7**, 473–483.
- Geyer, C. J. (1998). Markov chain monte carlo lecture notes.
- Geyer, C. J. (2011). Introduction to markov chain monte carlo. In *Handbook of Markov Chain Monte Carlo*, chapter 1, pages 3–48. CRC press, New York.
- Gilbert, P. and Varadhan, R. (2019). *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1.1.
- Gilks, W. R., Thomas, A., and Spiegelhalter, D. J. (1994). A language and program for complex bayesian modelling. *Journal of the Royal Statistical Society: Series D (The Statistician)* **43**, 169–177.
- Goudie, I. and Ashbridge, J. (2000). A conditionally-unbiased estimator of population size based on plant-capture in continuous time. *Communications in Statistics-Theory and Methods* **29**, 2605–2619.

- Goudie, I., Jupp, P., and Ashbridge, J. (2007). Plant-capture estimation of the size of a homogeneous population. *Biometrika* **94**, 243–248.
- Goudie, I., Pollock, K. H., and Ashbridge, J. (1998). A plant-capture approach for population size estimation in continuous time. *Communications in statistics-theory and methods* **27**, 433–451.
- Grant, W. S., Milner, G. B., Krasnowski, P., and Utter, F. M. (1980). Use of biochemical genetic variants for identification of sockeye salmon (*oncorhynchus nerka*) stocks in cook inlet, alaska. *Canadian Journal of Fisheries and Aquatic Sciences* **37**, 1236–1247.
- Gustafson, P. (2023). Parameter restrictions for the sake of identification: Is there utility in asserting that perhaps a restriction holds? *Statistical Science* **38**, 477–489.
- Hamazaki, T. and DeCovich, N. (2014). Application of the genetic mark–recapture technique for run size estimation of yukon river chinook salmon. *North American Journal of Fisheries Management* **34**, 276–286.
- Hess, J. E., Whiteaker, J. M., Fryer, J. K., and Narum, S. R. (2014). Monitoring stock-specific abundance, run timing, and straying of chinook salmon in the columbia river using genetic stock identification (gsi). *North American Journal of Fisheries Management* **34**, 184–201.
- Hopper, K. (1991). Monitoring and evaluating the 1990 s-night in new york city. In *Final report on Joint Statistical Agreement 90-18 between the Bureau of the Census and the Nathan Kline Institute for Psychiatric Research*. Orangeburg, New York.
- Hopper, K., Shinn, M., Laska, E., Meisner, M., and Wanderling, J. (2008). Estimating numbers of unsheltered homeless people through plant-capture and postcount survey methods. *American Journal of Public Health* **98**, 1438–1442.
- Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika* **52**, 225–247.
- Jones, G. L. (2004). On the markov chain central limit theorem. *Probability Surveys* **1**, 299–320.
- Kelter, R. (2024). The bayesian simulation study (basis) framework for simulation studies in statistical and methodological research. *Biometrical Journal* **66**, 2200095.
- Kéry, M. (2010). *Introduction to WinBUGS for ecologists: Bayesian approach to regression, ANOVA, mixed models and related analyses*. Academic Press.

- Kéry, M. and Schaub, M. (2011). *Bayesian population analysis using WinBUGS: a hierarchical perspective*. Academic Press.
- Koljonen, M.-L., Pella, J. J., and Masuda, M. (2005). Classical individual assignments versus mixture modeling to estimate stock proportions in atlantic salmon (*salmo salar*) catches from dna microsatellite data. *Canadian Journal of Fisheries and Aquatic Sciences* **62**, 2143–2158.
- Kruschke, J. B. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 2nd edition.
- Kuismin, M., Saatoglu, D., Niskanen, A. K., Jensen, H., and Sillanpää, M. J. (2020). Genetic assignment of individuals to source populations using network estimation tools. *Methods in Ecology and Evolution* **11**, 333–344.
- Laska, E. M. and Meisner, M. (1993). A plant-capture method for estimating the size of a population from a single sample. *Biometrics* **49**, 209–220.
- Link, W. A., Yoshizaki, J., Bailey, L. L., and Pollock, K. H. (2010). Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. *Biometrics* **66**, 178–185.
- Liu, J., Nordman, D. J., and Meeker, W. Q. (2016). The number of mcmc draws needed to compute bayesian credible bounds. *The American Statistician* **70**, 275–284.
- Luikart, G. and England, P. R. (1999). Statistical analysis of microsatellite dna data. *Trends in Ecology & Evolution* **14**, 253–256.
- Lukacs, P. M. and Burnham, K. P. (2005). Estimating population size from dna-based closed capture-recapture data incorporating genotyping error. *The Journal of Wildlife Management* **69**, 396–403.
- Martin, E. (1992). Assessment of s-night street enumeration in the 1990 census. *Evaluation review* **16**, 418–438.
- Martin, E., Laska, E., Hopper, K., and Wanderling, J. (1997). Issues in the use of a plant-capture method for estimating the size of the street dwelling population. *Journal of Official Statistics* **13**, 59.
- McCandless, L. C., Patterson, M. L., Currie, L. B., Moniruzzaman, A., and Somers, J. M. (2016). Bayesian estimation of the size of a street-dwelling homeless population. *Journal of Modern Applied Statistical Methods* **15**, 15.

- McCarthy, M. A. (2007). *Bayesian methods for ecology*. Cambridge University Press.
- Meyn, S. P. and Tweedie, R. L. (2009). *Markov chains and stochastic stability*. Cambridge University Press, Cambridge.
- Millar, R. (1991). Selecting loci for genetic stock identification using maximum likelihood, and the connection with curvature methods. *Canadian Journal of Fisheries and Aquatic Sciences* **48**, 2173–2179.
- Millar, R. B. (1987). Maximum likelihood estimation of mixed stock fishery composition. *Canadian Journal of Fisheries and Aquatic Sciences* **44**, 583–590.
- Milner, G., Teel, D., Utter, F., and Burley, C. (1981). *Columbia River stock identification study: validation of genetic method*. National Marine Fisheries Service.
- National Law Center on Homelessness & Poverty (2017). *Don't Count on It: How the HUD Point-in-Time Count Underestimates the Homelessness Crisis in America*. Washington, DC.
- Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The computer journal* **7**, 308–313.
- Nichols, J. D. and MacKenzie, D. I. (2004). Abundance estimation and conservation biology. *Animal biodiversity and conservation* **27**, 437–439.
- Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. John Wiley & Sons.
- OrgCode Consulting Inc. (2012). *Red Deer Point In Time [PIT] Homeless Count 2012 Final Report*. Red Deer, Canada.
- OrgCode Consulting, Inc. (2013). *The State of Homelessness in Kingston, 2013*. Kingston, Canada.
- Östergren, J., Palm, S., Gilbey, J., and Dannewitz, J. (2020). Close relatives in population samples: evaluation of the consequences for genetic stock identification. *Molecular ecology resources* **20**, 498–510.
- Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife monographs* pages 3–135.

