COUNTING ALL THE IMAGINARY FISH AND MORE

by

Jonathan Babyn

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
November 2023

One Fish, two fish, red fish, blue fish - Dr. Seuss

Contents

List of Tables

List of Figures

Abstract

Avoiding overexploitation of marine resources requires being able to accurately estimate the size and health of a population. This thesis presents methods that improve upon existing fisheries stock assessment methods. Having an estimate of the ages of fish in a sample can simplify stock assessment model development and allow more insight into stock structure. We first present a spatial Age Length Key (ALK) model that accommodates physical barriers to fish movement such as islands or bays. By incorporating spatial information and accounting for barriers in the construction of ALKs more accurate estimates of age can be obtained. We then turn our attention to effective population size, a concept used by conservationists and geneticists to summarize the overall genetic health of a population by giving the size of the population under the "ideal" Wright-Fisher model. We show how using CKMR alone it is possible to estimate the effective population size as well as the variance in number of offspring enabling new insights into the population. Finally, we examine the applicability of CKMR to a population similar to the Sable Island grey seal colony through individual based simulation.

List of Abbreviations and Symbols Used

AD Automatic Differention.

ADMB AD Model Builder.

AIC Akaike Information Criterion.

ALK Age Length Key.

CART Classification and Regression Tree.

CKMR Close-kin Mark-Recapture.

CMP Conway-Maxwell Poisson.

CRL Continuation Ratio Logit.

CV cross validation.

DFO Fisheries and Oceans Canada.

FEM Finite Element Method.

FOAM Female Only Abundance Model.

GAM Generalized Additive Model.

GCV Generalized Cross Validation.

GF Gaussian Random Field.

GFB Gaussian field model with barrier support.

GMRF Gaussian Markov Random Field.

GPGCP Grandparent-Grandchild Pair.

HS Half-sibling.

HSP Half-Sibling Pair.

ICES International Council for the Exploration of the Sea.

IPM Integrated Population Model.

LD Linkage Disequilibrium.

MLE Maximum Likelihood Estimation.

MR Mark-Recapture.

NAFO Northwest Atlantic Fisheries Organization.

PMF Probability Mass Function.

PO Parent-Offspring.

POP Parent-Offspring Pair.

RMSE Root Mean Squared Error.

RV Research Vessel.

SMM Stratified Mean Method.

SPDE Stochastic Partial Differential Equation.

TMB Template Model Builder.

TRO Total Reproductive Output.

Acknowledgements

Thanks to everyone that helped along the way.

Chapter 1

Introduction

In the inaugural address of the 1883 Fisheries Exhibition in London the English biologist Thomas Huxley said "I believe, then, that the cod fishery, the herring fishery, the pilchard fishery, the mackerel fishery, and probably all the great sea fisheries, are inexhaustible; that is to say, that nothing we do seriously affects the number of the fish. And any attempt to regulate these fisheries seems consequently, from the nature of the case, to be useless" (Huxley 1883). Time of course proved Huxley wrong, in the early 1990s most of the cod fisheries off the coast of Newfoundland collapsed in part due to overfishing, and remain closed to fishing to this day over thirty years on (Myers et al. 1997; Verma 2022). More recently the stock of Atlantic mackerel off the coast of Nova Scotia was closed due to concerns it was depleted (*Atlantic Mackerel Commercial Fishery and Bait Closure* 2022). To maintain fishery populations in a sustainable way it is crucial that we understand their key aspects such as total abundance, growth rate, recruitment (number of new individuals entering the fishery), and age composition.

Fisheries science combines elements of ecology, mathematics, statistics, and marine biology in an attempt to paint an accurate picture of the health of fish stocks and avoid their over-harvesting. Fisheries science began to develop over the course of the late 19th and 20th centuries as governments around the world started taking notice of declining fish populations. This led to the foundation of societies like the U.S Fish Commission and the International Council for the Exploration of the Sea (ICES) with the purposes of managing and investigating fisheries. Over time researchers began to develop tools to answer key questions such as the work of Hoffbauer

in the 1890s who noticed that rings develop annually on pond carp scales providing a proficient aging tool, or Reibisch who noted that the bones of some fish are laid down in layers, again enabling estimates of age to be found (T. D. Smith 1994). Over the course of the 20th century mathematical modelling and statistics started to play a larger role in helping to answer some of the questions posed by fisheries scientists. This research led to growth models as developed by Von Bertalanffy (1938), tools for estimating abundance like virtual population analysis developed by Gulland (1965), further development on Mark-Recapture (MR) by Chapman (1951), stock-recruitment relationships as suggested by Beverton and Holt (2012), among others. These methods tended to be deterministic in nature rather than statistical and gave no consideration to observational error present in the data. As the course of the 20th century progressed stochastic elements were introduced into the methods used by fishery scientists thanks to work by individuals like Ricker and Doubleday (Quinn 2003). Fournier and Archibald (1982) started the modern practice of integrated analysis for fisheries stock assessment modelling, where data in the model are kept as raw as possible to try and let the model capture as much uncertainty as possible through the use of joint likelihood functions. State space stock assessment models started to appear in the early 1990s through the works of Sullivan (1992) and Gudmundsson (1994) allowing for better capture of unobserved processes like fishing selectivity through the use of random effects. New tools like AD Model Builder (ADMB) and Template Model Builder (TMB) have simplified the process of developing state space stock assessment models and have led to an increasing number being in use (Nielsen and Berg 2014; Miller and Stock 2020; Cadigan 2015).

A contributing factor to the collapse of the cod stocks off the coast of Newfoundland was that the stock assessment methods in use at the time were overestimating abundance while underestimating fishing mortality (Myers et al. 1997). Had more accurate methods been available then perhaps it could have been avoided. There have

been calls for advancements in stock assessment models and methods to incorporate new sources of information with the hope of uncovering new insights and resolving potential biases (Quinn 2003; Punt, Dunn, et al. 2020). It has been recognized that stocks may not be spatially homogeneous across their management area and it may be necessary and worth the increased complexity of incorporating spatial information to account for this variability (Punt, Dunn, et al. 2020). Genetic data are another new potential source of information for stock assessment that could allow for insights into things like the familial structure of the population (Quinn 2003) and enable new techniques (Bravington et al. 2016). This thesis develops stock assessment methods that incorporate data from both spatial and genetic sources to glean new insights into populations, test new approaches and improve their accuracy.

1.1 Age Structured Population Dynamics

Age structured population dynamics models provide a compromise between the over-simplicity of models that do not differentiate between individuals and more complex length or stage based ones (Quinn 2003). Age-structured methods require an estimate of age to be known. For fish this is often done by taking a subsample and examining a hard part such as a scale or otolith (ear bone) in a lab and then using an ALK or some other method (Aanes and Vølstad 2015). While the concepts reviewed here are for age structured populations they can easily be extended to length, two sex or stage-based ones.

The numbers of individuals at age in population at time t, or numbers at age at

time t can be represented as the vector,

$$\boldsymbol{N}_t = \begin{bmatrix} N_{1,t} \\ N_{2,t} \\ N_{3,t} \\ \vdots \\ N_{A-1,t} \\ N_{A,t} \end{bmatrix}$$

where $N_{i,t}$ is the number of individuals in the ith age class at time t. In a discrete time setting the numbers at age in time $t+1$ can be found through

$$\boldsymbol{N}_{t+1} = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 & \dots & \beta_{A-1} & \beta_A \\ \phi_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \phi_2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \phi_3 & \dots & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & & \dots & \phi_{A-1} & 0 \end{bmatrix} \begin{bmatrix} N_{1,t} \\ N_{2,t} \\ N_{3,t} \\ \vdots \\ N_{A-1,t} \\ N_{A,t} \end{bmatrix}.$$

The leading matrix here is referred to as the Leslie matrix, the top row β_is being the average per captia fecundity produced by individuals aged i and the off diagonal ϕ_i terms represents the proportion of individuals aged i surviving to $i+1$ (Caswell 2000). In fish populations fecundity can often increase with size (thus indirectly with age) and capturing this can better reflect the population (Quinn 2003). The Leslie matrix can also be turned into a stochastic matrix by having the terms inside be represented by random functions that could potentially vary in time (Caswell 2000).

The above representation is essentially a discrete time matrix version of the cohort equations commonly used in stock assessment models (Miller and Stock 2020; Nielsen

and Berg 2014; Cadigan 2015),

$$N_{a,t+1} = N_{a,t}e^{-Z_{a,t}} \tag{1.1}$$

where $Z_{a,t}$ is the total mortality rate of individuals aged a in time t. This can easily be seen by realizing that $e^{-Z_{a,t}}$ is just the proportion of individuals aged a that survive to the next time step. However, rather than directly get the numbers of individuals in the youngest age class represented (recruits) from fecundity directly a recruitment process or function may be used (Miller and Stock 2020; Nielsen and Berg 2014; Cadigan 2015).

1.2 Close-kin Mark-Recapture

This thesis relies heavily on the work of Julius Skaug (2001) which introduced the concepts behind CKMR. Replacing the physical tags or marks of MR with the kinship relationship between pairs of individuals such whether they are a Parent-Offspring Pair (POP), HSP, unrelated or some other relationship status allowing for estimates of adult abundance, survival and fecundity (Bravington et al. 2016). As developments in genotyping have driven down prices the technique has started to see more traction and use across a range of species ranging from mosquitoes, bats, southern bluefin tuna and more (Bravington et al. 2016; Sharma et al. 2022; Lloyd-Jones et al. 2023). While the work presented here assumes that kinship status individuals is known exactly we review some concepts and terminology to understand how they may be found in practice. MR has been used for decades by fisheries scientists and biologists to estimate values like abundance and survival. At it's simplest, MR involves sampling individuals from a closed population (no deaths or immigration) and marking them in some way, e.g., with a tag or brand, releasing them back into the population then after sufficient time for the marked individuals to mix back into the population,

taking a second independent sample of the population. By counting the the number of marked individuals found in the second sample an estimate of the total population abundance, $\hat{N}$, can be found using the Lincoln-Petersen estimator,

$$\hat{N} = \frac{n_1 n_2}{m}$$

where n_1 and n_2 are the number of individuals sampled in the first and second sample respectively and m is the number of marked individuals found in the second sample (Lohr 2009). MR has been extended to work in scenarios where there are more than two samples (Darroch 1958), the population is no longer closed (Darroch 1959; Jolly 1965; Seber 1965), unequal capture probabilities (Pollock 1982), among others which have allowed for terms such as survival to be estimated alongside abundance.

As previously mentioned above, CKMR replaces the physical tags in MR with the kinship relationship between individuals. The CKMR analogue to the MR Lincoln-Petersen estimator given above would be to take a sample of adults, n_a, and juveniles, n_j, and count how many POPs are found between the two samples as H then the adult abundance in the population is

$$\hat{N}_a = \frac{2 n_j n_a}{H}$$

where the two is an adjustment for the fact each juvenile has two parents (Bravington et al. 2016).

In wildlife populations determining the kinship relationship requires using the properties of genetics and we review some of the terminology here. Most animals are diploids and receive one copy of each gene from their mother and another copy of each gene from their father. The location of a gene is called the locus. Within a population there can be multiple different variants of any given gene, referred to as alleles. The

pair of alleles that individuals inherit at each locus make up that individuals genotype. In practice the entire genome of an individual is not typically sampled but instead select loci of interest called markers. Determining what genotypes an individual has at the sampled markers is the process of genotyping.

Most implementations of CKMR have taken a pairwise approach to kinship finding (Bravington et al. 2016; Thomson et al. 2020; Lloyd-Jones et al. 2023) which while potentially less accurate than methods that consider all sampled individuals simultaneously (Wang 2004) is considerably more performant especially on the sample sizes necessary for CKMR (Bravington et al. 2016). To keep things relatively simple we will ignore issues like genotyping errors, mutations and assume genes are unlinked.

Following Thompson (1975), suppose we have two individuals A and B and we wish to know their relationship $\aleph(A, B)$. Each pair of individuals can have either zero, one or two genes be shared at each locus, we use κ_0, κ_1 and κ_2 to represent the probability of sharing zero, one or two genes respectively. Table 1.1 gives the values of κ_0, κ_1 and κ_2 for some common kinship relationships. Note that the relationships between individuals are reciprocal and that it's not possible for instance to determine based on genetics alone which individual in a POP is the parent and which is the offspring. The table also illustrates that some relationships like uncle/niece and grandparent/grandchild can have the same values of κ_0, κ_1 and κ_2 meaning it's not possible to distinguish between the two solely using genetics, and in certain scenarios may not be possible to distinguish at all.

If A and B have genotypes G_1 and G_2 at Locus j respectively and r(A,B) is the number of genes shared by A and B then the likelihood of relationship $\aleph$ between A and B is

$$L_j(\aleph) = P[G(A) = G_1, G(B) = G_2 | \aleph(A, B) = \aleph]$$

Table 1.1: Different possible relationships of A and B. κ_0,κ_1 and κ_2 are the probabilities of sharing 0, 1 or two genes (Thompson 1975).

Relationship of A to B ($\aleph$)	Code Letter	κ_0	κ_1	κ_2
Unrelated	U	1	0	0
Offspring, parent	Q	0	1	0
Full Sibling	B	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Self, Identical twin	R	0	0	1
Niece,Nephew, Uncle	N	$\frac{1}{2}$	$\frac{1}{2}$	0
Grandparent, grandchild	N	$\frac{1}{2}$	$\frac{1}{2}$	0
Half-sib	N	$\frac{1}{2}$	$\frac{1}{2}$	0
First Cousin, Parent's Half-sib, Half-sib's Child	C	$\frac{3}{4}$	$\frac{1}{4}$	0
Double First Cousin	D	$\frac{9}{16}$	$\frac{6}{16}$	$\frac{1}{16}$
Half-sibs whose parents are sibs or PO	NB,NQ	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$
Half-sibs whose parents are half-sibs	NN	$\frac{7}{16}$	$\frac{1}{2}$	$\frac{1}{16}$

$$= \sum_{i=0}^{2} P[G(A) = G_1, G(B) = G_2 | r(A,B) = i] P[r(A,B) = i | \aleph(A,B) = \aleph]$$

$$= \sum_{i=0}^{2} \kappa_i P_i(G_1, G_2)$$

where

$$P_i(G_1, G_2) = P[G(A) = G_1, G(B) = G_2 | r(A,B) = i]$$

Then since we have assumed that the markers are unlinked, the likelihood of being relationship $\aleph$ over all s loci is just

$$L_s(\aleph) = \prod_{j=1}^{s} (L_j(\aleph))$$

We can then use the likelihood-ratio test to determine the likelihood of a pair of individuals being one kinship relationship versus another. For instance if we are interested in comparing if a pair of individuals are kinship relationship $\aleph_1$ versus $\aleph_2$ using

$$\Lambda(r(A,B) = i) = \frac{L_s(\aleph_1)}{L_s(\aleph_2)}.$$

We can then apply the above process or similar to individuals that we have sampled in a pairwise fashion to determine the kinship relationships that we are interested in. These then form our observed kinship pairs which can be used in a CKMR estimator or model to estimate terms like adult abundance, survival and fecundity (Bravington et al. 2016). Most applications of CKMR have taken a pseudo-likelihood (also known as composite or quasi-likelihood) approach to estimating the parameters due to the complexity of considering the joint distribution for the kinship relationships among all sampled individuals (Bravington et al. 2016; Lloyd-Jones et al. 2023; Thomson et al. 2020; Sharma et al. 2022). Instead, only the kinship relationships between pairs of individuals are considered. Further details on the pseudo-likelihood and how the observed kinship pairs are used to estimate parameters are given in Chapters 3 and 4.

The above has focused on estimating the abundance or population size of the population, that is how many individuals are there in the population. Chapter 4 of this thesis deals with the concept of *effective population size* often denoted as N_e. Effective population size is the size of the population if it were to follow a Wright-Fisher model, which is how big the population would be if there were discrete generations, random mating and Poisson distributed numbers of offspring. Effective population size can be defined in terms of the rate of inbreeding or genetic drift occurring in the population and because of this N_e can help serve as an indication of how they might be impacting the population. For some populations N_e calculated from the perspective of inbreeding or genetic drift will be the same but this is not always the case and how to find N_e will depend on the population in question (Crow, Kimura, et al. 1970; Felsenstein 2005).

1.3 Overview

Chapter 2 presents a model for spatial ALKs that supports physical barriers using an approximation to a GF. By incorporating catch locations, and accounting for the movement of fish caused by barriers such as islands or bays, we demonstrate a way to improve the accuracy of age estimation. Chapter 3 demonstrates the link between N_e and CKMR using sibling comparisons from the same cohort. There we show how to estimate N_e and the variance in number of offspring using CKMR. Chapter 4 is a simulation based evaluation of how well CKMR might fare on a grey seal population similar to the one present on Sable Island with particular emphasis on a juvenile only sampling scheme. Chapter 5 reports our conclusions.

Chapter 2

A Gaussian Field Approach to Generating Spatial Age Length Keys

2.1 Introduction

Stock assessments are tools that can allow us to understand the overall health of a fish stock. They enable quantifying the abundance, age and length compositions of the population and determine indications of whether the stock is facing overexploitation (Worm et al. 2009). They play a key part in helping to rebuild and maintain fisheries around the world (Worm et al. 2009; Hilborn and Ovando 2014). Stock assessment methods have evolved from simple methods based only on catch data to models that integrate additional sources of data, to modern state-space approaches (Aeberhard et al. 2018) that allow for increasing levels of inference and precision.

Age structured methods can greatly simplify stock assessment models as ages link directly to the numbers of survivors in each year. However, for most species accurately and easily determining the age of a fish can be a time consuming and expensive process that often requires an expert counting the number of growth rings on an otolith or similar procedure. Measuring the length of a fish is much easier, less lethal and it can be done on site for low cost. In order to take advantage of the benefits of age structured methods, approaches for estimating the age of a fish from its cheaply measured length are commonly used, like Age Length Keys (ALKs) (Aanes and Vølstad 2015).

ALKs have been used to estimate the age of fish for over 80 years (Fridriksson

1934). They are based on the idea that the proportion of fish at age a, p_a is equal to

$$p_a = \sum_{i=1}^{K} k_i p_{a|i} \tag{2.1}$$

where i indexes discrete length bins $i = 1$ to K and k_i is the proportion of fish in length bin i, $p_{a|i}$ is the observed conditional probability (or proportion) of being age a given membership in length bin i. An ALK is then simply a matrix of proportions of fish at age a given length i. To convert the sampled length frequencies to estimates at age, the length frequencies are multiplied by the ALK to get the numbers at age. An example of a traditional ALK along with an example of a smooth model based ALK is shown in Table 2.1.