- Pella, J. and Masuda, M. (2001). Bayesian methods for analysis of stock mixtures from genetic characters. *Fishery Bulletin* **99**, 151–167.
- Pestal, G., Schwarz, C. J., and Clark, R. A. (2020). Taku river sockeye salmon stock assessment review and updated 1984-2018 abundance estimates. Technical Report 32, Pacific Salmon Commission, Vancouver, British Columbia.
- Petersen, C. G. J. (1896). The yearly immigration of young plaice in the limfjord from the german sea. *Rept. Danish Biol. Sta.* **6**, 1–48.
- Plummer, M., Best, N., Cowles, K., Vines, K., et al. (2006). Coda: convergence diagnosis and output analysis for mcmc. *R news* **6**, 7–11.
- Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria.
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rice, K. and Ye, L. (2022). Expressing regret: a unified view of credible intervals. *The American Statistician* **76**, 248–256.
- Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid markov chains. *Electronic Communications in Probability* **2**, 13 – 25.
- Roberts, G. O. and Rosenthal, J. S. (2004). General state space markov chains and mcmc algorithms. *Probability Surveys* **1**, 20–71.
- Royle, J. A. and Dorazio, R. M. (2008). *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Elsevier.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.
- Rukhin, A. (1975). Statistical decision about the total number of observable objects. *Sankhyā: The Indian Journal of Statistics, Series A* **37**, 514–522.
- Schaub, M. and Kéry, M. (2021). *Integrated population models: Theory and ecological applications with R and JAGS*. Academic Press.
- Schofield, M. R. and Bonner, S. J. (2015). Connecting the latent multinomial. *Biometrics* **71**, 1070–1080.

- Schorcht, W., Bontadina, F., and Schaub, M. (2009). Variation of adult survival drives population dynamics in a migrating forest bat. *Journal of Animal Ecology* **78**, 1182–1190.
- Seber, G. A. (1965). A note on the multiple-recapture census. *Biometrika* **52**, 249–259.
- Seber, G. A. F. (1982). *The estimation of animal abundance and related parameters*. Charles Griffin, London, 2nd edition.
- Simtech Solutions, Inc. (2023). The counting us mobile app.
- Skalski, J. and Robson, D. (1982). A mark and removal field procedure for estimating population abundance. *The Journal of Wildlife Management* **46**, 741–751.
- Smouse, P. E., Waples, R. S., and Tworek, J. A. (1990). A genetic mixture analysis for use with incomplete source population data. *Canadian Journal of Fisheries and Aquatic Sciences* **47**, 620–634.
- Stan Development Team (2024). *Stan Modeling Language Users Guide and Reference Manual, Version 2.34*. <https://mc-stan.org>.
- Taberlet, P., Waits, L. P., and Luikart, G. (1999). Noninvasive genetic sampling: look before you leap. *Trends in ecology & evolution* **14**, 323–327.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics* **22**, 1701–1728.
- Turkman, M. A. A., Paulino, C. D., and Müller, P. (2019). *Computational Bayesian statistics: an introduction*, volume 11 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, Cambridge.
- United States Department of Housing and Urban Development (2008). A guide to counting unsheltered homeless people, 2nd revision.
- Vaart, A. W. v. d. (1998). *Asymptotic Statistics*, chapter 10, page 138–152. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Vale, R., Fewster, R., Carroll, E., and Patenaude, N. (2014). Maximum likelihood estimation for model mt, α for capture–recapture data with misidentification. *Biometrics* **70**, 962–971.

- Vansteelandt, S., Goetghebeur, E., and Verstraeten, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* **56**, 1126–1133.
- Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for markov chain monte carlo. *Biometrika* **106**, 321–337.
- Waits, L. P. and Paetkau, D. (2005). Noninvasive genetic sampling tools for wildlife biologists: a review of applications and recommendations for accurate data collection. *The Journal of Wildlife Management* **69**, 1419–1433.
- Williams, B. K., Nichols, J. D., and Conroy, M. J. (2002). *Analysis and management of animal populations*. Academic press.
- Wright, J. A., Barker, R. J., Schofield, M. R., Frantz, A. C., Byrom, A. E., and Gleeson, D. M. (2009). Incorporating genotype uncertainty into mark–recapture-type models for estimating abundance using dna samples. *Biometrics* **65**, 833–840.
- Yip, P. and Fong, D. Y. (1993). Estimating population size from a removal experiment. *Statistics & probability letters* **16**, 129–135.
- Yip, P. S., Xi, L., Fong, D. Y., and Hayakawa, Y. (1999). Sensitivity-analysis and estimating number-of-faults in removal debugging. *IEEE Transactions on Reliability* **48**, 300–305.
- Yoshizaki, J. (2007). *Use of natural tags in closed population capture-recapture studies: Modeling misidentification*. Phd thesis, North Carolina State University, Raleigh, NC.
- Yoshizaki, J., Brownie, C., Pollock, K. H., and Link, W. A. (2011). Modeling misidentification errors that result from use of genetic tags in capture–recapture studies. *Environmental and Ecological Statistics* **18**, 27–55.
- Youngflesh, C. (2018). Mcmcvis: tools to visualize, manipulate, and summarize mcmc output. *Journal of Open Source Software* **3**, 640.

APPENDICES

Appendix A

Derivation of MLE for Model \mathcal{M}_{basic}

Based on the joint likelihood of the parameters of interest γ described in Section 2.3.1, it is easy to get the log-likelihood as

$$\begin{aligned}
 l(\gamma; y, m^{yes}, m^{mb}, m^{no}) &= m^{yes} \log\{p^c(1 - p^{mb})\} + m^{mb} \log p^{mb} + m^{no} \log\{(1 - p^c)(1 - p^{mb})\} \\
 &\quad + \log\{(H + m^{mb})!\} - \log\{(H + m^{mb} - y + m^{yes})!\} \\
 &\quad + (y - m^{yes}) \log p^c + (H + m^{mb} - y + m^{yes}) \log(1 - p^c)
 \end{aligned} \tag{A.1}$$

Taking partial derivatives on Equation (A.1) with respect to p^{mb} and p^c and letting them equal to 0, we arrive at the ML estimator for p^{mb} as $\frac{M^{mb}}{M^{yes} + M^{mb} + M^{no}} = \frac{M^{mb}}{M}$ and the ML estimator for p^c (with known H) as $\frac{Y}{H + M^{yes} + M^{mb} + M^{no}} = \frac{Y}{H + M}$.

To find the ML estimator for H , we consider the ratio

$$\begin{aligned}
 \frac{L(H + 1)}{L(H)} &= \frac{\binom{H + 1 + m^{mb}}{y - m^{yes}} (1 - p^c)^{H + 1 + m^{mb} - y + m^{yes}}}{\binom{H + m^{mb}}{y - m^{yes}} (1 - p^c)^{H + m^{mb} - y + m^{yes}}} \\
 &= \frac{H + 1 + m^{mb}}{H + 1 + m^{mb} - y + m^{yes}} (1 - p^c) < 1 \text{ when } H > \frac{y - m^{yes} - m^{mb} p^c}{p^c} - 1.
 \end{aligned}$$

This implies a ML estimator of H (when p^c is known) as $\lfloor \frac{Y - M^{yes} - M^{mb} p^c}{p^c} \rfloor$. The ML estimator for p^{mb} does not depend on the other two parameters, but the ML estimators for H

and p^c depend on each other. Since the ML estimators need to satisfy all three expressions simultaneously, we can solve for H and p^c to yield the ML estimators for each parameter: $\hat{p}^c = \frac{M^{yes}}{M^{yes}+M^{no}}$, $\hat{p}^{mb} = \frac{M^{mb}}{M}$ and $\hat{H} = \lfloor \frac{Y-M^{yes}-M^{mb}\hat{p}^c}{\hat{p}^c} \rfloor$, where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x .

Appendix B

Alternative Computational Methods for Bayesian Inference

We explored two alternative computational approaches to fit our models. These approaches are detailed in this section, and evaluated in a simulation study in Appendix C.

B.1 Bayesian Normal Approximation (BNA)

Bayesian normal approximation is an approach that constructs an approximate representation of the posterior distribution using a multivariate normal distribution. The mean of the distribution is approximated by the vector of posterior mode of the parameters, obtained via numerical optimization of the posterior distribution. The variance-covariance matrix of the distribution is defined as the inverse of the negative Hessian matrix of the log posterior density at the modes. A more comprehensive description of this technique is given in [Gelman et al. \(2013\)](#).

Similarly to the MLE method described in Section 2.3.1, we can express the posterior distribution of our proposed models in a general form $\pi(\boldsymbol{\gamma}|\mathbf{x}) \propto \pi(\boldsymbol{\gamma})P(\mathbf{X} = \mathbf{x}|\boldsymbol{\gamma}) = \pi(\boldsymbol{\gamma}) \sum_{\mathbf{z} \in \Omega} P(\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}|\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ denotes the model parameters with a prior distribution $\pi(\boldsymbol{\gamma})$, \mathbf{X} represents the data, \mathbf{Z} stands for latent variables, and Ω denotes the set of values that \mathbf{Z} can take. We also apply a log transformation on the counts and a logit transformation on the probabilities to remove any constraints on their bounds, avoiding computational issues related to boundary constraints. The numerical method to apply this approach is the same as the MLE approach detailed in Section 2.3.1. Furthermore, the prior settings for the parameters remain the same with those described in Section 2.4.

B.2 Uncertainty Propagation Method (UP)

While inference for model \mathcal{M}_{basic} is relatively straightforward using MCMC algorithms, the specification of models \mathcal{M}_{id} and \mathcal{M}_{class} using probabilistic programming languages can be relatively complicated because of the equality constraints in Equations (2.10) and (2.13), as discussed in Section 2.3.2. Instead of using the *dsum* function in JAGS, another simple solution is to employ an uncertainty propagation (UP) method to obtain an approximate posterior inference for H . To illustrate this method, we use model \mathcal{M}_{id} as an example.

Initially, it is important to recognize that if p^c and H^c were observed, it would be feasible to construct an approximate representation of the posterior distribution $\pi(H|p^c, H^c)$ using a normal distribution due to the Bernstein–von Mises theorem, as follows:

$$H|H^c, p^c \sim N\left(\hat{H}_0, \frac{\hat{H}_0(1-p^c)}{p^c}\right), \quad (\text{B.2.1})$$

where $\hat{H}_0 = H^c/p^c$ is a variant of the MLE for H (Rukhin, 1975) based on the binomial distribution in Equation (2.8). A proof of the asymptotic correspondence of \hat{H}_0 with the MLE is presented in the box below.

Proof

Suppose we have $H^c \sim \text{Binom}(H, p^c)$. Given p^c and H^c , we use the method introduced in the Appendix of the main paper to derive the MLE of H :

$$\begin{aligned} \frac{L(H+1)}{L(H)} &= \frac{\frac{(H+1)!}{H^c!(H+1-H^c)!} (p^c)^{H^c} (1-p^c)^{H+1-H^c}}{\frac{H!}{H^c!(H-H^c)!} (p^c)^{H^c} (1-p^c)^{H-H^c}} \\ &= \frac{H+1}{H+1-H^c} (1-p^c) < 1 \text{ when } H > \frac{H^c}{p^c} - 1, \end{aligned}$$

which leads to the ML estimator of H as $\lfloor \frac{H^c}{p^c} \rfloor$.

The approximate MLE, \hat{H}_0 , has variance

$$\text{Var}(\hat{H}_0) = \frac{\text{Var}(H^c)}{(p^c)^2} = \frac{H(1-p^c)}{p^c},$$

which is estimated by $\frac{\hat{H}_0(1-p^c)}{p^c}$ in Equation (B.2.1). Hence, given observed values of p^c and H^c , we could sample from the approximate posterior distribution by sampling directly from Equation (B.2.1).

In practice, while p^c and H^c are not directly observable, we can sample from their posterior distribution conditional on the observed data \mathbf{x} , $\pi(H^c, p^c|\mathbf{x})$. This is possible because p^c and H^c are completely informed when Equation (2.8) is removed from our model. In fact, the role of Equation (2.8) is strictly to expand H^c into H via p^c , offering no insights into any model parameters other than H . To summarize, fitting model \mathcal{M}_{id} without Equation (2.8) and marginalizing over p^{ilc}, p^{mb} and $M^{mb,c}$ provides a posterior sample from $L(H^c, p^c|\mathbf{x})$. Once an MCMC sample is obtained (first step), the values can be plugged into Equation (B.2.1) to simulate posterior samples from H (second step), which result in the desired approximate posterior distribution $\pi(H|\mathbf{x})$. Essentially, uncertainty from the first step is propagated into the second step.

Instead of employing a two-step approach, the desired outcome can be achieved more straightforwardly in a single step by replacing Equation (2.8) in our model with the approximate normal distribution (B.2.1). This substitution simplifies the MCMC process and enhances computational efficiency. BUGS-based software can thus be used to directly sample from our approximate representation of $\pi(\boldsymbol{\gamma}|\mathbf{x})$, which can then be marginalized to sample from $\pi(H|\mathbf{x})$.

Finally, when implementing the proposed UP method in BUGS languages, the following grammatical nuance must be considered: Equations (2.10) and (2.13) need to be rearranged to shift H^c to the left side. This ensures that the software doesn't interpret H^c as an undefined node.

Appendix C

Supplementary Tables for Chapter 2

Table C.1: Results of the simulation studies for \mathcal{M}_{basic} using BNA. All the values are rounded to integers or 2 decimal points.

Method	M	Parameter	True Value	Median	SD	RBias	RRMSE	CP	LCI
BNA	15	H	150	149	27	-0.01	0.18	0.91	109
		p^c	0.7	0.72	0.11	0.03	0.15	0.99	0.43
		p^{mb}	0.2	0.24	0.10	0.19	0.49	0.94	0.39
BNA	100	H	1,500	1,498	112	0.00	0.08	0.94	442
		p^c	0.7	0.70	0.05	0.01	0.07	0.95	0.19
		p^{mb}	0.2	0.21	0.04	0.03	0.20	0.96	0.16

Table C.2: Results of the simulation studies for \mathcal{M}_{id} using BNA and UP method. All the values are rounded to integers or 2 decimal points.

Method	M	Parameter	True Value	Median	SD	RBias	RRMSE	CP	LCI
BNA	15	H	150	150	26	0.00	0.17	0.92	105
		p^c	0.7	0.71	0.11	0.02	0.14	0.98	0.42
		$p^{mb ni}$	0.2	0.28	0.15	0.39	0.73	0.93	0.55
		$p^{i c}$	0.8	0.79	0.04	-0.01	0.05	0.94	0.15
UP	15	\hat{H}	150	159	38	0.06	0.22	0.97	144
		\hat{H}_0	150	159	37	0.06	0.22	0.96	138
		p^c	0.7	0.68	0.11	-0.03	0.16	0.97	0.43
		$p^{mb ni}$	0.2	0.26	0.14	0.30	0.73	0.96	0.52
		$p^{i c}$	0.8	0.80	0.04	-0.01	0.05	0.95	0.14
BNA	100	H	1,500	1,500	105	0.00	0.07	0.93	414
		p^c	0.7	0.70	0.05	0.00	0.07	0.94	0.18
		$p^{mb ni}$	0.2	0.21	0.06	0.06	0.29	0.96	0.23
		$p^{i c}$	0.8	0.80	0.01	0.00	0.02	0.95	0.05
UP	100	\hat{H}	1,500	1,512	111	0.01	0.07	0.94	434
		\hat{H}_0	1,500	1,513	108	0.01	0.07	0.93	422
		p^c	0.7	0.70	0.05	-0.00	0.07	0.94	0.18
		$p^{mb ni}$	0.2	0.21	0.06	0.04	0.29	0.96	0.23
		$p^{i c}$	0.8	0.80	0.01	0.00	0.02	0.96	0.05

Table C.3: Results of the simulation studies for \mathcal{M}_{class} using BNA and UP method. All the values are rounded to integers or 2 decimal points.