ALKs are often constructed separately for different covariates such as sex, time of year and gear type depending on the species and application. Traditional ALKs can easily suffer from data gaps resulting in some fish not being assigned an age estimate. ALKs also suffer from low sample numbers particularly for rarer older age classes. An example of such a sampling artifact can be seen in Table 2.1a where any fish assigned to the 49 cm length bin will automatically be assumed to be age 9 despite the existence of shorter older fish.

Smooth ALKs have non-zero proportions at every possible length bin ensuring that all fish are assigned an age estimate. Kvist et al. (2000) suggested the possibility of generating smooth ALKs using a Continuation Ratio Logit (CRL) model. Smooth ALKs have been applied before to lesser sandeel, (Rindorf and Lewy 2001) North Sea haddock, (Stari et al. 2010; Berg and Kristensen 2012) cod, herring and whiting (Berg and Kristensen 2012) . They can also mitigate the effects of the sampling artifacts mentioned above, demonstrated in Table 2.1b where the probability of being a particular age given a length of 49cm is more suitably spread across nearby age

classes and not just associated with age nine.

Many fish species are known to aggregate in different locations and may be in different spots at different stages of their lifecycles (Parrish 1999). They may also have differing levels of growth depending on location (Punt, Haddon, et al. 2015). Incorporating spatial information into traditional ALKs requires dividing the study area into subareas. The greater the number of subareas the sparser the data, the more likely gaps and other issues are to arise in the ALK resulting in more missed or incorrectly aged fish. Berg and Kristensen 2012 presented a way of constructing ALKs for point referenced data using a Generalized Additive Model (GAM) and thin-plate regression splines. They found better internal and external consistencies for age based survey indices when using spatial ALKs, in addition to observing differences in ALKs constructed in different areas.

However, the method of Berg and Kristensen (2012) does not account for boundaries posed by physical barriers such as landmasses that may be present in the study area. These oversimplifications result in predictions that smooth under landmasses. That is, the spatial part of the model will ignore any landmasses and predicted probabilities will ignore marine distance which may result in poor estimates for samples on opposite sides of a large landmass for instance. This could be a problem if for example a large group of young fish inhabits a bay as smoothing under landmasses may artificially increase the probability of young fish living on the other side of the bay.

The spatial ALK approach presented here addresses the problem of smoothing over landmasses by using an approximation of a Gaussian Random Field (GF) that has support for physical barriers (Bakka, Vanhatalo, J. B. Illian, et al. 2019) . Desirably it still allows ALKs to be constructed at any location within the study area. Approximations of GFs have previously been proposed to help model spatial indices

of abundance (Thorson, Shelton, et al. 2015; Thorson and Barnett 2017), species distribution (Bakka, Vanhatalo, J. Illian, et al. 2016) along with other non-marine uses such as global temperature data (Lindgren and Rue 2011).

In Section 2.2 our spatial ALK is fully described and all necessary background knowledge provided. In Section 2.3.1 the proposed Gaussian field model with barrier support (GFB) is tested using simulated survey data to determine the benefit of using a spatial ALK instead of one that ignores all spatial structure. It is tested alongside a traditional ALK, a non-spatial CRL model and a spatial GAM ALK implementation. Finally in Section 2.3.2, the four methods are applied to two real datasets (Cod and American Plaice) from Fisheries and Oceans Canada (DFO) 's multi-species bottom trawl Research Vessel (RV) survey. The method proposed is implemented as an `R` package called `barrierALK` and is available on Github `https://github.com/jgbabyn/barrierALK`.

Table 2.1: Examples of traditional and smoothed ALKs. Columns represent ages and rows define three centimetre length bins. Zeros are omitted for readability.

	1	2	3	4	5	6	7	8	9	10
1										
4										
7	1.000									
10	0.933	0.067								
13	0.200	0.767	0.033							
16		0.567	0.433							
19		0.100	0.700	0.200						
22			0.333	0.533	0.133					
25			0.033	0.433	0.467	0.067				
28				0.200	0.400	0.367	0.033			
31				0.033	0.167	0.467	0.333			
34					0.033	0.367	0.433	0.100	0.067	
37					0.067	0.100	0.400	0.267	0.167	
40						0.036	0.107	0.393	0.357	0.107
43						0.071		0.357	0.500	0.071
46								0.286	0.429	0.286
49									1.000	
52										
55										

(a) A traditional ALK. Any fish measured to be less than 4cm or greater than 52cm would by missed by this ALK.

	1	2	3	4	5	6	7	8	9	10
1	1.000									
4	1.000									
7	1.000									
10	0.990	0.010								
13	0.064	0.918	0.018							
16		0.612	0.382	0.006						
19		0.047	0.798	0.150	0.004					
22		0.002	0.290	0.602	0.101	0.005				
25			0.031	0.473	0.408	0.083	0.005			
28			0.003	0.139	0.455	0.337	0.065	0.001		
31				0.027	0.209	0.476	0.273	0.010	0.005	
34				0.005	0.062	0.326	0.475	0.081	0.048	0.003
37				0.001	0.016	0.146	0.346	0.265	0.203	0.023
40					0.004	0.053	0.115	0.383	0.378	0.067
43					0.001	0.018	0.026	0.370	0.458	0.127
46						0.006	0.005	0.314	0.471	0.203
49						0.002	0.001	0.255	0.443	0.299
52						0.001		0.201	0.389	0.409
55								0.157	0.319	0.524

(b) An example of a smooth ALK. This ALK was constructed from a CRL model with length as the sole covariate using the same age-growth data as was used to construct the ALK in Table 1a. All lengths are represented and there is no longer any possibility of missing fish at the more extreme length bins. Effects of sampling artifacts like all 49cm fish being considered age 9 despite shorter fish being age 10 are reduced. Zeros are again omitted for readability.

2.2 Methods

2.2.1 Ordinal Regression and Continuation Ratio Logits

An ALK can be constructed using any classifier capable of handling multiple age classes and returning probabilities of a fish with a particular set of covariates (e.g., male, caught using gillnet, etc.) belonging to each age class. The ALK is simply those probabilities for every set of observed covariates. Ages naturally have an ordering associated with them, a seven-year-old fish must have first been a six year old fish and before that a five-year-old fish and so on. Ordinal regression can handle ordered categorical data and return class probabilities (Agresti 2003). It is also possible to incorporate spatial structure directly into ordinal regression models through the use of splines or random fields. This is not the case for some alternative multi-class classifiers like Classification and Regression Trees (CARTs). A number of different ordinal regression methods exist such as cumulative logits, adjacent-category logits, etc. Continuation Ratio Logits (CRLs) are one method that offers a few advantages (Agresti 2003) (Harrell 2014).

CRLs models have the advantage over other ordinal regression methods in that it is very easy to remove or loosen the proportional odds assumption. This allows for all covariates or some subset of covariates to be able to freely vary with every level of a category (Harrell 2014). In addition CRL models can be represented using $\mathcal{A} - 1$ binomial models which allows their fitting using any software capable of performing logistic regression (Agresti 2003).

Suppose the ith aged fish is observed to be age a, and there are $\mathcal{A}$ total ages. The CRL for the ith observation with the corresponding vector of covariates, $\boldsymbol{X_i}$, that must of course include length along with others (e.g. sex, gear type, etc.), is

$$\text{logit}(\pi_a[\boldsymbol{X}_i]) = P(x_i = a | x_i \geq a) \tag{2.2}$$

where

$$\pi_a[\boldsymbol{X}_i] = \frac{p_a[\boldsymbol{X}_i]}{p_a[\boldsymbol{X}_i] + ... + p_A[\boldsymbol{X}_i]} \tag{2.3}$$

and p_a is the proportion of fish at age a . In other words, the CRL for the ith observation is the probability of being age a given it is at least age a or greater (Agresti 2003).

The unconditional probabilities $P(x_i = a)$ can be found by

$$\begin{cases} \pi_a[\boldsymbol{X}_i], \quad a = R \\ \pi_a[\boldsymbol{X}_i] \sum_R^{a-1}(1 - \pi_i[\boldsymbol{X}_i]) \quad R < a < A \\ 1 - \sum_R^{A-1}(1 - \pi_i[\boldsymbol{X}_i]) \quad a = A \end{cases} \tag{2.4}$$

where R is the first age in the model, and A is the last. In aging data R can refer to the age of recruitment to the survey or fishery, A identifies a plus group or the final age involved (Berg and Kristensen 2012). In the models presented here, length has an individual parameter for each age group which allows for greater flexibility. This is known as relaxing the proportional odds assumption.

2.2.2 Random Fields

Random fields are collections of random variables, $\{X(\boldsymbol{s}), \boldsymbol{s} \in \boldsymbol{S}\}$. $\boldsymbol{S}$ is the set of indices and $\boldsymbol{s}$ is the index (Ross 2014). Typically, a set of indices is a set of locations, but could also incorporate time. The index set $\boldsymbol{S}$ can be a discrete set, continuous, finite or infinite. The random variables in random fields can follow any of the typical distributions used such as Gaussian, Student's t, Gamma, etc. Gaussian Random Fields (GFs) are those in which all the $X(\boldsymbol{s})$ are normally or Gaussian distributed (Rue and Held 2005). A GF can be specified by its mean function $\mu(s)$ and covariance function $\text{Cov}(s, t), \quad s, t \in \boldsymbol{S}$. A popular choice of covariance function for spatial data

is the Matérn covariance function,

$$c(s,t) = \sigma_u^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{8\nu} \frac{||s-t||}{r} \right) K_\nu \left(\sqrt{8\nu} \frac{||s-t||}{r} \right) \tag{2.5}$$

where Γ is the Gamma function, ν is a smoothness parameter, K_ν is the modified Bessel function of the second kind with order ν, r is the range parameter which is the spatial distance when the correlation is ≈ 0.13, σ_u is the marginal standard deviation (Bakka, Rue, et al. 2018) and $||s-t||$ is the distance between two points.

Using a GF directly is not computationally tractable for large problems. The cost to factorize the resulting dense covariance matrix is cubic in time. As a result, a number of alternative approaches to try and get around the high computational cost have been proposed. Gaussian Markov Random Fields (GMRFs) are GFs with the Markov property, that is

$$P(\boldsymbol{s}_i | \{\boldsymbol{s}_j : j \neq i\}) = P(\boldsymbol{s}_i | \boldsymbol{x}_j : j \in \mathcal{N}_i) \tag{2.6}$$

where the neighbours $\mathcal{N}_i$ to the location $\boldsymbol{s}_i$ are the points $\{\boldsymbol{s}_j, j \in \mathcal{N}_i\}$ that are close or connected to $\boldsymbol{s}_i$. Conditional on its neighbours the mean at a location is independent of all other locations. The Markov property ensures that the precision matrix (inverse of the covariance matrix) is sparse. The sparsity of the precision matrix reduces the memory needed and overall computational time required (Rue and Held 2005). However GMRFs have traditionally been limited for spatial applications as they require that areas be broken into predefined regions beforehand. This may be difficult to do in practice and it also limits the spatial resolution available.

2.2.2.1 Gaussian Random Field Approximation using Stochastic Partial Differential Equations

Lindgren et al. (2011) found an explicit link between GFs and GMRFs when using a Matérn covariance function. A valid positive semi-definite covariance matrix is the result of the solution of a set of Stochastic Partial Differential Equations (SPDEs), which creates an approximation of a GF using a GMRF. This allows the benefits of modelling as a GF with the computational speed of a GMRF.

The SPDE method requires creating a Delaunay triangulation or mesh of the study area such as through `R-INLA`'s `inla.mesh.2d` function, which can then be used to generate the matrices that are used for the Finite Element Method (FEM) solution to the SPDE. The Matérn field is the solution $u(\boldsymbol{s})$ to the SPDE

$$u(\boldsymbol{s}) - \nabla \cdot \frac{r^2}{8} \nabla u(\boldsymbol{s}) = r\sqrt{\frac{\pi}{2}}\sigma_u \mathcal{W}(\boldsymbol{s}) \tag{2.7}$$

assuming that the smoothness parameter $\nu = 1$, where ∇ is defined as $\left(\frac{\delta}{\delta x}\right)$, r is the range parameter, σ_u is the marginal standard deviation of the model component u, $\mathcal{W}(\boldsymbol{s})$ refers to white noise. The approximation to the spatial GF $\tilde{u}(\boldsymbol{s})$ is distributed $\mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}(\sigma_u, r)^{-1})$ with $\boldsymbol{Q}(\sigma_u, r)$ is the precision matrix that results from the FEM solution to the SPDE with hyperparameters σ_u and r.

The mesh helps ensure that the resulting precision matrix is sparse as well. Every node in the mesh is an element in the resulting covariance matrix. The more nodes in the mesh, the denser it is and the better the approximation to the GF. This comes as a tradeoff, as computational time increases non-linearly with the addition of more nodes.

Recently Bakka et. al (2019) extended the SPDE GF approximation method of Lindgren et. al (2011) to support physical barriers in a spatial GF such as the problem

presented by coastlines. Their method has several advantages over other proposed methods of incorporating boundary information into a spatial model like the soap film smoother proposed by Wood et al. (2008). It is robust to the selection of boundary polygons, takes similar amounts of computational time and is not particularly hard for the applied practitioner to use beyond defining the mesh and barrier polygons (Bakka, Vanhatalo, J. B. Illian, et al. 2019). Under the assumption that the smoothness parameter $\nu = 1$, the barrier Matérn field is the solution $u(\boldsymbol{s})$ to the SPDE

$$u(\boldsymbol{s}) - \nabla \cdot \frac{r^2}{8} \nabla u(\boldsymbol{s}) = r\sqrt{\frac{\pi}{2}}\sigma_u \mathcal{W}(\boldsymbol{s}) \quad \text{for } s \in \Omega_n \tag{2.8}$$

$$u(\boldsymbol{s}) - \nabla \cdot \frac{r_b^2}{8} \nabla u(\boldsymbol{s}) = r_b\sqrt{\frac{\pi}{2}}\sigma_u \mathcal{W}(\boldsymbol{s}) \quad \text{for } s \in \Omega_b \tag{2.9}$$

where Ω_n is the set of nodes outside the boundary, Ω_b is the set of nodes inside the boundary, and r_b is not a new range parameter but instead a predetermined fraction of r. In this case $r_b = \frac{1}{10}r$. This has the effect of essentially making the decorrelation range close to zero for points that fall within the barrier creating the desired boundary properties. Other parameters are the same as described above. Further details on solving the SPDE for the GF approximation can be found in Lindgren and Rue (2011), Bakka, Vanhatalo, J. B. Illian, et al. (2019) and Bakka (2018).

Prediction and fitting of points that do not fall exactly on mesh node are handled by projector matrix $\boldsymbol{A}$. $\boldsymbol{A}$ is also a sparse matrix that has the same number of rows as data being predicted or fit and a column for every node in the mesh. Every row of $\boldsymbol{A}$ has either one or three non-zero entries. If the data point falls exactly at a mesh node then the non-zero entry will be 1 at that node's column. For points not at a mesh node, the three non-zero entries are the distances from the three vertices of the triangle in the mesh that the point lies in. The $\boldsymbol{A}$ matrix is multiplied against the observed random effects for nodes in the mesh and estimates of the random effects at

each point are found and usable in the model (Bakka, Vanhatalo, J. B. Illian, et al. 2019).

2.2.3 GF Spatial Age-Length Key

The GFB model presented here and implemented in `barrierALK` combines the CRL and barrier approach together. This GFB model is

$$\text{logit}(\pi_a[\boldsymbol{X}_i]) = \alpha_a + \beta_a l_i + \xi_{a,\boldsymbol{s}} \tag{2.10}$$

where α_a is the intercept for age a, β_a is the length parameter for age a and $\xi_{a,\boldsymbol{s}}$ is the spatial intercept resulting from the GF at location $\boldsymbol{s}$:

$$\xi_{a,\boldsymbol{s}} = \begin{cases} \text{MVN}\left(\mathbf{0}, \frac{\sigma_u^2}{(1-\varphi_a^2)} c(s,0)\right) & a = 1 \\ \text{MVN}\left(\varphi_a \xi_{a-1,s}, \sigma_u^2 c(s,0)\right) & a > 1. \end{cases} \tag{2.11}$$

The φ_a allows for spatial correlation between age classes, if it exists. What this means in practical terms is that if age structure varies in space, φ_a can measure how correlated that relationship may be.

2.2.4 Estimation

Estimation is performed using the `R` package TMB. TMB uses the Laplace approximation to approximate the integrals in the log likelihood resulting from the random effects that need to be integrated out. TMB uses Automatic Differention (AD) to generate the derivatives for a given objective function which can result in a speed up when paired with an optimizer capable of utilizing derivative information (Kristensen et al. 2016).

When fitting ordinal regression models via Maximum Likelihood Estimation (MLE),

optimization algorithms will often fall into local minima resulting in unrealistic parameter estimates which also has the effect of reducing the predictive accuracy of the model. Penalizing the log likelihood can improve parameter estimates for classes with low numbers of observations (Harrell 2014, p. 323, 209–213). The penalized log likelihood is written as

$$\log L - \frac{1}{2}\lambda \boldsymbol{\beta}' \boldsymbol{P} \boldsymbol{\beta} \tag{2.12}$$

where L is the likelihood from an unpenalized model, $\boldsymbol{\beta}$ being the vector of the fixed effects coefficients, λ the penalty factor chosen by cross validation and $\boldsymbol{P}$ the penalty matrix. Parameters relating to continuous variables are scaled in $\boldsymbol{P}$ by their standard deviation, parameters relating to categorical variables use the penalty function $\sum_i^c (\beta_{fi} - \bar{\beta}_f)^2$ where f is a categorical variable in the model with c levels, $\bar{\beta}_f$ is the mean of all c β_{fi}. This shrinks parameters towards the mean parameter value which avoids biasing towards a specific level (Harrell 2014, p. 209-213) (Verweij and Van Houwelingen 1994).