Method	M	Parameter	True Value	Median	SD	RBias	RRMSE	CP	LCI
BNA	30	H	300	294	40	-0.02	0.12	0.99	218
		p_{easy}^c	0.9	0.86	0.08	-0.04	0.08	0.94	0.33
		p_{hard}^c	0.4	0.48	0.13	0.19	0.35	0.95	0.48
		$p^{mb ni}$	0.2	0.23	0.10	0.14	0.49	0.95	0.41
		$p^{i c}$	0.8	0.80	0.03	-0.00	0.03	0.96	0.10
UP	30	\hat{H}	300	327	103	0.09	0.21	0.96	323
		\hat{H}_0	300	328	100	0.09	0.21	0.95	312
		p_{easy}^c	0.9	0.84	0.08	-0.06	0.09	0.94	0.32
		p_{hard}^c	0.4	0.42	0.13	0.04	0.32	0.96	0.49
		$p^{mb ni}$	0.2	0.22	0.10	0.09	0.49	0.98	0.38
BNA	100	H	1,500	1,486	124	-0.01	0.08	0.98	634
		p_{easy}^c	0.9	0.90	0.04	-0.00	0.05	0.96	0.17
		p_{hard}^c	0.4	0.42	0.08	0.06	0.20	0.95	0.29
		$p^{mb ni}$	0.2	0.21	0.06	0.06	0.29	0.96	0.23
		$p^{i c}$	0.8	0.80	0.01	-0.00	0.02	0.95	0.05
UP	100	\hat{H}	1,500	1,538	156	0.03	0.10	0.96	606
		\hat{H}_0	1,500	1,538	152	0.03	0.10	0.95	589
		p_{easy}^c	0.9	0.89	0.04	-0.01	0.05	0.96	0.16
		p_{hard}^c	0.4	0.41	0.08	0.00	0.19	0.96	0.30
		$p^{mb ni}$	0.2	0.21	0.06	0.04	0.29	0.95	0.23
		$p^{i c}$	0.8	0.80	0.01	-0.00	0.02	0.95	0.05

Table C.4: Estimation of the homeless population size H using the Chapman-Bailey estimator. All the values are rounded to integers.

City	Maybe as seen		Maybe as not seen	
	Estimate	95% CI	Estimate	95% CI
Chicago	7	(3, 20)	42	(23, ∞)
New Orleans	63	(56, 72)	76	(65, 91)
Phoenix	96	(81, 118)	102	(85, 129)
New York	1,520	(1,368, 1,721)	1,670	(1,624, 2,233)
Los Angeles	257	(212, 335)	289	(231, 402)

Appendix D

Determining ψ in the AR(1) Prior Specification for $\pi_{k,t}$

Determining an appropriate value for ψ is crucial to ensure that $\pi_{k,t}$ can encompass values across the entire range of $[0,1]$. To explore various options, we arbitrarily selected four values for ψ : 0.5, 2, 5, and 10, and conducted a simulation to assess their performance. With $K = 4$ (as in the Taku River data example from Section 3.5), we simulated 10,000 $\pi_{1,1}$'s for each ψ value. Figure D.1 depicts the density plots for $\pi_{1,1}$ under different ψ settings. It is evident that a small $\psi = 0.5$ confines the range of $\pi_{1,1}$ mostly below 0.5, whereas $\psi \geq 5$ renders values between 0.25 and 0.75 quite improbable. Conversely, when $\psi = 2$, the density curve is more evenly distributed across the range $[0,1]$, albeit showing an expected bias towards $1/K$. Consequently, $\psi = 2$ emerges as a reasonable choice for the scenario $K = 4$, which we choose for our analyses.

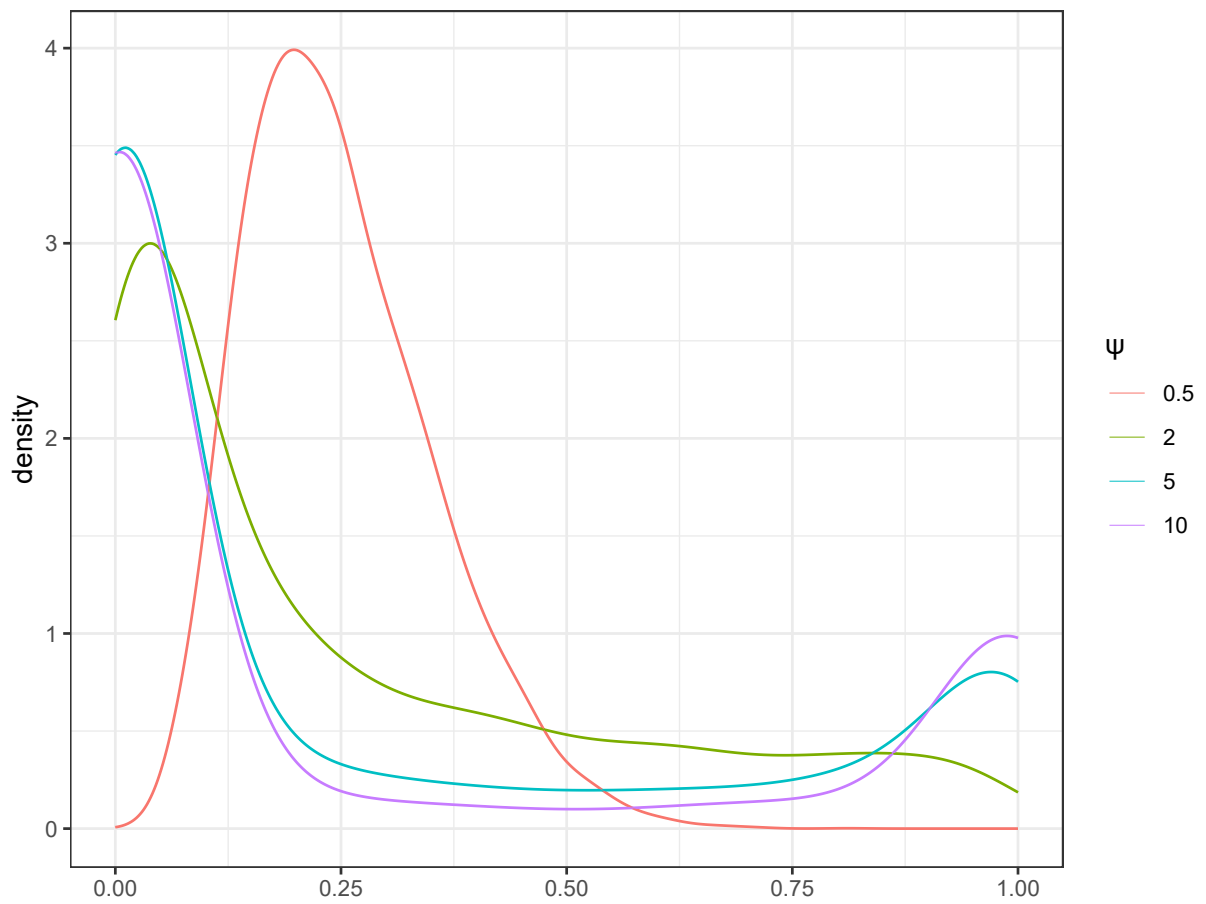


Figure D.1: Density plot for $\pi_{1,1}$ with different choices of ψ in the time series prior.

Appendix E

Supplementary Figures for Chapter 3

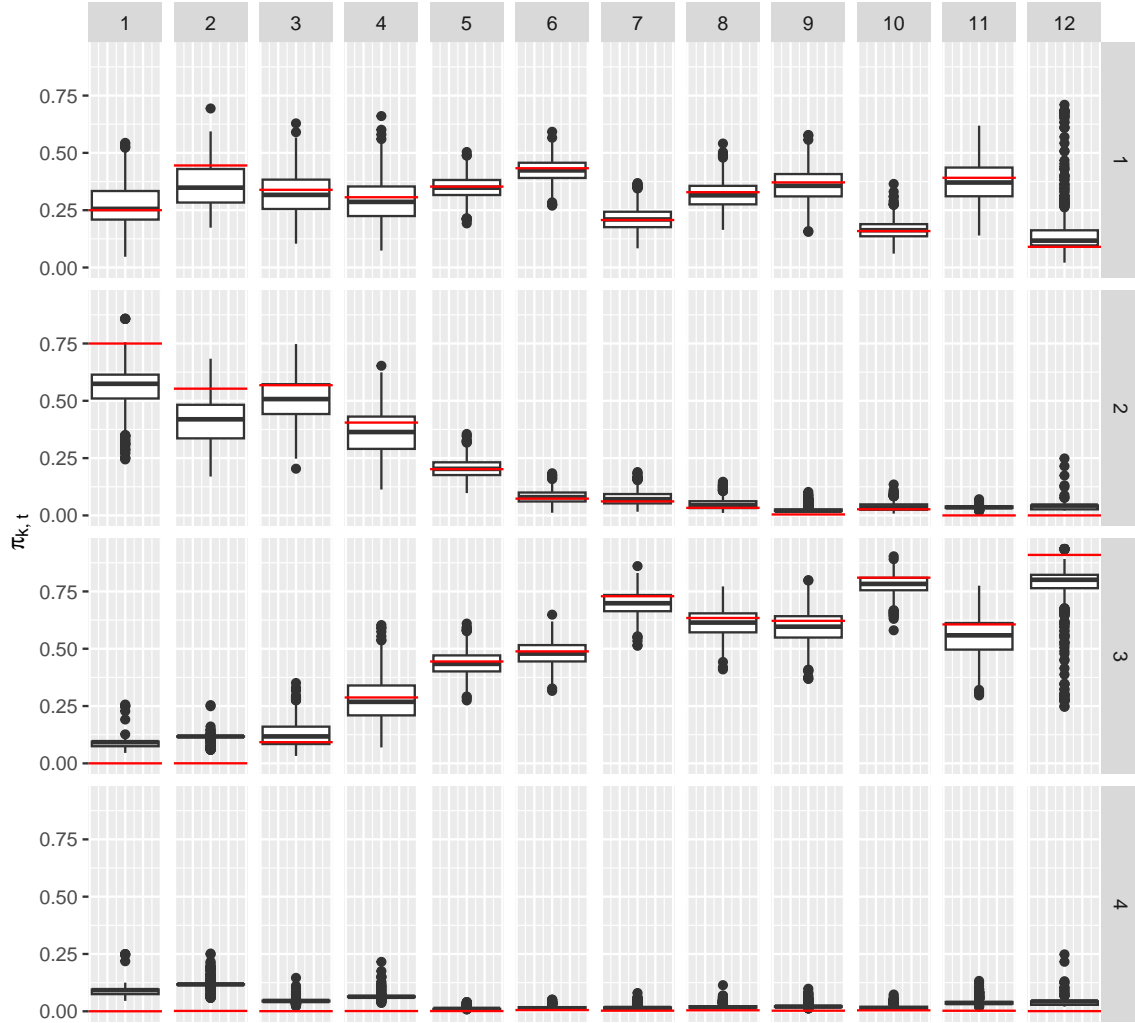


Figure E.1: Distribution of the posterior mean for $\pi_{k,t}$ from the approximate reverse Dirichlet-multinomial model with Dirichlet prior in the simulation study, for stocks $k = 1$ to 4 in weeks $t = 1$ to 12. Red horizontal lines indicate the true $\pi_{k,t}$ values.

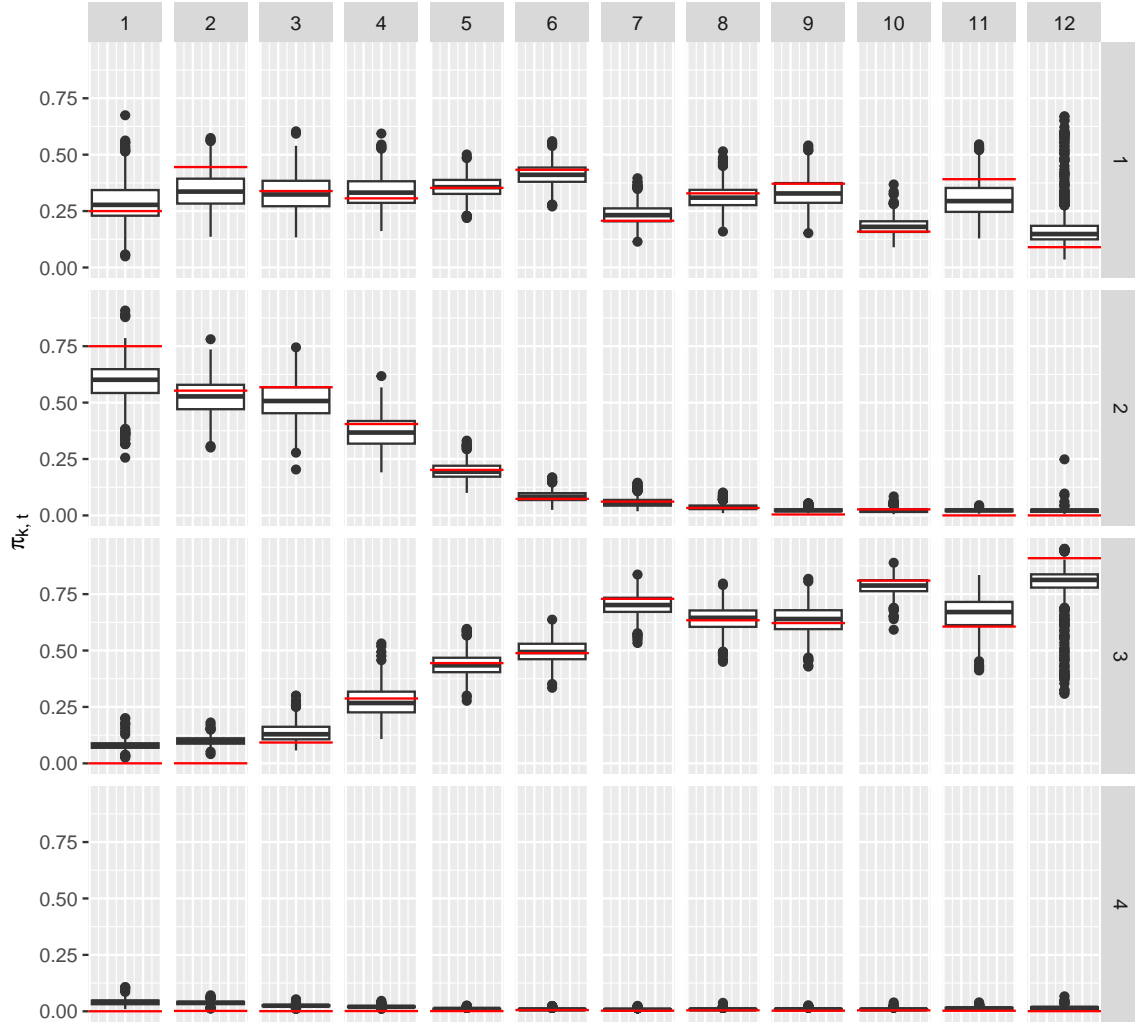


Figure E.2: Distribution of the posterior mean for $\pi_{k,t}$ from the approximate reverse Dirichlet-multinomial model with AR(1) prior in the simulation study, for stocks $k = 1$ to 4 in weeks $t = 1$ to 12. Red horizontal lines indicate the true $\pi_{k,t}$ values.