The optimal value of λ can be chosen by k-fold cross validation (CV). This can be quite time intensive for large spatial models. Due to the presence of random effects in the spatial ALK model, an approach like Generalized Cross Validation (GCV) which would avoid the time consuming k-fold CV can also not be used due to the difficulty in finding an influence matrix if it even exists at all. Instead a modified Akaike Information Criterion (AIC) is proposed here in place of k-fold CV. The modified AIC used here is defined to be

$$\text{LR} \quad \chi^2 - \text{effective degrees of freedom} \tag{2.13}$$

where LR χ^2 is the likelihood ratio value comparing the null model containing only an intercept to the model with the final penalized parameters ignoring the penalty function and the effective degrees of freedom that result from taking the penalization

into account that are found by

$$\text{trace}(I(\boldsymbol{\beta}^P)V(\boldsymbol{\beta}^P)) \tag{2.14}$$

where $\boldsymbol{\beta}^P$ is the vector of penalized fixed effects parameters resulting from MLE, $I(\boldsymbol{\beta}^P)$ is the information matrix resulting from the model using the penalized parameters but ignoring the penalty function and $V(\boldsymbol{\beta}^P)$ is the covariance matrix for the penalized parameters of the model when taking the penalty function into account (Gray 1992; Verweij and Van Houwelingen 1994). The model that maximizes the modified AIC over the selection of possible λ values is the version of the model most likely to result in the best predictive accuracy for a new data set. This method has been shown to be asymptotically equivalent to CV approaches for selecting the penalty factor (Harrell 2014, p. 209-213). While the modified AIC approach does not explicitly account for random effects in the model, comparisons are only made between models based on the same data and with the same number of random effects, only the value of λ changes. The values of λ considered for penalization for the non-spatial CRL and the proposed GFB model are 0, 0.001, 0.01, 0.1, 0.25, 0.5, 1, 2, 5 and 10.

2.2.5 Simulation Study

Survey data was simulated using a modified version of the `SimSurvey R` package available on GitHub. `SimSurvey` was originally designed with the purpose of testing different stratified random sampling survey designs for a research vessel survey aimed at estimating abundance. `SimSurvey` is capable of generating data quite similar to those resulting from a stratified random survey design like the multi-species bottom trawl research vessel survey that DFO performs annually in the Newfoundland region and elsewhere. With simulated data it's possible to know the true age structure and

the abundance numbers for the population. Further details on the simulation study are provided in Appendix A (Regular et al. 2020).

The population simulated with `SimSurvey` is similar to the cod population living in Northwest Atlantic Fisheries Organization (NAFO) subdivision 3Ps. The fish were set to grow according to a Von Bertalanffy growth curve with an asymptotic length (L_∞) of 120 cm and K parameter of 0.5. The population is spatially distributed within grid cells grouped into strata based on depth and a stratified random survey is taken by sampling random cells. A subsample of fish from sampled locations is obtained based on length stratified sampling and considered to have been "aged". This subsample of fish is what is used to construct the four different ALKs described below.

Spatial methods have an opportunity to improve the estimate of age structure by being better able to discriminate between age classes with overlapping length distributions by taking into account the location where sampling occurred. The simulated population does not let length at age vary from location to location rather the distribution of age classes varies spatially.

The four different aging methods applied to the simulated survey data are the traditional ALK, a smooth ALK made from a CRL model involving only length as a covariate,

$$\text{logit}(\pi_a[\boldsymbol{X}_i]) = \alpha_a + \beta_a l_i \tag{2.15}$$

a GAM based model similar to the one presented in Berg and Kristensen (2012),

$$\text{logit}(\pi_a[\boldsymbol{X}_i]) = \alpha_a + \beta_a l_i + f(\boldsymbol{s}) \tag{2.16}$$

where f is a function of location $\boldsymbol{s}$ using thin-plate regression splines and finally the model proposed here as described in Equation 2.10. An automatic selection of the maximum basis dimension (k) is used for the thin-plate regression splines with AIC

smoothness selection. The automatic selection is based on the method used by the DATRAS package discussed in Berg and Kristensen (2012) where the maximum basis dimension is the number of unique observations of the covariates appearing in the smooth terms minus one (i.e the number of unique (non-zero) tow locations minus one) unless there are less than 10 unique locations for which then it falls back to a GLM. If including spatial information increases the accuracy of predicting what age class a fish belongs to, then the stratified survey estimates of abundance at age should also be closer to true abundance numbers at age. For all of the approaches age 10 was taken to be a plus group.

The data were simulated in a simplified area with a large peninsula-like landmass represented by rectangles. A plot of the landmass and an example simulation mesh can be found in Appendix A. In practice any physical boundary can be defined with the only caveat being that more detailed boundaries require more nodes in the mesh and increase the computational time required. A simplified boundary was chosen for computational convenience to reduce the required time needed to run the model hundreds of times. 500 simulated surveys were performed, and stratified survey estimates of abundance were created for each survey using the same methods outlined in S. Smith and Somerton 1981 and the same method applied to DFO 's bottom trawl multi-species survey conducted in the Newfoundland region annually (Ings et al. 2019). The simulation does not account for observation error and each simulation has a new realization of the true abundance in each run. The Root Mean Squared Error (RMSE) is calculated on the true total population at age available to the survey (adjusting for selectivity) and the stratified survey estimates resulting from the age frequencies made from each of the four methods, the GFB, GAM, non-spatial CRL and traditional ALK.

For the simulated survey 96 tows were conducted per survey in 48 strata based on depth. A mean of 2774.26 simulated fish were caught and measured in each survey

with a mean of 454.38 of those being “aged" and used for constructing the ALKs. Length stratified sampling was used in selecting the subsamples to be aged. Out of 500 simulations, 73 failed to due to too low sampling numbers. Specifically in those simulations, zero catches of older age classes occurred which prevents the models from being used as estimates of coefficients for those age classes can not be obtained. In practice this could be avoided by reducing the number of age groups below where the low sampling numbers occur.

For each simulation the total stratified abundance at age is calculated. This follows the methods outlined in S. Smith and Somerton 1981 that developed from stratified random sampling techniques described in greater detail in books like Cochran (1977) and Lohr (2009). The survey area is divided into N trawlable units and H strata, where N_h is the number of trawlable units in strata h. The true mean catch at age a (Y_{ah}) in survey in stratum h is found by

$$\bar{Y}_{ah} = \frac{\sum_{i=1}^{N_h} y_{ahi}}{N_h}, \tag{2.17}$$

where y_{ahi} is the survey catch at age in the ith unit. The total population estimate at age a is then

$$\hat{Y}_a = N \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y}_{ah} \tag{2.18}$$

(Cochran 1977; Lohr 2009; S. Smith and Somerton 1981).

2.2.6 Application

Atlantic cod and American Plaice are distributed throughout NAFO subdivision 3Ps, but during most years abundances is highest at particular locations such as the Halibut Channel (Cod) or the southeast slope of St. Pierre Bank (Plaice). The ALK methods discussed above were applied to both Cod and American Plaice data from DFO ’s multi-species bottom trawl RV survey of NAFO subdivision 3Ps (Ings et al.

2019; M. J. Morgan et al. 2020). For the model based methods, the same model formulations given in Equations 2.10, 2.15 & 2.16 were applied to both datasets along with the empirical ALK. Cod data are from the start of DFO 's inshore/offshore survey in 1996 to 2018. American Plaice data are limited until 2013 due to the lack of aged otoliths (M. J. Morgan et al. 2020). Samples for otolith collection for both species were subject to length-stratified sampling and the number of otoliths collected varies from year to year (Ings et al. 2019). Cod otolith collection uses a sampling scheme requiring otoliths to be collected from five different areas around NAFO subdivision 3Ps Ings et al. (2019). However American Plaice collection simply requires otoliths to be collected from the entire area. In 2006 the survey was unsuccessfully completed (Ings et al. 2019). The survey areas for the two species are similar but not the same since the survey area for cod does not include all of strata (Ings et al. 2019; M. J. Morgan et al. 2020). Since the study areas differ in size, two different meshes were required for each of the applications. For cod a more detailed boundary and higher density mesh for a more exact approximation was used, while the Plaice analysis was performed using a less dense mesh with a less detailed boundary. All methods were applied to each year of data independently of one another.

The mesh design can have an impact on the performance of the model. If a mesh is not dense enough, the approximation may not work well. Mesh designs can sometimes also impact the convergence of the model. Care should also be taken to ensure that all points that should be outside the boundary, are in fact. Further details on how to create a mesh for a barrier model can be seen in the referenced tutorial (Bakka no date). As with the simulation in the previous section, age frequencies were generated using the same four methods and then stratified estimates of abundance were obtained.

2.3 Results

2.3.1 Simulation Study

The two spatial models performed similarly with the GFB model having a lower RMSE (between the true abundance at age and the estimated abundance at age across the same survey) than the traditional ALK in 76.9% of simulations and the GAM model 67.8% of simulations. Overall, 40.5% of the time the GFB model had the lowest error across all models and the GAM model 36.8% of the time. The upper part of Figure 2.1 is a pairs plot of RMSE comparing each method against one another, while the diagonal elements are density plots of the RMSE of each model. The bottom of Figure 2.1 is a boxplot of the RMSE for each method. Overall the two spatial models yield tighter bounds than the non-spatial models.

The bottom row of Figure 2.2 is the true simulated spatial distribution for fish aged 4,5 & 6 for one simulation. The ages are somewhat overlapping in their spatial distribution but are centred in different areas. The spatial ALKs are better able to discriminate between ages by considering the location of the samples. This is evident in the top row of Figure 2.2 which shows the probability of being a fish being age 4, 5 or 6 given a length of 40 cm across the simulated study area as obtained by the GFB model. The probability of being in that age class increases when predicting over the main bulk of that age class. The overall trend of the GAM approach is very similar to that of the GFB model while also displaying evidence of undesired smoothing underneath landmasses. Both the GFB and GAM yield a higher probability for 5-year-olds south of the peninsula than might be expected by the true proportions. This is due to the fact that the survey caught a larger share of age 5s than for other age classes.

Each fish's length in the population is simulated when the population is generated and then distributed spatially into the simulation grid cells such that the age and

length for every fish in the simulation is known. The true proportion at age in each of the 4833 simulation grid cells was compared to the predicted proportion at age for each of the four methods. Figure 2.3a captures the difference between the true proportion of an age and the predicted proportion from the GFB model in each of the simulation grid cells for the same simulation used in Figure 2.2. If the model was perfect then the entire map would be a single solid color representing zero difference.

Figure 2.3b is the same style of plot, except the predicted proportions come from the traditional ALK. Results are very similar to those for the non-spatial CRL model. Compared to the GFB method it has larger differences in proportion for the older ages. Younger ages have more of the difference in age proportion spread out across the space than concentrated in a single region than the GFB model.

Table 2.2: The percentage of the 160 grid cells surrounding the simulation bay where the median RMSE in each cell from all simulations is lower for either the GAM or GFB models by age. For ages four and up the GFB model outperforms the GAM which is the majority of biomass available to the survey.

	1	2	3	4	5	6	7	8	9	10
GAM	100.000	100.000	92.500	38.125	28.750	28.750	31.875	10.000	1.875	41.250
GFB	0.000	0.000	7.500	61.875	71.250	71.250	68.125	90.000	98.125	58.750

To assess how well the GFB model improves performance near a landmass the RMSE for the 160 simulation grid cells surrounding the bay was calculated. Figure 2.4 shows the median RMSE across all simulations in each of those grid cells by age from each of the four methods. The median RMSE is typically much lower in the two spatial methods than the two non-spatial ones. Table 2.2 shows the percentage of the 160 cells where either the GFB model or GAM model has a lower median RMSE. For ages 3 and below the GAM model has a lower median RMSE for most cells but for ages 4 and up the GFB model outperforms it. On average ages 4 and up make up over 74% of the biomass available to the survey in those 160 grid cell suggesting the GFB model represents an improvement on the majority of fish.

Overall the simulation showed that both spatial ALKs methods are capable of

improving estimates of abundance at age from a stratified random survey. This suggests that the age frequencies created by the spatial ALKs are more accurate. For ages that make up a larger share of the abundance like ages 4 through 7 the reduction in error is very noticeable as can be seen in the example between the GFB model and traditional ALK in Figure 2.3 and around the landmass for all methods in Figure 2.4. However for other age classes like one and two the differences can be minor.

2.3.2 Application

2.3.2.1 American Plaice

American Plaice are associated with fine substrates and both juveniles and adults frequently occur in the same habitats (M. Morgan 2000; Johnson 2004). They do not conduct extensive annual migrations (Johnson 2004). The model based methods were fit from ages one to thirteen except for a handful of years where no age one Plaice otoliths sampled. In those cases the models were run on ages two through thirteen. Estimates of total abundance at age for American Plaice for the four different method are shown in Figure 2.5. With the exception of age one plaice, the four methods result in very similar estimates for the majority of the time series of total abundance at age (obtained by aggregating across space using the Stratified Mean Method (SMM)). Closer to the end of the time series there is a divergence for some age groups like 7 and 8 due to data sparsity.

Despite the similarity in aggregated abundance metrics across methods, when looking at the spatial GAM and GFB results there are differences in the predicted probabilities. For example in Figure 2.6 , the 2003 survey year the unconditional probabilities of an American Plaice being age 8 with a length of 35 cm were predicted across the space for both the GFB and GAM models. There does appear to be evidence of the probabilities being smoothed underneath the peninsula and into the

bay in the GAM version that does not occur with the GFB models due to its support for physical barriers. Both models make it clear that the probabilities are spatially varying and there is a dependence on location. When looking at other ages not shown here, the two spatial methods do not always agree, the GAM method will sometimes predict almost flat gradients across space while the GFB for the same year will vary more across space. Figure 2.6 also showcases examples of spatial ALK constructed from the GFB and GAM models at two different points. One at the tip of Placentia Bay, and one in Fortune Bay. Each curve represents the proportion at each length taken up by that age (like the columns of an ALK) while each vertical slice at a length must sum to one (like the rows of an ALK). Since age 13 was used as a plus group as fish get longer they end up more likely to land there. The two ALKs are different at each of the two points. For instance, the ALKs estimated by the GFB model shows more overlap between ages 1 and 2 in Placentia Bay than Fortune Bay while still having more overlap between the first two ages when compared against the ALKs at the same points from the GAM model.

2.3.2.2 Cod

In contrast to American Plaice, Atlantic Cod occur over a broader range of substrates with juveniles and adults overlapping in some broad areas whereas only juveniles may be more frequently sampled at some locations closer to the coast (Dalley and Anderson 1997; Fahay et al. 1999). Adult Atlantic Cod conduct extensive migrations from areas offshore to shallow coastal locations and there is evidence to suggest some alongshore movements as well (Fahay et al. 1999; Brattey et al. 2002).

For the Cod data, age eleven was taken to be the plus group for every year and the models were fit to ages one through eleven for all years. As for the American Plaice data the stratified estimates of abundance were generated using all four methods and are shown in Figure 2.7. Each of the four aging methods result in similar trends.

The two non-spatial methods both show very similar trends with almost completely overlapping lines in most years. The spatial methods, GAM and GFB are largely similar in trend but differ occasionally from the non-spatial methods in (e.g., age 10 and 11 in 2004).

ALKs were generated at two different locations within 3Ps. Figure 2.8 displays the locations used to create these ALKs using 3 cm length bins along with corresponding visual representations. ALKs constructed at each of the two areas are different from one another. The GAM model finds proportion of fish being age 10 and 11 in Fortune Bay to be essentially zero and a similar situation occurs in Placentia Bay except for the last few age groups.

The unconditional probabilities of Cod being a certain age given various lengths were examined spatially for both the GFB and GAM methods. Both of the spatial methods suggest that there is a difference spatially in age for Cod of a given length. An example of this for the probabilities of being age 5 for 50 cm Cod can be seen in the left hand side of Figure 2.8 for both the GAM and GFB models. The GFB finds a lower probability of Cod being age 5 in the tip of Fortune Bay, the GAM model however shows evidence of smoothing underneath the landmass. Using the GFB model it can be seen that Cod on the northwestern portion of the survey area also have a much lower probability of being age 5 at 50 cm than those that live in the rest of survey area.

The simulation study presented in Section 2.2.5 was designed to have an age and length structure that mimics the 3Ps cod population with a stratified survey design similar to the one used in the region. Based on the results of the simulation study it is expected that for most years the GFB model should provide more accurate estimates of the abundance at age for ages 4 and up.

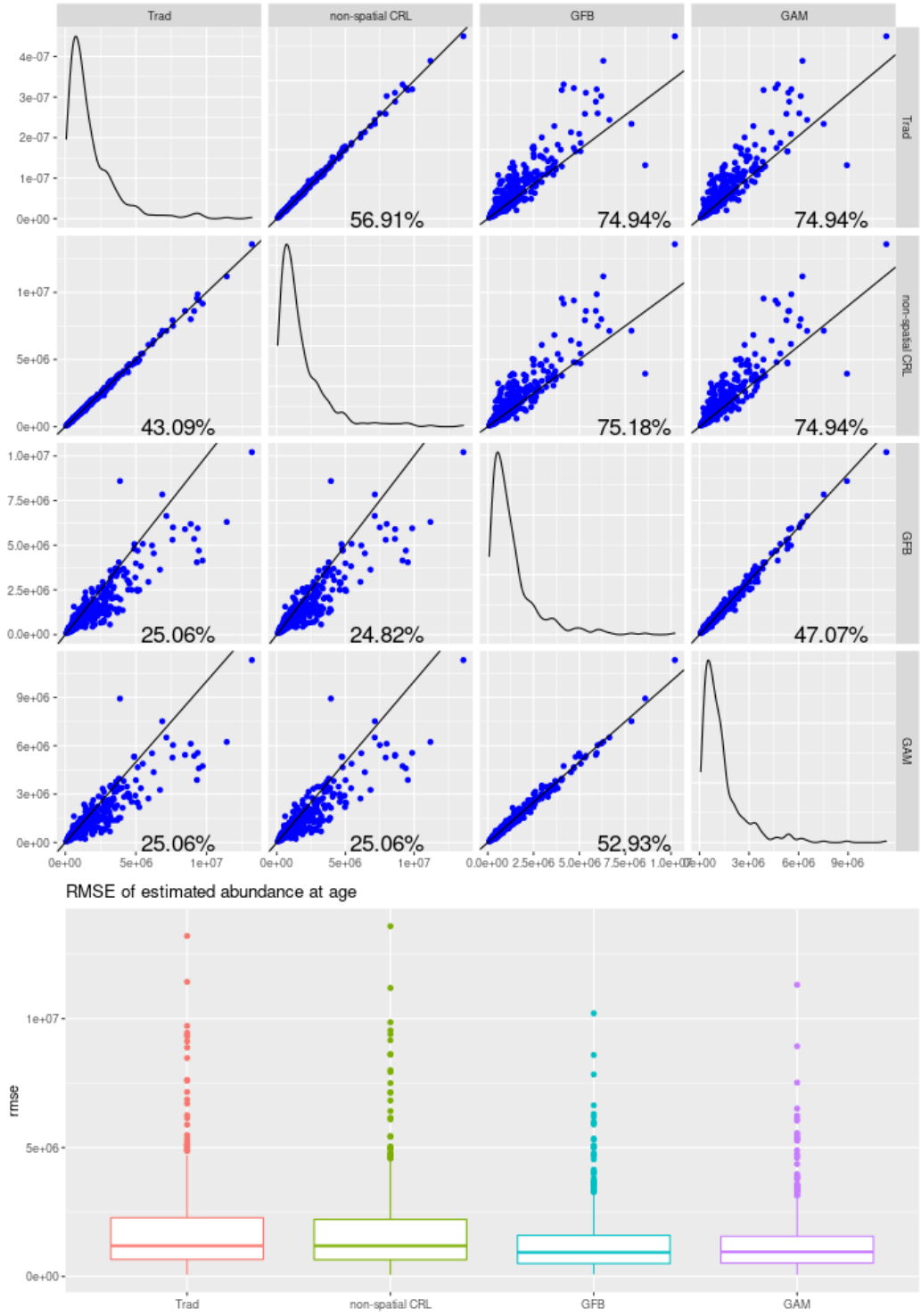


Figure 2.1: The RMSE of the stratified survey estimate of abundance at age versus the true total abundance at age in the simulation. The top figure compares the RMSE of each of the methods against one another with a diagonal line illustrates where points would lie if the two methods were identical. The percentages are the percent of simulations where the RMSE of the model on the x-axis is less than or equal to the RMSE of the model on the y-axis. Plots along the diagonal of the top figure are density plots of the RMSE for each method. The bottom part of the figure is a boxplot of the RMSE for each method.