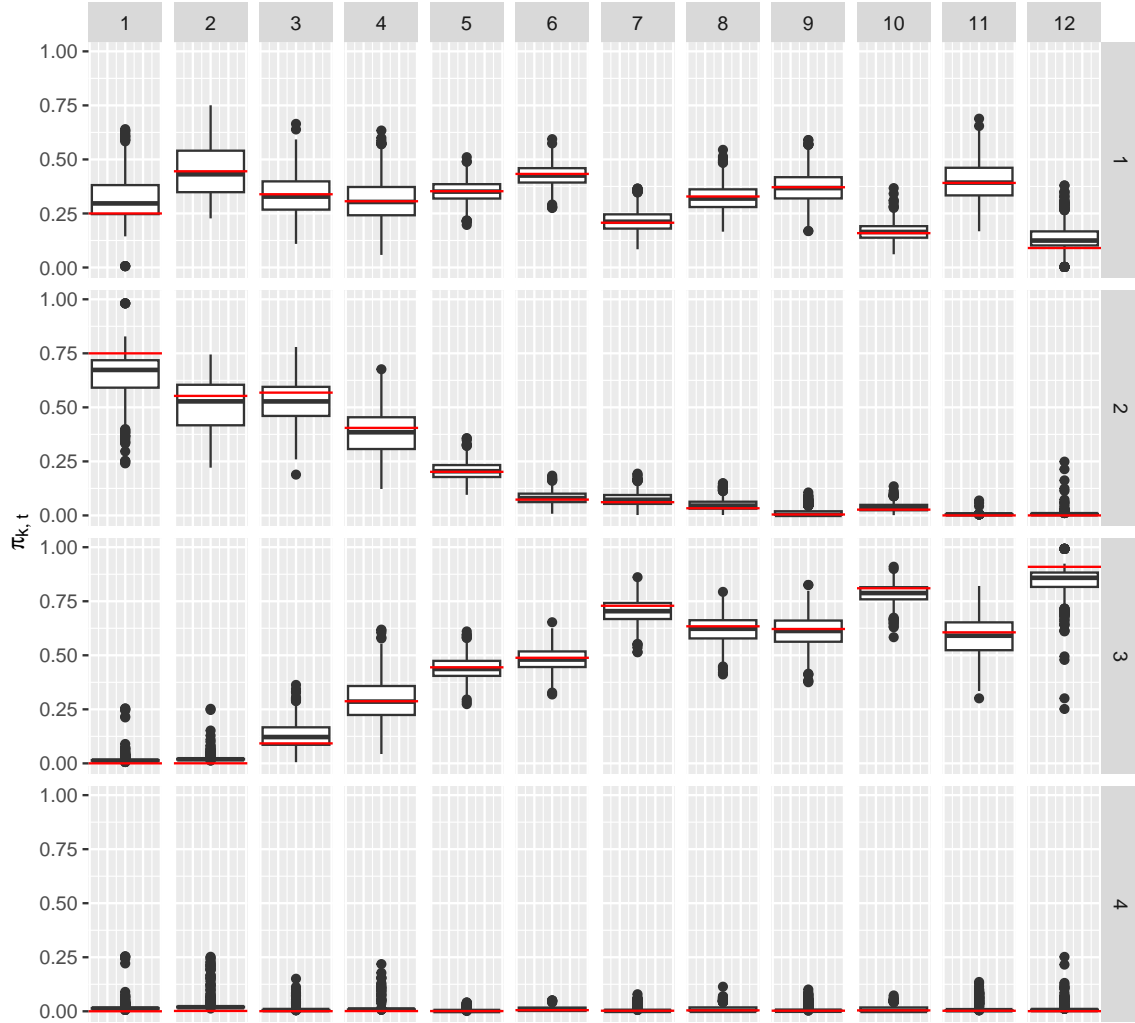


Figure E.3: Distribution of the posterior mean for $\pi_{k,t}$ from the moment-matching Dirichlet model with Dirichlet prior in the simulation study, for stocks $k = 1$ to 4 in weeks $t = 1$ to 12. Red horizontal lines indicate the true $\pi_{k,t}$ values.

Appendix F

CLT for the PQ Estimator

Let g be a real-valued function on the state space of a Markov chain $\{X_n\}$. The following theorem is adapted from [Geyer \(1992\)](#)'s Theorem 2.1.

Theorem 1 *If $\{X_n\}$ is a stationary, irreducible, reversible Markov chain and $\text{Var}\{g(X)\} < \infty$, then as $n \rightarrow \infty$,*

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n g(X_i) - E\{g(X)\} \right) \xrightarrow{d} N(0, \sigma_{CLT}^2). \quad (\text{F.1})$$

Theorem 1 presents a version of the Markov chain central limit theorem (CLT). For further details on the Markov chain CLT under various conditions, such as ergodicity, ergodicity of degree 2, uniform ergodicity, geometric ergodicity, and Harris ergodicity, refer to [Tierney \(1994\)](#); [Chan and Geyer \(1994\)](#); [Roberts and Rosenthal \(1997, 2004\)](#); [Jones \(2004\)](#). [Geyer \(1998\)](#) provided the expression of σ_{CLT}^2 for stationary Markov chains as

$$\sigma_{CLT}^2 = \text{Var}\{g(X_i)\} + 2 \sum_{k=1}^{\infty} \text{Cov}\{g(X_i), g(X_{i+k})\}, \quad (\text{F.2})$$

which does not depend on the value of i .

Suppose $\mathbf{g}(x)$ is a vector with components $g_r(x)$, $r = 1, \dots, R$. [Geyer \(2011\)](#) described the multivariate Markov chain CLT as follows.

Theorem 2 *If $\{X_n\}$ is a stationary, irreducible, reversible Markov chain and $\text{Var}\{g_r(X)\} < \infty$ for $r = 1, \dots, R$, then as $n \rightarrow \infty$,*

$$\sqrt{n} \left(\begin{pmatrix} \frac{1}{n} \sum_{i=1}^n g_1(X_i) \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n g_R(X_i) \end{pmatrix} - \begin{pmatrix} E\{g_1(X)\} \\ \vdots \\ E\{g_R(X)\} \end{pmatrix} \right) \xrightarrow{d} N(0, \Sigma_{CLT}), \quad (\text{F.3})$$

where

$$\Sigma_{CLT} = \text{Var}\{\mathbf{g}(X_i)\} + 2 \sum_{k=1}^{\infty} \text{Cov}\{\mathbf{g}(X_i), \mathbf{g}(X_{i+k})\}. \quad (\text{F.4})$$

Though Equation (F.4) resembles Equation (F.2), the main difference is that in Equation (F.4), $\text{Var}\{\mathbf{g}(X_i)\}$ denotes the matrix with components $\text{Cov}\{g_r(X_i), g_s(X_i)\}$ and $\text{Cov}\{\mathbf{g}(X_i), \mathbf{g}(X_{i+k})\}$ denotes the matrix with components $\text{Cov}\{g_r(X_i), g_s(X_{i+k})\}$.

Consider $g_1(x) = x$ and $g_2(x) = x^2$. Based on the strong law of large numbers for Markov chains (Meyn and Tweedie, 2009), we have

$$\begin{aligned} \bar{X}_n &= \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{a.s.} E(X_i) = \mu_X, \\ \overline{X_n^2} &= \frac{1}{n} \sum_{i=1}^n X_i^2 \xrightarrow{a.s.} E(X_i^2) = \mu_X^2 + \sigma_X^2, \end{aligned}$$

and

$$S_n \xrightarrow{a.s.} \sqrt{E(X_i^2) - E(X_i)^2} = \sigma_X,$$

where μ_X and σ_X^2 denote the true posterior mean and variance, respectively. Then by Theorem 2 we have

$$\sqrt{n} \left\{ \begin{pmatrix} \bar{X}_n \\ \overline{X_n^2} \end{pmatrix} - \begin{pmatrix} \mu_X \\ \mu_X^2 + \sigma_X^2 \end{pmatrix} \right\} \xrightarrow{d} N(\mathbf{0}, \Sigma_{PQ}).$$

Referring to Equation (F.4), we can express the components of Σ_{PQ} as presented in Equations (4.1), (4.2) and (4.3).

Appendix G

Estimating Leisler’s Bats Population Using Jolly-Seber and Data Augmentation

In this appendix, we detail the Jolly-Seber model through superpopulation parameterization used in [Kéry and Schaub \(2011, Chapter 10\)](#) to analyze the Leisler’s Bats dataset. This approach involves data augmentation, where “all-zero” encounter histories are added to account for individuals who were never observed during the study period.

G.1 Notations

We first define the key notations of the model:

- T : Number of sampling occasions
- N_s : Superpopulation size (total number of individuals ever alive during the study)
- M : Size of the augmented dataset, of which N_s are genuine and the rest are pseudo-individuals
- $z_{i,t}$: Latent state of individual i at occasion t (1 if alive and present in the population, 0 otherwise)
- $y_{i,t}$: Observation indicator (1 if individual i is captured at occasion t , 0 otherwise)

- $p_{i,t}$: Capture probability for individual i at occasion t
- $\phi_{i,t}$: Survival probability for individual i at occasion t
- γ_t : Removal entry probability at occasion t , which is the probability that an available individual in M enters the population at occasion t
- B_t : Number of individuals entering the population at occasion t

G.2 Model Specification

G.2.1 Prior

For the survival probability $\phi_{i,t}$, we set its prior based on a mean survival probability $\bar{\phi}$ throughout the study period:

$$\begin{aligned}\text{logit}(\phi_{i,t}) &= \text{logit}(\bar{\phi}) + \epsilon_t \\ \bar{\phi} &\sim \text{Uniform}(0,1) \\ \epsilon_t &\sim N(0, \sigma^2) \\ \sigma &\sim U(0,5)\end{aligned}$$

For the capture probability $p_{i,t}$, we set the prior assuming all weeks have the same capture probability:

$$\begin{aligned}p_{i,t} &= \bar{p} \\ \bar{p} &\sim \text{Uniform}(0,1)\end{aligned}$$

For the removal entry probability γ_t , we use a uniform prior: $\gamma_t \sim \text{Uniform}(0,1)$.

G.2.2 Likelihood

The likelihood of the observed data given the model parameters is defined as follows:

1. Initial State:

$$z_{i,1} \sim \text{Bernoulli}(\gamma_1)$$

2. State Transition:

$$z_{i,t+1} \mid z_{i,t}, \dots, z_{i,1} \sim \text{Bernoulli}(z_{i,t}\phi_{i,t} + \gamma_{t+1} \prod_{k=1}^t (1 - z_{i,k}))$$

3. Observation Process:

$$y_{i,t} \mid z_{i,t} \sim \text{Bernoulli}(z_{i,t}p_{i,t})$$

4. Recruitment:

$$B_t = \sum_{i=1}^M (1 - z_{i,t-1})z_{i,t}$$

5. Superpopulation Size:

$$N_s = \sum_{t=1}^T B_t$$

Appendix H

Supplementary Figures for Chapter 4

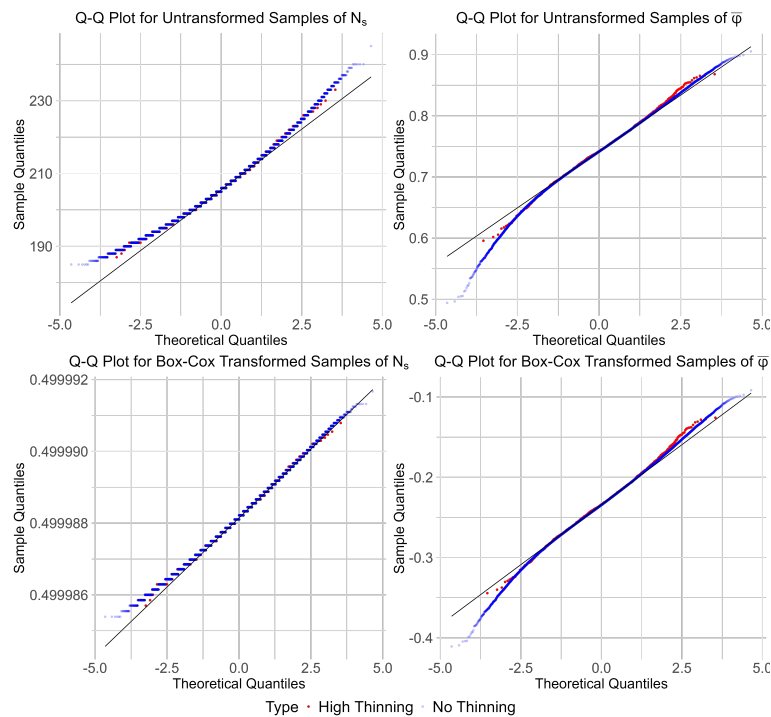


Figure H.1: Q-Q plots for N_s and $\bar{\phi}$ under different thinning and transformation scenarios. “High thinning” refers to running 100,000 iterations per chain and retaining a total of 2,500 samples. “No Thinning” refers to running 100,000 iterations per chain and retaining all the samples.