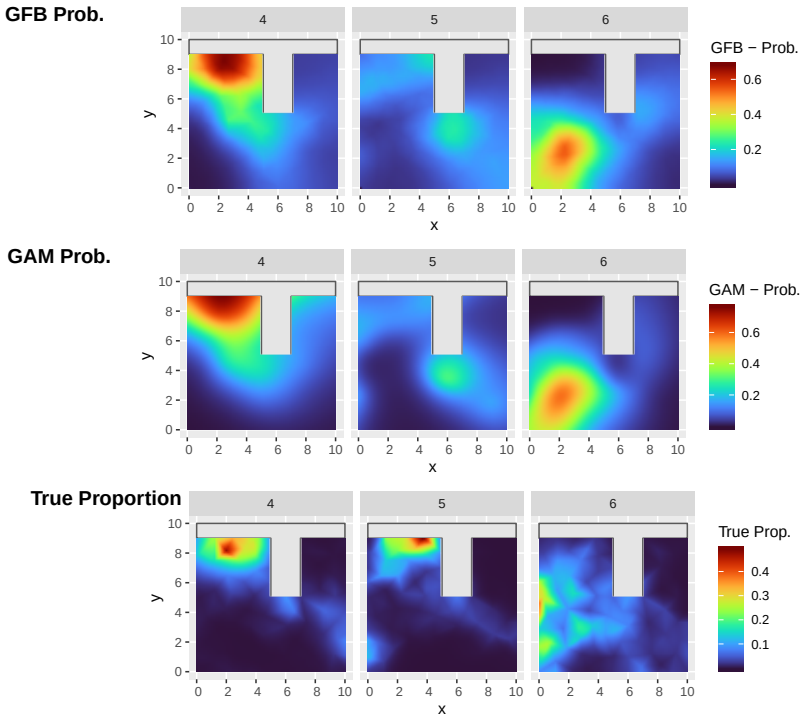
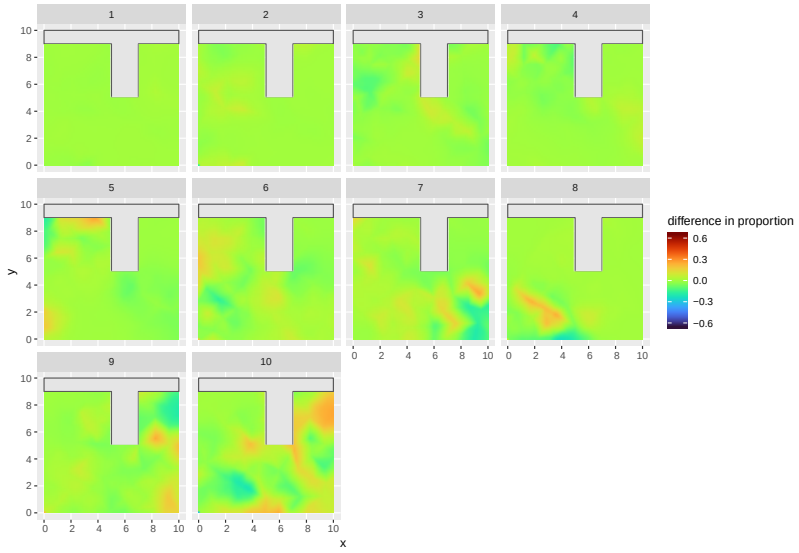
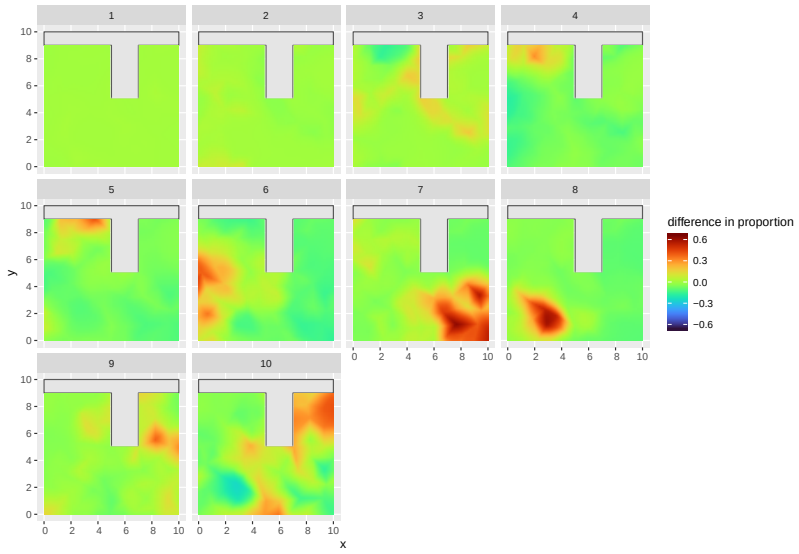


Figure 2.2: Top row is predicted probabilities in each grid cell of fish being aged 4,5, and 6 in one simulation with a length of 40 cm as predicted by the GFB model. Middle is the predicted probabilities for the same as predicted by the GAM model and the bottom row is the true simulated proportion of fish aged 4,5 and 6 in each grid cell as distributed by SimSurvey . The non-spatial CRL model found the probability of 23.1%, 10.9% and 18.6% for fish being aged 4, 5 and 6 respectively having a length of 40 cm. The traditional ALK method found the probability of 27.8%, 5.6% and 22.2% for fish being aged 4,5 and 6 respectively.

.

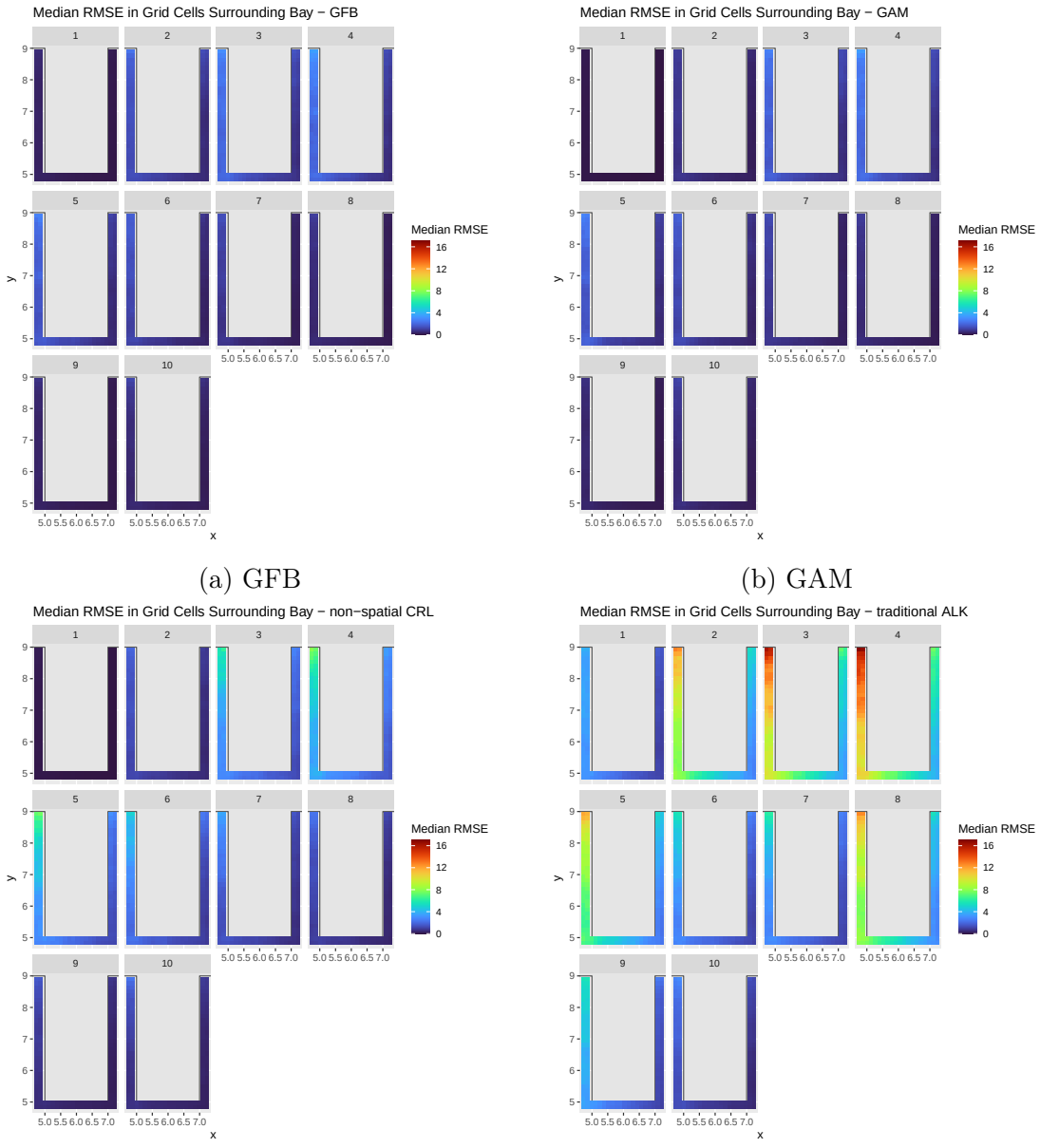


(a) Estimation error (true minus predicted) for the GFB model. A new spatial ALK is generated and applied to each simulation grid cell.



(b) Estimation error (true minus predicted) from the traditional ALK. The same global traditional ALK was applied to each simulation grid cell. The non-spatial CRL model results in a very similar plot to the traditional ALK. The traditional ALK struggles more with older ages than the GFB model.

Figure 2.3: The difference between the true proportion for each age in every grid cell and the proportion predicted by GFB model and traditional ALK for the same simulation as in Figure 2.2 for fish available to the survey. A perfect model would be a flat green with no difference in proportion between the two.



(a) GFB

(b) GAM

(c) non-spatial CRL

(d) traditional ALK

Figure 2.4: The median RMSE across all the simulations for the 160 grid cells surrounding the bay for each of the four methods by age. Both spatial models perform considerably better in each cell than the non-spatial methods. The x and y are the spatial coordinates used in the simulation and grey is the landmass.

Figure 2.5: Total abundance estimates by age for American Plaice within NAFO division 3P for the years from 1996 to 2013, not including 2006. Abundance estimates at age are largely similar across all four methods and follow the same general trends.

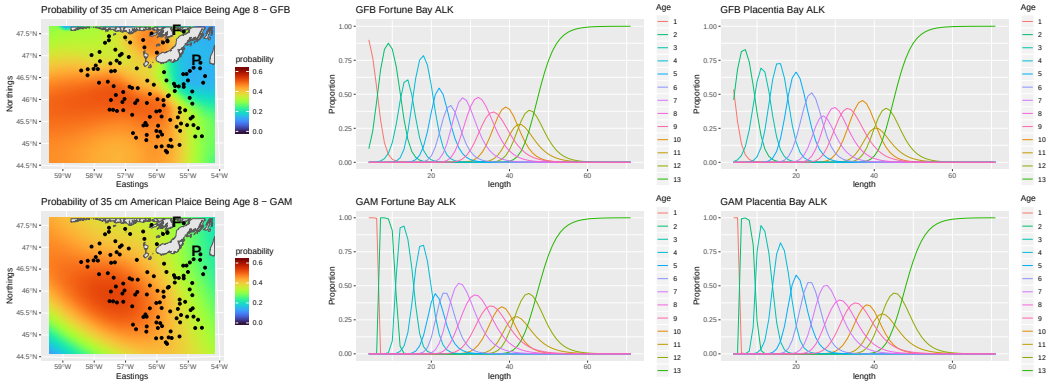


Figure 2.6: A visual representation of three ALKs generated using the GF and GAM methods at the two locations shown (P-Placentia Bay, F-Fortune Bay) for the 2003 survey year. Each vertical slice of the graphs in columns 2 and 3 must sum to 1 and is like a row in an ALK. The maps are also surfaces showing the predicted probabilities of American Plaice being 8 years old with a length of 35 cm. ALKs constructed at different locations result in different glspl:alk with either model. Points represent sampling locations from which otoliths were collected during that year.

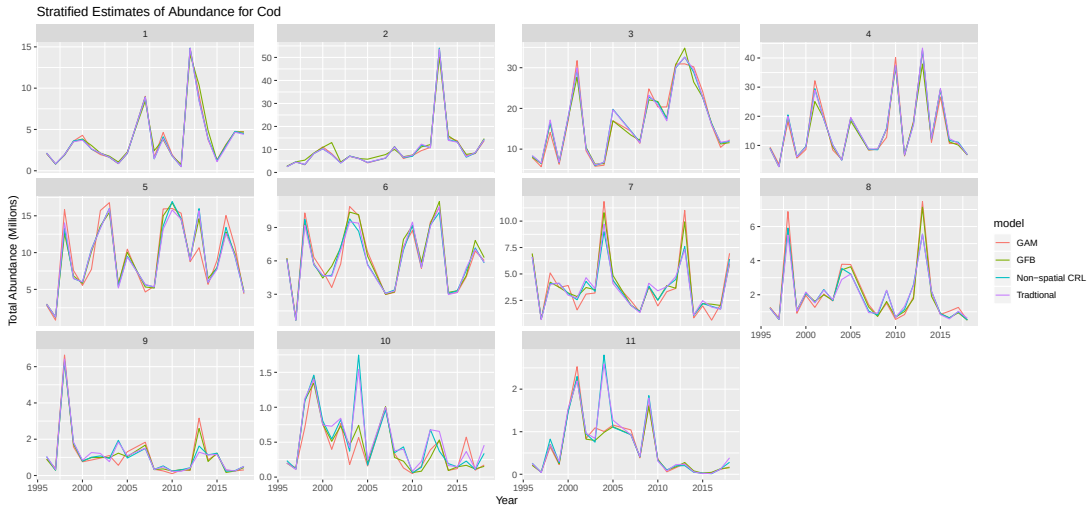


Figure 2.7: Total abundance estimates by age for the cod dataset for the years from 1996 to 2018, not including 2006. Trends across years are broadly similar between methods.

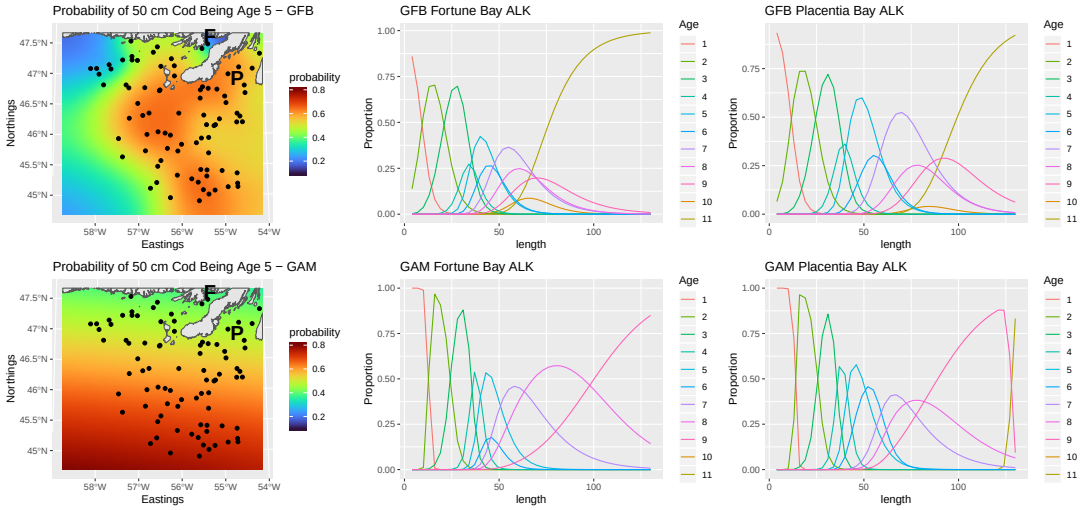


Figure 2.8: A visual representation of three ALKs generated using the GF and GAM methods at the two locations shown (P-Placentia Bay, F-Fortune Bay) for the 2011 survey year. Each vertical slice of the graphs in columns 2 and 3 must sum to 1 and is like a row in an ALK. The maps are also surfaces showing the predicted probabilities of cod being 5 years old with a length of 50 cm. ALKs constructed at different locations result in different ALKs with either model. Points represent sampling locations from which otoliths were collected during that year.

2.4 Discussion

Errors related to the process of obtaining age estimates are often ignored in age structured stock assessment models. Model based approaches offer an avenue for incorporating such errors even when a non-spatial method of aging is used. If errors in the aging process are ignored estimates obtained from the stock assessment model provide a false sense of precision and may impact derived quantities (e.g., spawning stock biomass). The simulation study here also suggests that spatial methods have the potential to further reduce errors resulting from applying an ALK. Future work includes incorporating a spatial ALK model into a stock assessment model directly to see exactly how derived quantites and associated errors may be affected by the age estimation process.

Incorporating more spatial information into the stock assessment process has the potential to increase precision. It is still not uncommon for stock assessment models to simply ignore or aggregate over space (Punt 2019), often taking an areal approach instead of a pointwise one. Our GFB model offers another choice of spatial ALK that could be integrated into stock assessment models particularly when there may be landmasses present in the study area.

In addition, both the simulation study and the application suggest that even in cases where it is not possible or desirable to fit a spatial ALK model due to data limitations or other constraints it may still be worthwhile using a model based approach to construct ALKs in order to gain the aforementioned benefits of smoothing and bridging of gaps.

Spatial ALKs can provide improved estimates at age over non-spatial methods. The simulation study showed that over three quarters of the time using a spatial method to generate the ALK had a reduced RMSE for the true abundance numbers at age as compared to traditional methods. It also suggested that spatial methods can

reduce the error across all ages in both the entire study area, and in spatial pockets for most ages. Examining the probabilities of a fish being a given length may indeed give insight into where certain age classes may be distributed during the time of the survey.

The GFB had the lowest RMSE more often than the other three models indicating it led to abudance at age estimates closer to the true values. Since the simulations distributed the fish randomly among the survey grid, in some simulations the majority of fish may not have been close to the landmass limiting the performance benefits of the GFB model to be similar to the GAM model. The GFB model also had the lowest maximum RMSE among all models suggesting it will not perform worse in most scenarios. The GFB model also provides more realistic plots of the probabilities of age given length by preventing them from smoothing beneath landmasses.

When looking at the two applications presented, neither of the spatial methods made obviously large changes to the abundance indices at age. However, when examining the predicted probabilities at age there is a clear indication that they do vary spatially. While there are similarities between the two spatial methods for the bulk of the probabilities predicted, the fact that the GFB model supports physical barriers is evident in how the bay is treated as captured in Figures 2.6 & 2.8. ALKs are demonstrated to vary with space.

This work makes evident that applying a non-spatial ALK may have ramifications for calculating indices of abundance. Combined with the fact that errors from the aging process are often ignored, there is a strong argument for integrating spatial ALK models directly into stock assessment models.

Chapter 3

Estimating Effective Population Size Using Close Kin Mark-Recapture

3.1 Introduction

Close-kin Mark-Recapture (CKMR), a method for estimating population abundance as well other demographic parameters including fecundity and survival, is based on the principle that every individual carries (i.e., tags) the genotype of each of its parents. CKMR can be applied to populations with overlapping generations. Genomic information is used to identify kin-pairs (e.g., Parent-Offspring Pairs (POPs), Half-Sibling Pairs (HSPs)) and their number is interpreted in a Mark-Recapture (MR) framework. CKMR thus replaces the physical tags or marks used in traditional MR with the close-kin relationships between individuals, offspring "tag" their parents, siblings "tag" each other, etc. In contrast to traditional MR, CKMR allows for advantages such as post-mortem sampling, lack of bias arising from misreporting or discards, no chance of tag loss and reduced harm to animals by eliminating the need for physical tags or branding (Bravington et al. 2016).

CKMR has been applied to a wide variety of organisms (mostly aquatic) including Southern Bluefin Tuna (Bravington et al. 2016), White Sharks (Hillary et al. 2018), Antarctic Blue Whales, (Bravington et al. 2016), Atlantic Salmon (Wacker et al. 2021), Brook Trout (Ruzzante, McCracken, Førland, et al. 2019), Artic Grayling (Prystupa et al. 2021), Speartooth Shark (Patterson et al. 2022), Flying Foxes (Lloyd-Jones et al. 2023) and Mosquitoes (Sharma et al. 2022). CKMR has been used to

estimate demographic parameters including adult abundance, survival and fecundity taking advantage of the role these parameters play in the probability of two individuals having a given kinship relationship (such as Parent-Offspring (PO) or Half-sibling (HS)). In a maximum likelihood or Bayes framework, we can obtain parameter estimates such that the expected number of kin pairs from the kinship probabilities best match the observed numbers of kin pairs.

HSPs are the main driver behind CKMR models' ability to estimate survival, an important parameter in population abundance models. Detecting a HSP from two different cohorts implies that the shared parent of the half-sibling individuals must have been alive to breed in both birth years. So, the probability of all potential parents surviving from the older sibling's birth year to the younger sibling's birth year must factor into the probability of a pair of individuals from two different cohorts being a HSP. As the age gap between siblings increases, the probability of the potential parent surviving decreases which is reflected in the lower number of observed HSPs. This provides insight into the adult survival rate.

HSPs from the same birth cohort (i.e., individuals born the same year) have thus far been considered more of a nuisance than a help. The number of expected sibling pairs from a within-cohort sample can result from non-independent sampling of individuals from the same cohort (e.g., sampling multiple individuals from a school of siblings travelling together) or from demographic reasons (e.g., lucky litter effect). In the demographic case, the number of expected sibling pairs from a within-cohort sample is related to the variance in number of offspring produced by parents, not just the mean, and so typically will not correspond to the expected number found using formulae derived for individuals from different cohorts (Bravington et al. 2016; Waples and Feutry 2022). Naively using formulae designed for different cohorts with comparisons for within-cohort individuals will result in incorrect or unavailable estimates of model parameters. Bravington et al. (2016) and Waples and Feutry

(2022) have suggested avoiding within-cohort comparisons entirely because of the issues caused by this variance. However, this may not be possible in many scenarios, e.g., when age is uncertain. Some CKMR models have taken the approach of using additional parameters to account for the difference in observed numbers of sibling pairs from the expected number using the formulae designed primarily for comparisons from individuals of different cohorts (Thomson et al. 2020; Patterson et al. 2022; Hillary et al. 2018).

Waples and Feutry (2022) discussed the link between within-cohort CKMR comparisons and effective population size N_e, the size of the population under an idealized Wright-Fisher model (i.e., assumes a constant population size, random mating, Poisson distributed numbers of offspring, and non-overlapping generations) were it to have the same rate of inbreeding or variance in allele frequency. N_e can serve as an indication of the extent of genetic drift or inbreeding occurring in a population. Estimates of N_e have been shown to be related to the total variance of reproductive success of individuals in the population, σ^2_{tot} (Hill 1979; Felsenstein 2005). This is the variance in the total number of offspring produced by individuals over their reproductive lifetimes (Hill 1979; Felsenstein 2005). Waples, Do, et al. (2011) developed a method for calculating N_e given a set of life table parameters. Here we take a similar tack but use CKMR to estimate the life table parameters at the same time.

In Section 3.2 we briefly discuss how σ^2_{tot} can be used to find N_e and how it can be decomposed in terms of the mean and variance of number of offspring at age in a given year. In Section 3.3, we demonstrate how to find the expected number of within-cohort comparisons and how they are linked to the mean and variance in the number of offspring born from all breeding adults, as well as survival in a given year. We show how, when combined with knowledge of survival and fecundity, we can transform the mean and variance in number of offspring at age in a given year into σ^2_{tot} and use it to estimate N_e. In Section 3.4 we further demonstrate these methods

using a more realistic age structured CKMR model implemented in Template Model Builder (TMB) using data generated with an individual based simulation where every member of the population is tracked from birth to death.

3.2 Decomposing Effective Population Size $\mathbf{N_e}$

It has been shown that N_e can be written in terms of the mean and variance in number of offspring produced during an individual's lifetime, μ_{tot} and σ^2_{tot} respectively, (Hill 1979; Crow, Kimura, et al. 1970; Waples, Do, et al. 2011) or in terms of the variance in number number of offspring and fecundity at age (Felsenstein 1971; Engen et al. 2005a). N_e can be defined in terms of the rate of inbreeding or in terms of the sampling variance of the allele frequency. Here we take the latter approach. Note that under an ideal Wright-Fisher population the variance of the allele frequency is

$$Var(p) = \frac{p(1-p)}{2N_{WF}}$$

where p is the allele frequency and N_{WF} is the number of individuals in the idealized population. The variance N_e is found by setting the true variance in allele frequency equal to $Var(p)$ and solving for the value of N_{WF} required for that to be true (Crow, Kimura, et al. 1970). The method to calculate N_e will vary from population to population as it can depend on factors like survival, fecundity, sex ratio and mating structure, requiring some understanding of the target population. In Appendix C we also provide an intuitive link between N_e and the allele frequency to μ_{tot} and σ^2_{tot}.

For now, we will focus on a relatively simple case of a two sex age-structured population with overlapping generations with equal numbers of males and females and equal numbers of progeny from each sex. Then the variance N_e in year y can be

written as

$$N_{e,y} \approx \frac{(N_{1,y-1} - 1)\mu_{tot} L}{1 + \frac{\sigma^2_{tot}}{\mu_{tot}}} \tag{3.1}$$

where L is the generation length which is defined as

$$L = \sum_{a=1}^{A} a(\prod_{b=1}^{b-1} \phi_b)\beta_a \lambda^{-a} \tag{3.2}$$

where A is oldest age class. It also represents the average age of parents at the stable age distribution. β_a is the average per capita fecundity of individuals aged a and ϕ_b is the probability of an individual surviving from age b to $b+1$. L is an adjustment for the fact that the population has overlapping generations and λ is the overall growth rate of the population (Hill 1979; Crow, Kimura, et al. 1970). The growth rate λ is the solution to the characteristic equation $1 = \sum_a (\prod_{b=1}^{a-1} \phi_b)\beta_a \lambda^{-a}$ since the population is age-structured (Caswell 2000).

In order to estimate N_e we require estimates of μ_{tot}, σ^2_{tot}, ϕ_a, β_a as well as $N_{1,y-1}$. The latter two are achievable using CKMR without within-cohort comparisons. Survival at age, ϕ_a can be estimated using CKMR when using non-lethal sampling. This is possible thanks to the survival terms present in the POP probabilities when the sampling year of the potential parent is before the birth year of the potential offspring. When combined with cases where the potential parent is sampled after the birth of the potential offspring it becomes possible to identify survival at age. In many populations, using the aggregate adult survival found when using lethal CKMR may still suffice for a suitable estimate of N_e. In the following two subsections we show how to write σ^2_{tot} and μ_{tot} in terms of the mean and variance of offspring produced by individuals that die at age a assuming that survival and reproduction are independent. We then further break it down in terms of the mean and variance in number of offspring produced by each individual at age.

3.2.1 Mean and Variance of Total Lifetime Reproductive Success

Let X be the total number of offspring produced by an individual during their lifetime, then $Var(X) = \sigma^2_{tot}$ and $E[X] = \mu_{tot}$. We can decompose $Var(X)$ in terms of the number of offspring produced by individuals given the age at which they died using the law of total variance. Let D be the random variable representing the age at which an individual died then

$$Var(X) = E_D[Var(X|D)] + Var_D(E[X|D]). \tag{3.3}$$

By noticing that the age at which an individual dies is a countable partition of X we can rewrite the first term using the law of total expectation as

$$E_D[Var(X|D)] = \sum_{a=0}^{A} Var(X|D = a)\Delta_a \tag{3.4}$$

where Δ_a is the probability of an individual dying at age a defined as

$$\Delta_a = Pr(D = a) = \phi_1\phi_2...\phi_{a-1}(1 - \phi_a). \tag{3.5}$$

The second term of Equation 3.3 can be expressed as

$$\begin{aligned} Var_D(E[X|D]) &= E_D[E[X|D]^2] - (E_D[E[X|D]])^2 && (3.6)\\ &= \sum_{a=0}^{A} E[X|D = a]^2\Delta_a - \left(\sum_{a=0}^{A} E[X|D = a]\Delta_a\right)^2 && (3.7)\\ &= \sum_{a=0}^{A} E[X|D = a]^2\Delta_a - E[X]^2 && (3.8)\\ &= \sum_{a=0}^{A} E[X|D = a]^2\Delta_a - \mu^2_{tot} && (3.9) \end{aligned}$$

again from the law of total expectation. Putting the two parts back together we get

$$\sigma^2_{tot} = \sum_{a=0}^{A} Var(X|D=a)\Delta_a + \sum_{a=0}^{A} E[X|D=a]^2\Delta_a - \mu^2_{tot}. \tag{3.10}$$

Equation 3.10 yields the variance in total reproductive success, σ^2_{tot}, in terms of the mean and variance in number of offspring produced by individuals over their lifetime given the age at which they died (noting from Equation 3.7 that μ_{tot} can be expressed that way) as well as the probability of them dying at that age.

We can write $E[X|D=a]$ as a sum of the expected number of offspring produced by individuals given age i each year up to the age that they died as

$$E[X|D=a] = \sum_{i=0}^{a} E[R|P_a = i]$$

where R is the reproductive output of an individual in a given year and P_a is the age of the adult.

If each individual's breeding events in a year are independent of their breeding events in prior years then

$$Var(X|D=a) = \sum_{i=0}^{a} Var(R|P_a = i).$$

If breeding events are not independent (e.g., because of physiological constraints leading to skip breeding) then it would be necessary to consider the covariance of reproductive output among the different age classes, see Waples and Feutry (2022) for further discussion. By combining the preceding equations we can express σ^2_{tot} and μ_{tot} in terms of the number of offspring produced at age.

As is discussed in greater detail in Section 3.3.2 the probability of two individuals sharing a parent from the same birth cohort is related to the variance in number of offspring produced per year by individuals in the population, that is $Var(R)$. In order

to estimate σ^2_{tot} and μ_{tot} we need to decompose $Var(R)$ into terms of the mean and variance in number of offspring produced by individuals in a single year at age. The age of the parent in a given year also forms a countable partition of R. This allows us to express $Var(R)$ in terms of the mean and variance in number of offspring given the age of the parent similar to how it was done above as,

$$Var(R) = \sum_{a=0}^{A} Var(R|P_a = a)Pr(P_a = a) + \sum_{a=0}^{A} E[R|P_a = a]^2 Pr(P_a = a) - E[R]^2. \tag{3.11}$$

Here $Pr(P_a = a)$ is the probability of an adult belonging to the ath age class in year y and can be found as

$$Pr(P_a = a) = \frac{N_{a,y}}{\sum_{i=0}^{A} N_{i,y}}.$$

We can also write $E[R]$ in terms of $E[R|P_a = a]$ and $Pr(P_a = a)$,

$$E[R] = \sum_{a=0}^{A} E[R|P_a = a]Pr(P_a = a). \tag{3.12}$$

3.3 Sibling Comparisons

We now explore how sibling comparisons between pairs of individuals born in different cohorts vary from within-cohort comparisons and how the latter relate to $Var(R)$. We first consider a simple scenario free from the effects of parental survival and varying average fecundity. Suppose we have two individuals i and j born in years g and h respectively where $g \leq h$ and we are interested in determining the probability that i and j share a mother based only on their birth years. Let's further suppose that all of the potential mothers of i and j survive to g, that on average they all have the same number of offspring each year and that the breeding events in years g and h are independent if $g \neq h$.

3.3.1 Comparisons between different birth years

If $R_{k,g}$ is the reproductive output of female k in year g the probability of individuals i and j sharing a mother when $g \neq h$ is

$$Pr(i \text{ and } j \text{ share a mother}|g,h) = \sum_{k}^{N_f} \left(\frac{R_{k,g}}{TRO_g} \times \frac{R_{k,h}}{TRO_h} \right) \tag{3.13}$$

where N_f is the number of potential mothers and $TRO_h = \sum_{l}^{N_f} R_l, h$ is the Total Reproductive Output (TRO) in year h. However, in practice the reproductive output of each female is unknown.

Since we assume that the mean number of offspring among the potential mothers are equal and that all mothers survive to the younger sibling's birth year, for now the probability of the two individuals sharing a mother can be written as

$$Pr(i \text{ and } j \text{ share a mother}|g,h) = \sum_{k=1}^{N_f} \frac{1}{N_f} \times \frac{1}{N_f} = \frac{1}{N_f} \tag{3.14}$$

and only depends on the number of potential mothers. This is a result of the fact that $R_{k,g}$ and $R_{k,h}$ are independent events.

In practice CKMR models operate on the number of observed kinship pairs. So if we have two independent random samples of individuals born in years g and h of sizes n_g and n_h respectively, then the expected number of pairs among the two samples sharing a mother is just the probability of two individuals sharing a mother multiplied by the number of pairwise comparisons,

$$E[\text{Number of pairs sharing a mother}|g,h] = \frac{n_g n_h}{N_f}. \tag{3.15}$$

3.3.2 Within-cohort comparisons

For within-cohort comparisons (i.e., $g = h$) Equation 3.15 does not apply due to the dependence between the births of individuals in the same cohort, since $R_{k,g} = R_{k,h}$. The probability that a pair of individuals born in the same cohort share the same mother is given by

$$Pr(i \text{ and } j \text{ share a mother} | h = g) = \sum_{k}^{N_f} \left(\frac{R_{k,h}}{TRO_h} \times \frac{R_{k,h} - 1}{TRO_h - 1} \right). \tag{3.16}$$

It can be shown with some effort (see Appendix B) that it is possible to rewrite the above probability in terms of the mean and variance in number of offspring from all mothers born in year h surviving to an age of our choosing from all mothers which we will term the reference age.

$$Pr(i \text{ and } j \text{ share a mother} | g = h) = \frac{1}{N_f} \times \left(1 + \frac{Var[R] - E[R]}{E[R]^2} \right) \tag{3.17}$$

where R is the number of offspring born in h from all mothers that survive to the chosen reference age. So long as survival is independent between individuals i and j at or above our chosen reference age, then the ages at which they were sampled do not matter (even if different). We can similarly find the expected number of sibling pairs from those born in the same year by

$$E[\text{Number of pairs sharing a mother} | g = h] = \binom{n_h}{2} \times \frac{1}{N_f} \times \left(1 + \frac{Var(R) - E[R]}{E[R]^2} \right). \tag{3.18}$$

Equation 3.18 is what can provide a CKMR model insight into $Var(R)$. Note that $E[R]$ can be constructed using Equation 3.12 using terms found with between cohort comparisons. Equation 3.18 only allows for estimation of a single parameter and so an assumption must be made about the relationship of the $Var(R|P_a = a)$ terms in

order to get an estimate of σ^2_{tot} using the decomposition discussed in Section 3.2.1. This could take the form of assuming that the variance in number of offspring at age follows a negative binomial distribution with a common overdispersion parameter θ, e.g.,

$$Var(R|P_a = a) = E[R|P_a = a] + \frac{E[R|P_a = a]^2}{\theta}$$

as is used in the age structured simulation of Section 3.4 or something like a mean-variance power relationship such as

$$Var(R|P_a = a) = cE[R|P_a = a]^\gamma$$

where c is some constant and γ is an exponent shared among all terms. What relationship is appropriate will depend on the species of interest, see Section 3.5 for further details.

3.3.3 Simple Simulation: Impact of variance on number of HSPs

To illustrate the impact of the variance of the offspring distribution on the number of sibling pairs sharing a parent we simulate parents having offspring according to the Conway-Maxwell Poisson (CMP) distribution. The CMP distribution allows for both over and under dispersion as compared to the standard Poisson distribution. The CMP distribution has Probability Mass Function (PMF)

$$f(x; \omega, \nu) = \frac{\omega^x}{(x!)^\nu} \frac{1}{W(\omega, \nu)} \tag{3.19}$$

where ν is the dispersion parameter and $W(\omega, \nu)$ is a normalizing constant to ensure the PMF sums to one. When $\nu = 1$ the CMP becomes the standard Poisson distribution with the variance equal to the mean. When $\nu < 1$ then the CMP distribution is overdispersed compared to the Poisson distribution with the variance greater than

the mean, and when $\nu > 1$ then the distribution is underdispered. We simulate offspring from 5,000 potential parents using the CMP distribution with a fixed mean of 5 and and ν values ranging from $log(-5)$ to $log(5)$ representing severe overdispersion to severe underdispersion.