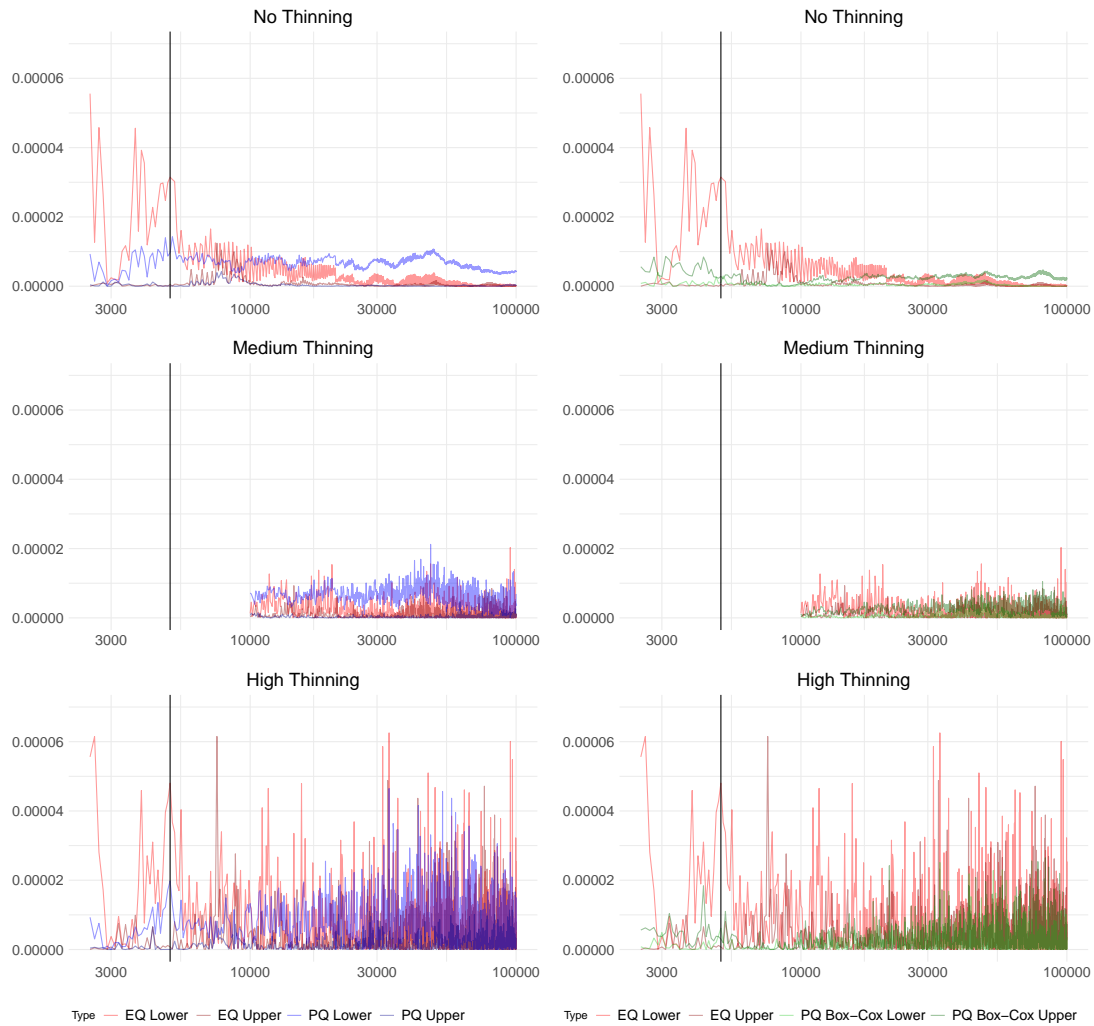


Figure H.2: Squared error in the 0.025 and 0.975 quantile estimates for $\bar{\phi}$ compared to the EQ estimate with full set of samples. The red and dark red lines represent EQ estimates for the 0.025 and 0.975 quantiles, respectively. The blue and dark blue lines represent PQ estimates for the 0.025 and 0.975 quantiles, respectively. The green and dark green lines represent PQ estimates for the 0.025 and 0.975 quantiles after Box-Cox transformation. The x-axis, on a \log_{10} scale, shows iterations from 2,500 to 100,000. The black vertical line indicates the 5,000th iteration, corresponding to the chain length used by [Kéry and Schaub \(2011\)](#).

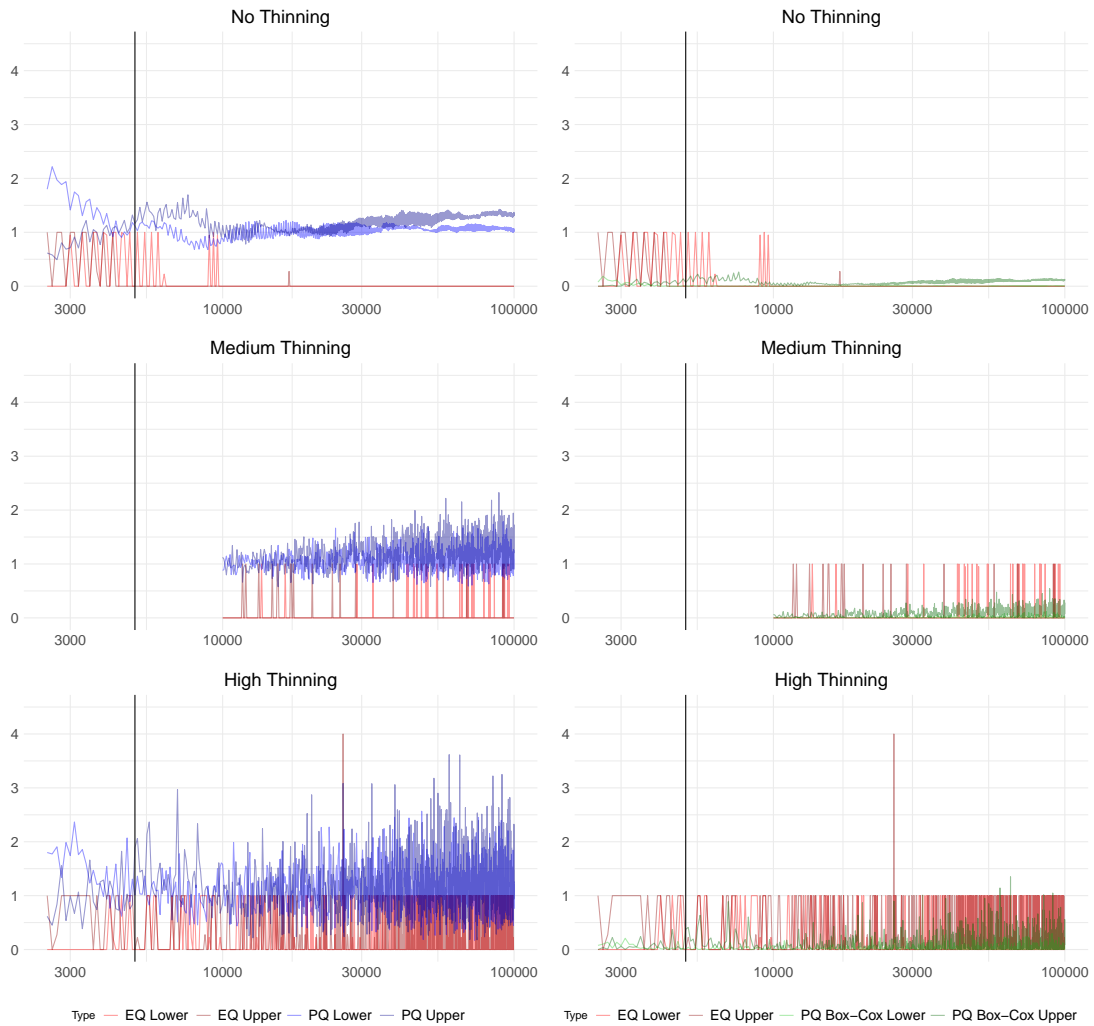


Figure H.3: Squared error in the 0.025 and 0.975 quantile estimates for N_s compared to the EQ estimate with full set of samples. The red and dark red lines represent EQ estimates for the 0.025 and 0.975 quantiles, respectively. The blue and dark blue lines represent PQ estimates for the 0.025 and 0.975 quantiles, respectively. The green and dark green lines represent PQ estimates for the 0.025 and 0.975 quantiles after Box-Cox transformation. The x-axis, on a \log_{10} scale, shows iterations from 2,500 to 100,000. The black vertical line indicates the 5,000th iteration, corresponding to the chain length used by [Kéry and Schaub \(2011\)](#).