Sets of within-cohort comparisons and between cohort comparisons are simulated and the number of sibling pairs found are recorded, with 125 samples for the within-cohort case in each simulation and 62 and 125 samples for the different cohort case to ensure that both scenarios have the same 7,750 pairwise comparisons. For each value of ν this is repeated 1,000 times. Figure 3.1 plots the mean number of pairs sharing a parent from the simulations at each value of $log(\nu)$ against the theoretical expectations given by Equations 3.15 and 3.18. This illustrates that the number of expected pairs sharing a parent only depends on the number of potential parents in the different cohort case (when average fecundity is equal for all individuals) but depends also on the variance (or overdispersion) as well as the number of potential parents in the within-cohort case. When the variance in number of offspring is equal to its mean, then the expected number of within-cohort sibling pairs is the same as if the between-cohort formula were applied: in this case, 1.55. When overdispersion occurs the expected number of sibling pairs sharing a mother in the within-cohort case will be greater than the expected number from the between cohort case tending towards the number of pairwise comparisons as variance in number of offspring tends towards infinity. When underdispersion occurs the expected number of pairs sharing a mother in the within-cohort case will be less than the between cohort case tending towards a lower limit when the variance in number of offspring is zero, in this case 1.24.

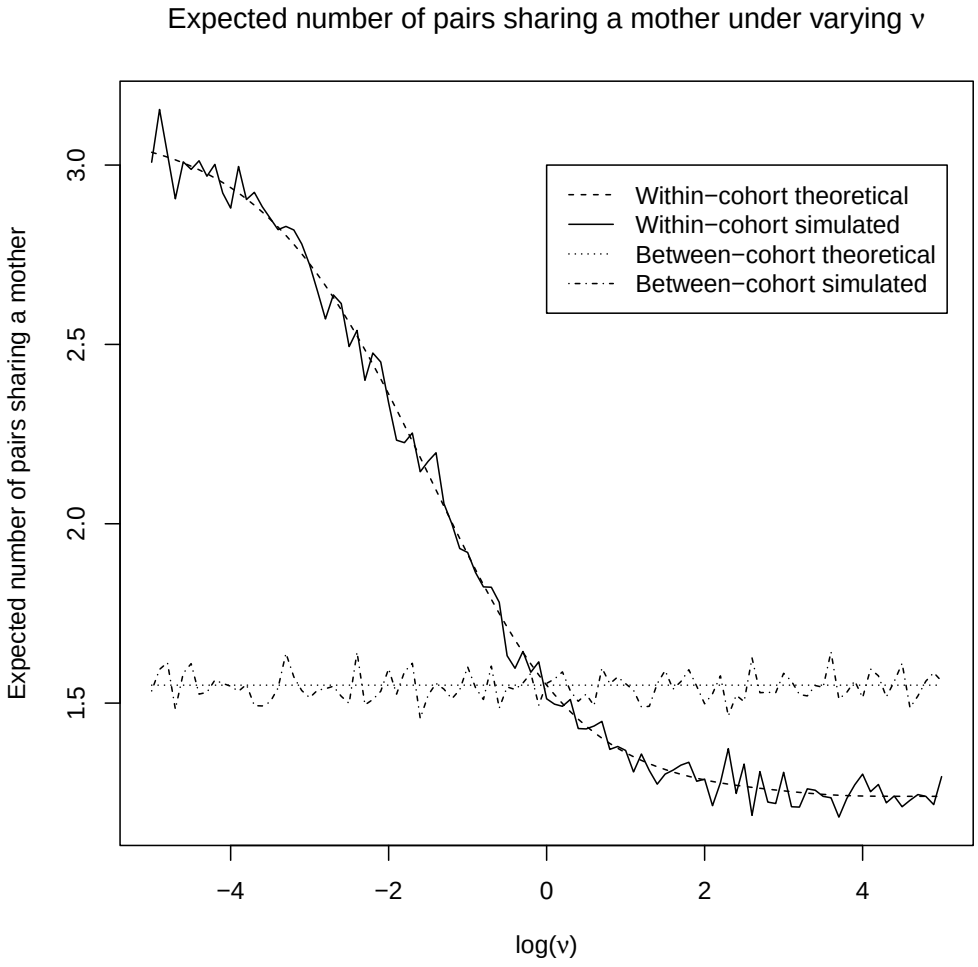


Figure 3.1: Theoretical and simulated expected number of pairs sharing a mother with a mean number of offspring of 5 and 7750 pairwise comparisons for a range of dispersion parameters.

3.4 Simulation

To test our methods on more realistic age structured populations we created an individual based simulation in `R` with similar life history characteristics to the Brook Trout populations in Ruzzante, McCracken, Førland, et al. (2019). The simulated fish are assumed to live up to three years becoming mature at age one with increasing fecundity at age. The number of offspring produced by each fish is set to follow a negative binomial distribution with a common overdispersion parameter θ for all age classes. Twenty-four populations are simulated under varying growth rates and overdispersion conditions. Ruzzante, McCracken, Parmelee, et al. (2016) estimated that the ratio of the variance of total reproductive success to the total mean number of offspring ranged from 5 to 200 in the Brook Trout populations suggesting overdispersion is a problem in the real populations used as a reference here. In our simulated populations, this ratio ranges from approximately 5 to 37. Three deterministic growth rate scenarios are targeted representing a declining, stable and growing population respectively. For each scenario we used a variety of overdispersion parameters. θ ranges between 0.1 and 5, see Table 3.1. Each population is simulated for 50 years, the first 35 years are simulated at a stable growth rate and the final 15 years using the population scenario values given in Table 3.1. The initial size for each population is set to be 10,000.

Each of the 24 populations has 1,000 independent CKMR samples (24,000 total) drawn from the final ten years of data. Sampling is simulated to be non-lethal, which allows both for model fits on the same population from different samples to be directly comparable and enables estimating survival at age. Approximately $2.5\sqrt{N_y}$ (where N_y is the total number of individuals in the population in each year) are taken for sampling to ensure adequate numbers of kin pairs. Age and kinship information are assumed to be known exactly.

We fit two models to each of the 24,000 samples. One model implements a version of the method proposed here for within-cohort comparisons and the other omits within-cohort comparisons as per Bravington et al. (2016) and serves as a baseline for comparison. The code for the simulations and model are available on github (`https://github.com/jgbabyn/varCKMR`).

We recorded the number of POPs and HSPs that were observed aggregated across each unique set of covariates that could result in a valid kin pair as well as the total number of pairwise comparisons used to find them. For POPs and HSPs from different cohorts the covariates used are the birth years and sampling years for each pair of individuals. For HSPs from the same cohort the sampling years are not used and the number of HSPs are only aggregated by birth year. Table 3.3 summarizes the covariates and data used by the model.

3.4.1 Age Structured Model

Both models discussed here are implemented identically except for their treatment of comparisons of individuals from the same cohort. They are implemented using the `R` package TMB which allows for automatic and accurate calculation of the derivatives of the objective function. They treat the sex of individuals as unknown as was the case in (Ruzzante, McCracken, Førland, et al. 2019), and set survival and fecundity parameters equal for both sexes. A table of the parameters estimated in the model is provided in Table 3.2. The model tracks fish aged one to three. The number of age one fish in year y is given by the $T\hat{R}O$ in year $y-1$

$$N_{1,y} = T\hat{R}O_{y-1} = \sum_{a=1}^{3} \beta_a N_{a,y-1} \tag{3.20}$$

where $N_{a,y}$ is the number of individuals aged a in the population in year y and β_a is the average fecundity at age a. Here the β_a are measuring absolute fecundity. For

Table 3.1: True parameters used to simulate the last 15 years of data. θ is assumed the same among all age classes. β_i is the mean fecundity at age, ϕ_i is the probability of an individual surviving from age i to $i+1$.

Growth Rate	θ	β_1	β_2	β_3	ϕ_1	ϕ_2
0.95	0.10	0.380	1.518	3.416	0.215	0.280
0.95	0.25	0.442	1.767	3.975	0.197	0.164
0.95	0.50	0.388	1.551	3.490	0.228	0.217
0.95	0.75	0.569	2.276	5.122	0.105	0.324
0.95	1.00	0.523	2.093	4.709	0.107	0.346
0.95	1.25	0.510	2.038	4.586	0.117	0.317
0.95	2.00	0.441	1.763	3.966	0.221	0.101
0.95	5.00	0.375	1.502	3.379	0.211	0.305
1.00	0.10	0.476	1.903	4.281	0.196	0.179
1.00	0.25	0.439	1.755	3.948	0.176	0.362
1.00	0.50	0.374	1.497	3.368	0.275	0.230
1.00	0.75	0.539	2.156	4.851	0.167	0.125
1.00	1.00	0.518	2.071	4.661	0.124	0.387
1.00	1.25	0.527	2.109	4.745	0.123	0.366
1.00	2.00	0.593	2.374	5.341	0.105	0.336
1.00	5.00	0.523	2.094	4.710	0.143	0.262
1.01	0.10	0.376	1.504	3.384	0.226	0.395
1.01	0.25	0.521	2.084	4.689	0.130	0.372
1.01	0.50	0.569	2.274	5.117	0.105	0.388
1.01	0.75	0.419	1.677	3.773	0.207	0.322
1.01	1.00	0.395	1.579	3.552	0.318	0.108
1.01	1.25	0.414	1.655	3.724	0.270	0.156
1.01	2.00	0.598	2.393	5.385	0.120	0.200
1.01	5.00	0.546	2.185	4.916	0.155	0.174

ages two and three it's the proportion that survive from the previous year

$$N_{a,y} = N_{a-1,y-1}\phi_{a-1} \quad a \neq 1. \tag{3.21}$$

Individuals aged 3 are assumed to die. For the initial year in the model the total number of individuals in that year $\boldsymbol{N}_{init}$ is treated as a parameter in the model and the numbers at age assumed to correspond to the fraction expected by survival at age. The ϕ_a terms in conjunction with the β_a are what describe the overall population trend.

The following kinship probabilities used in the model were tested against the age structured simulation. With the population dynamics in place the model computes the probability of a randomly selected potential parent that would be age a in potential offspring's birth year g as

$$Pr[i \text{ is } j\text{'s parent}|g,h,s] = \begin{cases} \left(\prod_{k=s-g}^{h-g}\phi_k\right)\frac{2\beta_a}{T\hat{R}O_h} & s < g \\ \frac{2\beta_a}{T\hat{R}O_h} & \text{otherwise} \end{cases} \tag{3.22}$$

where s is the sample year of the potential parent. Here, the leading product is the case when $s < g$ is accounting for the probability of the potential parent surviving to the birth year of the offspring. The remaining term is given by the average fecundity of the parent in the year the potential offspring was born over the expected number of offspring. Since the sex of the potential parents is unknown, the two is required as they could be either the mother or the father. The probabilities are then used to find the expected number of POPs by multiplying the probability with the number of pairwise comparisons performed. The expected number of POPs is compared against the observed number of POPs and they are assumed to follow a Poisson distribution when added to the likelihood because it's an accurate approximation to

the pairwise Bernoulli kinship trials due to the large number of trials and the small success probability.

For HSP kin pairs it is necessary to consider the reproductive output from all the potential parents that could have resulted in half siblings with the same age gap. Since we only consider differences between age classes here, rather than sum over all potential parents, we sum over age classes and multiply by the number of parents in that age class. It is also necessary to consider the survival of the potential parent from the birth year of the older half sibling to the younger sibling's which is given by the product of the survival of potential parent during that period. Similarly to the POP probability, since sex is unknown a two is required in the terms representing the rate of drawing an offspring from a given birth year given the age of the parent. Thus the probability that two individuals i and j where i is the older individual are a HSP from different cohorts is given by

$$Pr[i \text{ and } j \text{ share a parent}|g,h] = \sum_{a=1}^{A-(h-g)} \left(\prod_{b=a}^{a+(h-g)} \phi_b \right) \times N_{a,g} \times \frac{2\beta_a}{TRO_g} \times \frac{2\beta_{a+(h-g)}}{TRO_h}. \tag{3.23}$$

For the model that implements within-cohort comparisons the probability is similar to what is given in Equation 3.17 but with an adjustment for the fact that sex is unknown and each individual has two parents,

$$Pr(i \text{ and } j \text{ share a mother}|g=h) = \frac{1}{\hat{\mathbf{N}}_y/2} \times \left(1 + \frac{\hat{\sigma}_g^2 - \hat{\mu}_g}{\hat{\mu}_g^2}\right) \tag{3.24}$$

where $\hat{\sigma}_g^2$ is the variance in number of offspring produced by all adults relative to the first age class in the model in year g and

$$\hat{\mu}_g = \frac{\sum_a \beta^a N_{a,g}}{\sum_a N_{a,g}},$$

is the mean number of offspring produced by all adults relative to the first age class. Here, $\hat{\sigma}_g^2$ can be constructed by using Equation 3.11 and the corresponding required estimates.

As discussed in Section 3.3.2, $\hat{\sigma}_g^2$ is written in terms of the variance in number of offspring produced by individuals of each age class. $\hat{\sigma}_g^2$ and $\hat{\mu}_g$ are constructed each year in the model. The number of offspring produced by each age class is assumed to follow a negative binomial distribution with a common overdispersion parameter θ as was done in the simulation. θ is estimated by the model.

We take the same pseudo-likelihood approach used in Bravington et al. (2016) where we only consider the kinship relationship between pairs of individuals. If k_{ij} is the observed kinship between the pair of individuals i and j with the set of covariates z_i and z_j respectively then the pseudo log-likelihood of the parameter vector $\boldsymbol{\alpha}$ is

$$l_{Pse}(\boldsymbol{\alpha}) = \sum_{1 \leq i < j < j \leq n} \log Pr(K_{ij} = k_{ij} | z_i, z_j; \boldsymbol{\alpha}). \tag{3.25}$$

3.4.2 Effective Population Size

The model that performs within-cohort comparisons also estimates the variance N_e using the same form presented in Equation 3.1. This requires an estimate of σ_{tot}^2, μ_{tot}, L and $N_{1,y}$. $N_{1,y}$ follows from the abundance formulae given above and the rest are derived from other parameters estimated by the model using the methods described in Section 3.2. In the simulated populations presented here, breeding events are independent. Thus the $Var(X|D = a)$ terms in Equation 3.10 are

$$Var(X|D = a) = \sum_{i=1}^{a} 2\beta_i + \frac{(2\beta_i)^2}{\theta} \tag{3.26}$$

or just the sum of the variances for number of offspring produced by each age class. $2\beta_i$ is used since the model estimates the average per-capita fecundity and not the

average individual fecundity (which is double the average per-capita since it takes two parents to create one offspring). Similarly the ($E[X|D = a]$) terms are

$$E[X|D = a] = \sum_{i=1}^{a} 2\beta_i. \tag{3.27}$$

Finally the required Δ_a terms in Equation 3.10 are found using the estimate of survival at age, ϕ_a, which is used to find the probability of death at age a via Equation 3.5. Then all the pieces are assembled into $\hat{\sigma}^2_{tot}$ as per Equation 3.10. $\hat{\mu}_{tot}$ is similarly found by using $E[X|D = a]$ and Δ_a as seen in the latter half of Equation 3.7. The growth rate λ is obtained from the estimates of β_a and ϕ_a. Using the estimated values of the growth rate, survival and fecundity allows finding an estimate of the generation length, $\hat{L}$. We then obtain an estimate of the effective population size using

$$\hat{N}_{e,y} = \frac{(\hat{N}_{1,y-1} - 1)\hat{\mu}_{tot}\hat{L}}{1 + \frac{\hat{\sigma}^2_{tot}}{\hat{\mu}_{tot}}}. \tag{3.28}$$

Table 3.2: Parameters estimated by the models. The model that omits same cohort comparisons does not include θ.

Parameter	Description
N_{init}	The initial total number of individuals in the population
ϕ_1	Survival of age ones
ϕ_2	Survival of age twos
β_1	Average per-capita fecundity for age ones
β_2	Average per-capita fecundity for age twos
β_3	Average per-capita fecundity for age threes
θ	The overdispersion parameter in number of offspring

3.4.3 Results

Both models were fitted to all 1,000 samples from each of the 24 populations. All 24,000 model fits converged for both models.

Table 3.3: The covariates and data input into the model

Data/Covariates	Description
g	The birth year of individual i in a pair
h	The birth year of the individual j in a pair
s_i	The sampling year of individual i in a pair
s_j	The sampling year of individual j in a pair
$n_{POPs,g,h,s_{i,s_j}}$	The number of observed parent-offspring pairs at each set of covariates
$n_{HSPs,g,h,s_{i,s_j}}$	The number of observed half-sibling pairs from different cohorts at each set of
$n_{HSPsSC,g}$	The number of observed half-sibling pairs from the same cohort in each possib
n_{comp}	The number of pairwise comparisons for each set of covariates

We compared the Root Mean Squared Error (RMSE) between the simulated abundance at age and those estimated by both models as well as the model parameters and growth rate. Table 3.5 shows the proportion of the 1000 model fits for each of the 24 populations where the RMSE was lower for the model that included within-cohort comparisons. In the case of estimating the overall population abundance at age the RMSE was lower in the majority of the model fits in all but one of the populations, the RMSE for the growth rate and fecundity were lower for 22 and 19 of the populations, respectively. For survival the RMSE for the model without within-cohort comparisons was lower for 15 populations. Table 3.4 shows the 5th, 50th and 95th percentiles for the model estimates of fecundity at age, θ, survival as well as abundance at age in sampling year 5 in the population where $\theta = 1$ and the growth rate equal to 0.95 along with the true values. The percentiles are extremely similar for both models also indicating that model performance for other quantities like abundance and fecundity are not impacted when including within-cohort comparisons. Percentiles for the other remaining 23 populations are similar with the two models' performance being close. For $\theta = 5$ and $\theta = 2$ the 95th percentile estimates for θ can be extremely large, over 100,000, since the difference in lower rates of overdispersion is harder to detect from the data. While the estimate of θ may not be as accurate when θ is large, it does not have much impact on the predicted variance and the resulting estimate of σ^2_{tot}.

	WCM	WCEM	WCM	Simulated Truth	WCEM	WCM	WCEM
	5%	5%	50%		50%	95%	95%
β_1	0.454	0.455	0.497	0.523	0.500	0.542	0.550
β_2	1.809	1.807	2.176	2.093	2.188	2.591	2.612
β_3	3.518	3.554	4.721	4.709	4.749	6.156	6.207
ϕ_1	0.082	0.083	0.104	0.106	0.104	0.128	0.128
ϕ_2	0.293	0.292	0.385	0.346	0.385	0.513	0.511
$N_{1,5}$	5050.045	5038.881	5573.535	5417	5562.997	6132.643	6123.710
$N_{2,5}$	513.433	510.605	611.451	607	608.997	720.472	717.773
$N_{3,5}$	190.309	188.230	247.159	233	243.658	331.235	326.946
θ	0.806	NA	1.123	1.000	NA	1.605	NA

Table 3.4: 5th, 50th and 95th percentiles for model parameters and the abundance at age for sampling year 5 for the model with within-cohort comparisons (WCM) and the model without (WCEM) for the population when $\theta = 1$ and $\lambda = 0.95$.

However the 95th percentiles for other quantities of interest are still very close to those from the model that omits within-cohort comparisons.

We also created density plots to examine the distribution of estimates for the abundance at age, growth rate and model parameters and the variance of reproductive success across the 1,000 model fits for each population. Estimates for both models tended to be skewed for all populations. Figure 3.2 shows the density plots for the population at age in sampling year 5 with a growth rate of 0.95 and $\theta = 1$ for both models. The dashed lines indicate the true simulated abundance that year. With the combined within-cohort and between cohort comparisons the model is capable of getting reasonable estimates of population abundance. The model that includes within-cohort comparisons seems to result in narrower density plots.

Estimates of θ and σ^2_{tot} are quite good for cases when the overdispersion is more severe. Figure 3.3 shows the density plots for the population when the growth rate is 0.95 and $\theta = 1$. Again, the dashed lines are the true values. The modes of the density plots are close to the true values and show that reasonable results are achievable from a practical CKMR study of 10 years. There is however still a tendency for the estimates to be right skewed.

As mentioned above estimating the value of θ is more difficult when the amount of

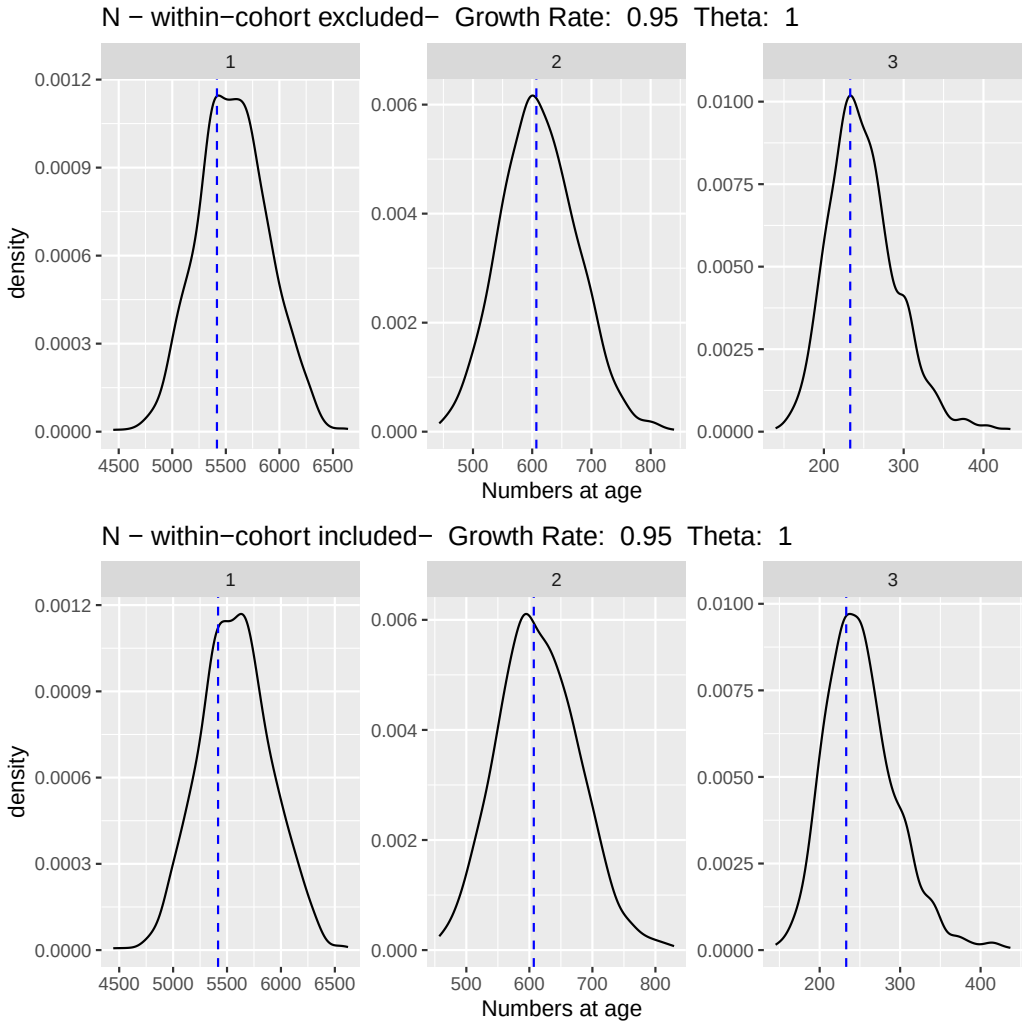


Figure 3.2: Density plots for the estimates of the numbers at age in sampling year 5 from both models for the population with a growth rate equal to 0.95 and $\theta = 1$. The dashed line indicates the true abundance.

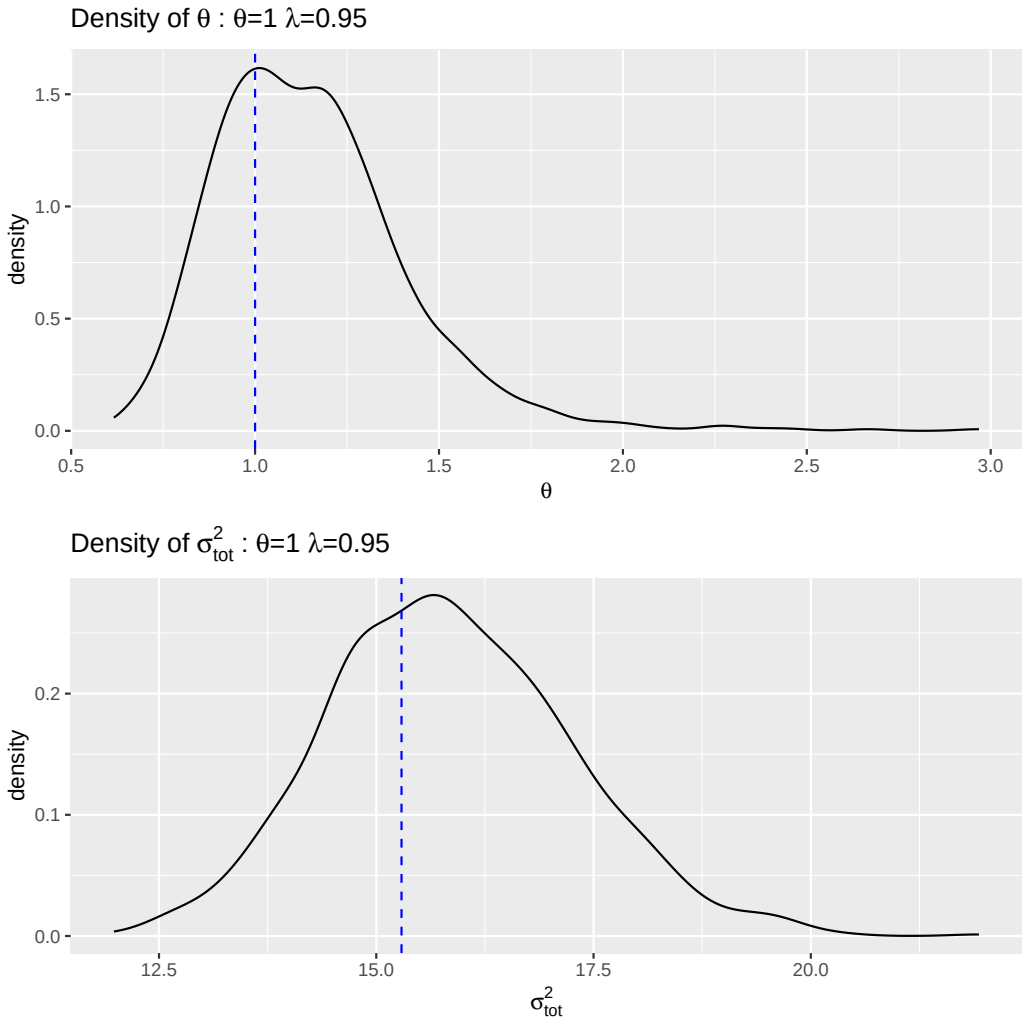


Figure 3.3: Density plots for the estimated values of θ and σ^2_{tot} for the population when the growth rate is 0.95 and $\theta = 1$.

overdispersion is low and estimates tend to be high but are not consistent. However, a bad estimate of θ does not seem to significantly affect estimates of abundance.

Overall, the addition of our method of estimating the number of siblings that share a parent in the same year does not have an adverse effect on estimating the abundance of a population using CKMR methods. Our method yields estimates of the variance of lifetime reproductive success and in scenarios where the overdispersion is high enough, these estimates are good.

Table 3.5: The proportion of the 1000 model fits for each population where the RMSE is lower for the model including within-cohort comparisons.

Population	$\boldsymbol{N}$	Growth Rate	ϕ	β
Growth Rate: 0.95 θ: 0.1	0.54	0.44	0.44	0.47
Growth Rate: 0.95 θ: 0.25	0.60	0.52	0.61	0.52
Growth Rate: 0.95 θ: 0.5	0.64	0.60	0.47	0.52
Growth Rate: 0.95 θ: 0.75	0.65	0.65	0.28	0.50
Growth Rate: 0.95 θ: 1	0.54	0.59	0.46	0.50
Growth Rate: 0.95 θ: 1.25	0.60	0.62	0.47	0.50
Growth Rate: 0.95 θ: 2	0.66	0.65	0.41	0.52
Growth Rate: 0.95 θ: 5	0.67	0.68	0.55	0.54
Growth Rate: 1 θ: 0.1	0.74	0.75	0.52	0.52
Growth Rate: 1 θ: 0.25	0.76	0.75	0.40	0.55
Growth Rate: 1 θ: 0.5	0.60	0.58	0.43	0.48
Growth Rate: 1 θ: 0.75	0.68	0.67	0.50	0.52
Growth Rate: 1 θ: 1	0.66	0.67	0.50	0.53
Growth Rate: 1 θ: 1.25	0.55	0.50	0.51	0.48
Growth Rate: 1 θ: 2	0.64	0.64	0.51	0.52
Growth Rate: 1 θ: 5	0.64	0.63	0.48	0.53
Growth Rate: 1.01 θ: 0.1	0.51	0.51	0.49	0.50
Growth Rate: 1.01 θ: 0.25	0.36	0.36	0.39	0.45
Growth Rate: 1.01 θ: 0.5	0.64	0.67	0.44	0.50
Growth Rate: 1.01 θ: 0.75	0.66	0.68	0.41	0.51
Growth Rate: 1.01 θ: 1	0.65	0.66	0.52	0.53
Growth Rate: 1.01 θ: 1.25	0.67	0.66	0.42	0.48
Growth Rate: 1.01 θ: 2	0.61	0.69	0.49	0.50
Growth Rate: 1.01 θ: 5	0.68	0.67	0.45	0.54

3.4.3.1 Effective Population Size

We also compared our estimate of effective population size against an empirical estimate of N_e using the adjusted Linkage Disequilibrium (LD) method outlined in Waples, Antao, et al. (2014) using three known life history traits. Each individual in our simulation was simulated with 100 unlinked genetic markers with two alleles each. 24,000 samples (1,000 for each of 24 populations) containing 15% of the age one individuals in every simulation year were used to estimate N_e using the adjusted LD method. A larger sample was required than what was used for the CKMR model to avoid infinite estimates of N_e. From these samples an estimate of raw number of breeders $\hat{N}_b$ was calculated with the LD method using the `R` package `strataG`. The adjusted $\hat{N}_b$ following Waples, Antao, et al. (2014) was calculated as

$$\hat{N}_{b(Adj3)} = \frac{\text{raw}\hat{N}_b}{0.991 - 0.206\log_{10}(AL) + 0.256\log_{10}(\alpha) + 0.137CV\beta} \tag{3.29}$$

which was then used to get the adjusted estimate of N_e,

$$\hat{N}_{e(Adj3)} = \frac{\hat{N}_{b(adj3)}}{0.833 + 0.637\log_{10}(AL) - 0.793\log_{10}(\alpha) - 0.423CV\beta} \tag{3.30}$$

where $\alpha = 1$ is the age of maturity, $AL = \text{max age} - \alpha + 1$ and $CV\beta$ is the coefficient of variation for the fecundity at age terms. These equations come from Waples, Antao, et al. (2014) fitting regression models of the observed N_e against the expected N_e.

The median, 5th and 95th percentiles of the $\hat{N}_{e(Adj3)}$ were computed from each of 1,000 samples for the 24 simulated populations. Figure 3.4 plots the value of N_e using the true population values as given by Equation 3.1 as well the median, 5th and 95th percentiles from the estimates using CKMR across the 1,000 CKMR samples and the median, 5th and 95th percentiles for the $\hat{N}_{e(Adj3)}$ on one of the 24 populations. As shown in Figure 3.4 N_e calculated using the true population values

agrees extremely well with the median estimates from the CKMR model, across all 24 populations. In most of the populations $\hat{N}_{e(Adj3)}$ is similar to the value given by the N_e formula in Equation 3.1 and the value estimated using the CKMR model, but in a few populations it deviates from the other two methods. The $\hat{N}_{e(Adj3)}$ method also seems to be more influenced by larger values of σ^2_{tot}

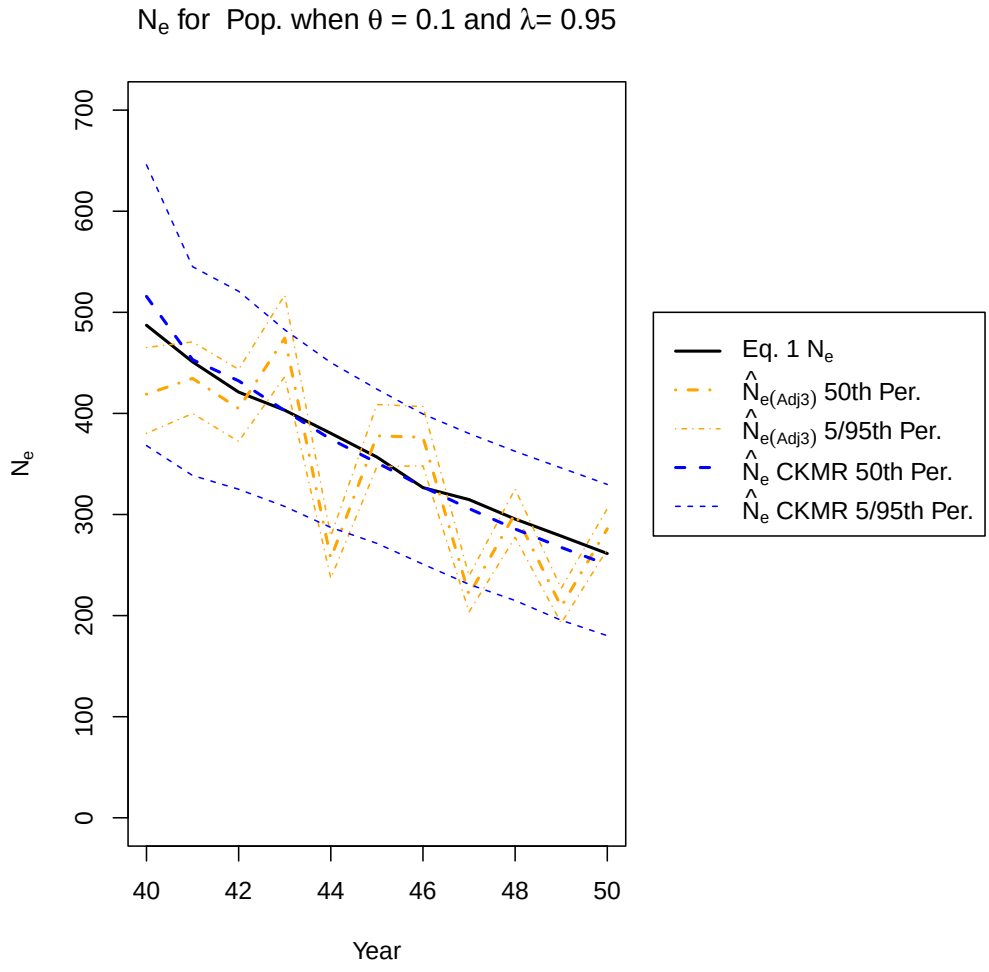


Figure 3.4: N_e computed from the true population values along with 5th, 50th and 95th percentiles for N_e estimated from the CKMR method presented here as well as for the adjusted LD method for one of the 24 populations.

3.4.4 Sensitivity Analysis of the Mean-Variance Assumption

So far, our model assumes that the mean-variance relationship of the reproductive output for an individual of age a follows a negative binomial distribution as was used in the simulation. That is

$$Var(R|P_a = a) = E[R|P_a = a] + \frac{E[R|P_a = a]^2}{\theta}. \tag{3.31}$$

In practice we do not know the true mean-variance relationship and so to test the impact of this assumption we tested a modified version of the model where Equation 3.4.4 was replaced by a power law relationship where

$$V(R|P_a = a) = cE[R|P_a = a]^{\gamma}. \tag{3.32}$$

However, there is only enough information in within cohort HSPs to estimate either the constant term c or the power term γ and not both simultaneously and so one of them must be fixed. We tried two scenarios, when $\gamma = 1$ and $\gamma = 2$. When $\gamma = 1$ then the mean-variance relationship is quasi-Poisson and the variance is proportional to the mean. This might occur if survival of offspring from a single individual in a given year are independent of one another at birth. If $\gamma = 2$ then this is akin to the case where some individuals are considerably more successful than others (i.e., the variance in number of offspring is much larger than the mean). This could occur in populations where the survival between offspring is highly correlated among siblings during early life such as in a species where an entire clutch of eggs might be eaten by a predator.

This assumption was tested on the 1,000 samples from the population where $\theta = 0.1$ and $\lambda = 0.95$. Figure 3.5 shows the 5th, 50th and 95th quantiles of N_e

estimated across the 1,000 samples as well as N_e calculated directly from the simulation parameters. The estimates from when the negative binomial is assumed and when $\gamma = 2$ are quite similar to each other as is N_e derived from the simulation parameters. In this population since θ is so small the impact of the $E[R|a]^2$ term is larger and so more akin to the power law scenario when $\gamma = 2$ than when $\gamma = 1$. While the method presented here does not seem particularly sensitive to the exact form of the mean-variance relationship, it needs to be capable of achieving similar magnitudes of the $Var(R|P_a = a)$ terms in order to get accurate estimates.

When looking at the other non-variance related terms estimated by the model the choice of the mean-variance relationship has very little impact. Table 3.6 shows the 5th, 50th and 95th percentiles for the fecundity, mortality and abundance at age estimates for sample year 5 of the simulation. Despite the gulf in N_e estimates when $\gamma = 1$ and $\gamma = 2$, non-variance related quantities result in very similar estimates suggesting that really only the variance and N_e estimates are impacted by the choice of the mean-variance relationship. The results from the model when the negative binomial variance relationship is assumed are extremely similar to when $\gamma = 2$ and so are omitted from the table.

Some care is required when determining how to structure the mean-variance relationship. Getting the form of the mean-variance relationship perfectly correct is not required, however it does need to get in the ballpark of where the true variance lies. But an incorrect choice of the mean-variance relationship is not a huge detriment to non-variance related parameters in the model.

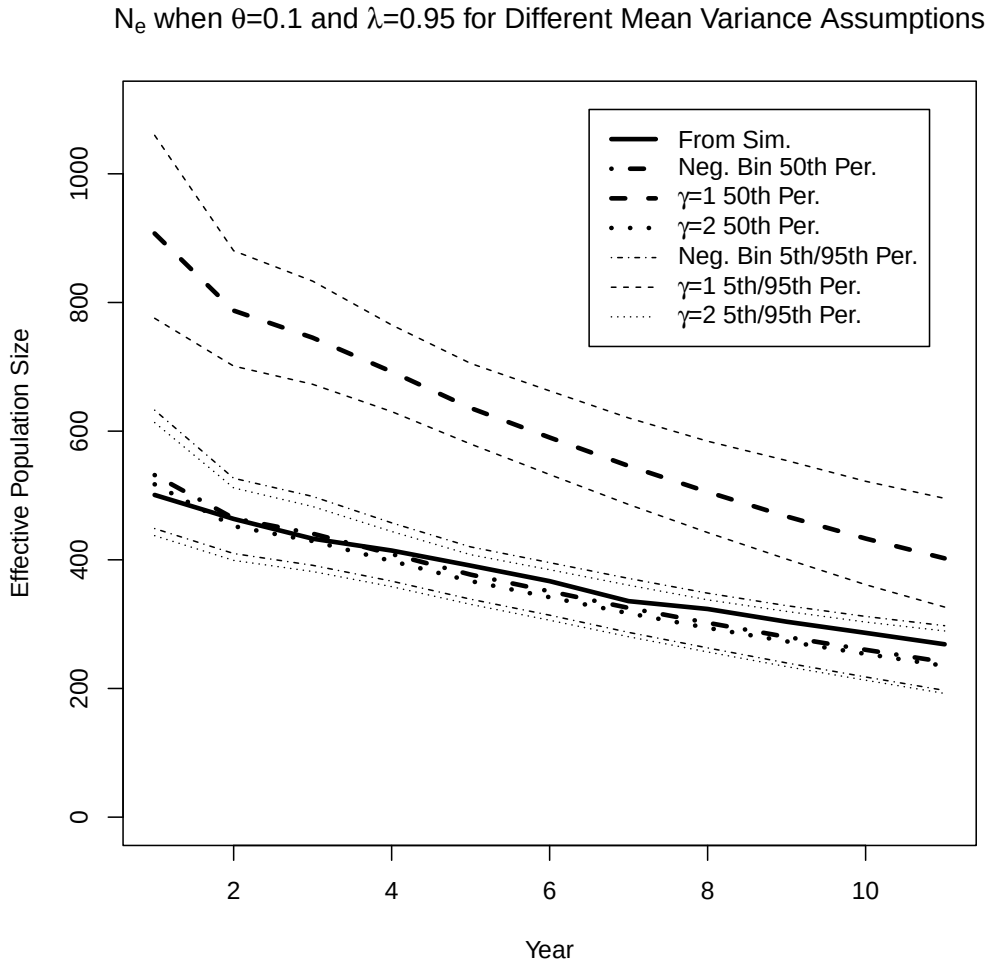


Figure 3.5: N_e for the population where $\theta = 0.1$ and $\lambda = 0.95$ estimated from the model from the three different variance relationship assumptions ($\gamma = 1, \gamma = 2$,Neg. Bin) from across 1000 samples as well as from the true population values.

	$\gamma = 1$	$\gamma = 2$	$\gamma = 1$	Simulated Truth	$\gamma = 2$	$\gamma = 1$	$\gamma = 2$
	5%	5%	50%		50%	95%	95%
β_1	0.232	0.236	0.337	0.380	0.341	0.454	0.456
β_2	1.01	1.04	1.62	1.52	1.66	2.54	2.58
β_3	1.18	1.11	3.23	3.42	3.08	6.60	6.41
ϕ_1	0.11	0.11	0.202	0.214	0.2	0.346	0.336
ϕ_2	0.741	0.807	0.31	0.281	0.326	0.15	0.153
$N_{1,5}$	3094.437	3092.366	4100.052	3951	4091.484	5615.102	5600.267
$N_{2,5}$	631.901	627.369	905.28	917	892.626	1271.023	1246.604
$N_{3,5}$	147.897	145.464	296.737	274	304.808	744.153	791.737

Table 3.6: The 5th, 50th and 95th percentiles for the model parameters (except θ) and the abundance at age for sampling year 5 for the models assuming a power law mean-variance relationship. Estimates for when the negative binomial variance relationship is assumed are very similar to when $\gamma = 2$ and are omitted.

3.5 Discussion

We have shown here how to estimate N_e using CKMR. This was done by decomposing the total reproductive variance into the mean and variance in number of offspring produced by each age class. By leveraging within-cohort HSPs we estimate the variance in number of offspring produced by all breeding adults in a given birth year which can be written in terms of the mean and variance in number of offspring from each age class. With the estimate of total reproductive variance in hand an estimate of N_e can be found. While we demonstrated this on a population where the fecundity, survival and number of individuals in each sex are equal this method could be extended to a population where this is not the case so long as sufficient information from mitochondrial DNA is available to distinguish between maternal and paternal HSPs from the same cohort.

In this paper we have demonstrated how the expected number of sibling pairs from the same cohort are related to the mean and variance in number of offspring produced by adults in a year. Previously it was suggested to exclude within-cohort CKMR comparisons to avoid issues caused by the mismatch in the observed number of sibling pairs versus the expected when not accounting for the variance in number

of offspring. However, we have demonstrated with an age structured individual based simulation that when within-cohort comparisons are specifically tailored to a CKMR model it can perform just as well as when within-cohort comparisons are omitted while also providing estimates of the variance in number of offspring suggesting there is little statistical cost to including them. Even in cases where estimates of the overdispersion parameters were quite far from the simulated value or when incorrect assumptions about the mean-variance relationship were used, estimates of the other quantities in the model were still similar to those obtained with the model with no within-cohort comparisons. Within-cohort comparisons on their own are not sufficient to estimate the variance in number of offspring in a given birth year. They must be combined with a known estimate of adult abundance or an additional data source, such as POPs, that allow for estimating adult abundance.

One of the necessary choices that must be made is the choice of the reference age. This is the age to which the fecundity and variance in number of offspring refer. Choosing a reference age where the survival between siblings is not independent will result in incorrect estimates of the variance in number of offspring. It's not too difficult to see from their definitions that while μ_{tot} and L do not depend on the choice of the reference age. As the unobserved survival of individuals below the reference age will be accounted for by the reduced number offspring of offspring that survive to the reference age. However, σ^2_{tot} can vary depending on the chosen reference age leading to different estimates of N_e. Obviously in order to fully capture the reproductive output of the population the reference age must at least include age at first maturity. Fisher (1939) recommended that the age before sexual maturity be used as the number of offspring surviving to maturity is what is important for calculating σ^2_{tot}. In order to be able to estimate σ^2_{tot} using within cohort HSP comparisons an assumption about the mean-variance relationship among the $Var(R|P_a = a)$ must be made. While it was shown in Section 3.4.4 that getting the form of the relationship exactly right

is not required, it is necessary to pick a form that can reasonably mimic the true $Var(R|P_a = a)$. In addition, using within cohort HSP comparisons on their own only allows for a single parameter related to the $Var(R|P_a = a)$ to be estimated. If this is the case, then it may be necessary to require some components to be fixed. A sensible default for many species may be

$$V(R|P_a = a) = cE[R|P_a = a]^2. \tag{3.33}$$

Which can apply to many animals where the success of one nest or litter can be all or nothing. A value of $\gamma = 2$ will also give a lower, more conservative estimate of N_e than alternatives.

Here we demonstrated how to estimate N_e when supplied with survival at age, ϕ_a, estimates of which are obtainable through non-lethal CKMR. In many populations it may be sufficient to instead use an aggregate estimate of adult survival as is achievable with lethal CKMR. If adult survival among age classes is thought to be very similar and not particularly variable then it could be an adequate replacement.

We also assumed that survival and fecundity are not varying over time. It would be possible to extend the ideas presented here to situations where this may not be the case. This would require estimating terms like survival and fecundity on a yearly basis, potentially with something like a random walk. This could also help account for populations where cohorts in some years considerably outperform cohorts in other years.

Full sibling kin pairs from the same cohort can also be used to inform the variance in number of offspring produced by males and females if random mating still applies. Even in species where parents breed with a single mate in a given year and all siblings from the same cohort are full siblings Equation 3.18 can still be used to find the variance in number of offspring among females (or males). Care must still be taken

to avoid issues where full sibling pairs are not sampled independently.

While the populations presented here were relatively simple and allowed for niceties not available in all populations (like the ability to estimate absolute fecundity), it has been made abundantly clear that it is possible to estimate all the pieces required for N_e using CKMR. Estimates of N_e that better account for fluctuating population size and time-varying vital rates may still be possible using the methods outlined in Engen et al. (2005b) combined with CKMR.

Chapter 4

Assessing the Feasibility of using CKMR on Sable Island Grey Seals

4.1 Introduction

Sable Island is a small remote crescent shaped island approximately 175 kilometres from mainland Nova Scotia. The island is home to the world's largest grey seal breeding colony (W. Bowen, McMillan, et al. 2007). The number of pups born on the island every year has increased exponentially since monitoring first began with a few hundred in the 1960's expanding into the tens of thousands during the 1990s with the last reported estimate of 2021 being over 80,000 pups and a total population estimate of over 300,000 individuals (Hammill et al. 2023).

Female grey seals are known to live to approximately 45 years of age and males to 35. Females start to have pups at 4 years or older with males reaching maturity at 7 to 9 years (den Heyer and W. Bowen 2017). Grey seals are considered capital breeders which means females spend stored reserves during lactation of pups. When pups have been weaned, females will abandon their pups and go back to sea (W. Bowen, Iverson, et al. 2006). Pups will remain on Sable Island for a few weeks after their mother leaves before heading out to sea themselves (Noren et al. 2008).

Currently the grey seal population on Sable Island is assessed using an Integrated Population Model (IPM) that combines estimates of pup production, survival, removals and reproductive rates to capture the dynamics of the population and uncertainty present (Hammill et al. 2023; Rossi et al. 2021). Prior to 1989 pup production

was estimated based on tagging, from 1989 aerial surveys have been run every few years (den Heyer, W. D. Bowen, et al. 2021). Tagging programs on Sable Island grey seals have been conducted since the 1960's and since 2002 over 6000 pups have been marked using hot iron brands. Resightings involve conducting whole island censuses searching for marked individuals multiple times in a breeding season (den Heyer and W. Bowen 2017). This Mark-Recapture (MR) effort has served as the source of survival and reproductive rates used in assessments (Hammill et al. 2023; den Heyer and W. Bowen 2017).

Conducting censuses of Sable Island for marked individuals is an expensive and time-consuming operation due to the remote nature of Sable Island and large breeding colony. Several weeks worth of supplies and fuel need to be transported to the island by helicopter and ship and a small team of 4-10 researchers are needed to scour the island (den Heyer 2023). Concerns have also been raised about the use of hot iron branding on marine mammals (Dalton 2005).

Close-kin Mark-Recapture (CKMR) could offer a potentially cheaper replacement for estimating abundance and survival that also reduces harm to grey seals. CKMR replaces the physical tags used in MR with the kinship relationships between pairs individuals detected through genetics (Bravington et al. 2016). Genetic samples can be collected from pups with less potential impact to the animal than brands. It would also be easier to collect these samples from weaned pups, after the mothers have left, to dcrease the impact on maternal investment and pup survival. A few hundred or thousand samples could be collected during this period over the course of a few days instead of the 4-7 weeks required to check the whole island for marked individuals during breeding season.

To assess how CKMR might work on the Sable Island grey seal colony we have created an individual based simulation that reflects what is known about the population demographic rates. In Section 4.2 we outline how this simulation is setup

and run and describe the sampling schemes under consideration. In Section 4.3 we present our prototype two sex model and result of fitting this model to the simulated data. Then in Section 4.4 we consider a model that only estimates female abundance. Section 4.5 contains our conclusions.

4.2 Simulation

We used a custom individual based simulation written in `R` to try and assess how well CKMR might perform for the Sable Island Grey Seal population. The simulation aims to mimic the Sable Island colony in size and structure. The maximum lifetime for a simulated seal is assumed to be 45 years old. Male adult survival is assumed to be lower than female survival as indicated by MR studies (den Heyer and W. Bowen 2017) and the IPM (Rossi et al. 2021). Females are simulated to have at most one pup in a given year as occurs in nature (W. Bowen, Iverson, et al. 2006). Males can potentially sire more than one pup in a year with multiple females. Male grey seals of an intermediate size are known to be more likely to sire pups than those of smaller and larger sizes (Lidgard et al. 2005). Males in the simulation fall into one of two size classes to account for this feature of the population. Males falling in the "intermediate" size class have an increased chance of being selected as the father of a pup. Males in the 20-24 age range also have an increased chance of siring more offspring. 50 populations are simulated for 300 years with certain parameters varying among them, discussed in Subsections 4.2.1 and 4.2.2. The first 200 years of each simulated population were used as a burn in period to allow for mixing among individuals. The burn in period keeps the population relatively low, around a few thousand individuals. After the burn in period the population is allowed to grow to be more reflective of the Sable Island grey seal colony. To simplify things removals have not been considered. Figure 4.1 shows the 5th, 50th and 95th percentiles of total abundance from the 50 simulated populations for the years following the burn

in period against estimates of total abundance given in Figure 4 of Hammill et al. (2023). The populations are a reasonable approximation to the size and dynamics of the Sable Island grey seal colony as suggested by the IPM used in the stock assessment.

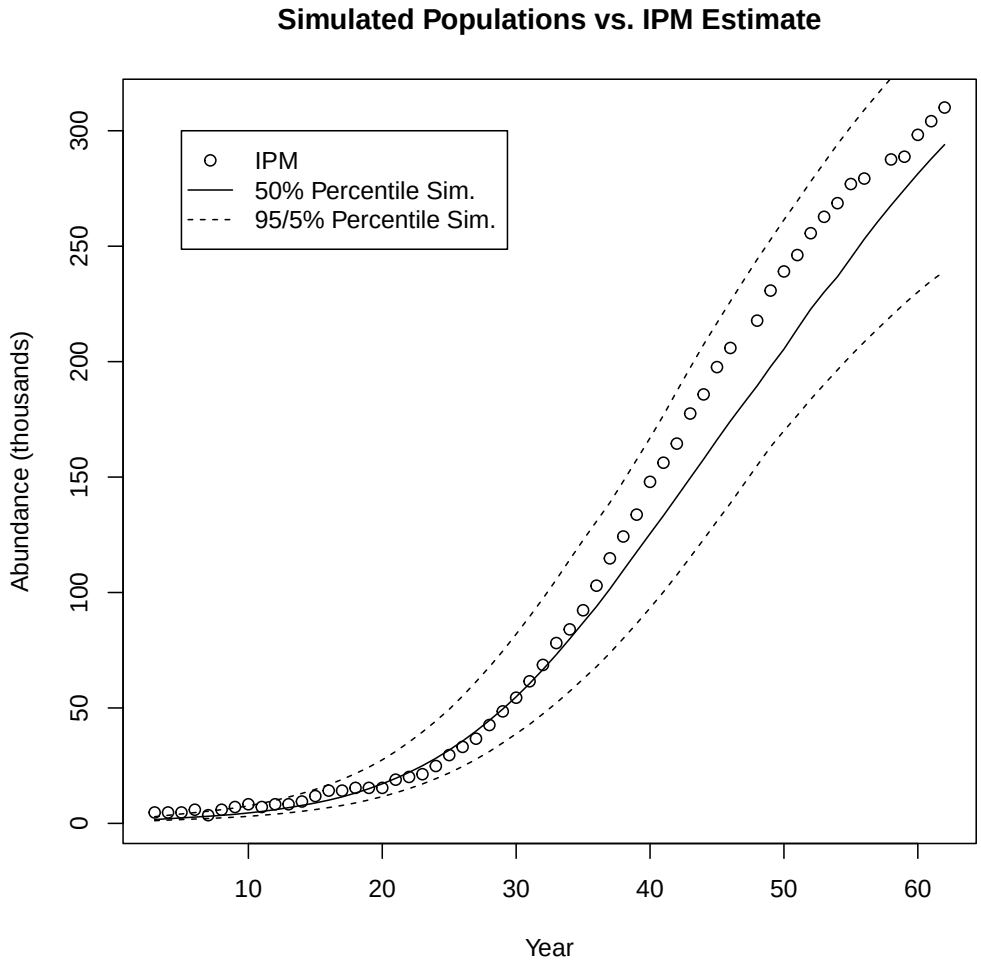


Figure 4.1: The 5th, 50th and 95th percentiles of the total abundance from the 50 simulated populations versus the Sable Island colony abundance estimate from the IPM (values taken from Figure 4 of Hammill et al. 2023).

4.2.1 Survival

Juvenile survival was split into a density dependent component (via an adjusted Ricker curve) and a density independent component following Rossi et al. (2021).

Age	0	1-9	10-14	15-19	20-24	25-29	30+
$\phi_{\text{a,m}}$	Unif(0.67,0.9)$S_t^{(D)}/2$	0.99	0.99	0.96	Unif(0.92,0.95)	Unif(0.76,0.82)	Unif(0.5,0.65)
$\phi_{\text{a,f}}$	Unif(0.67,0.9)$0.99S_t^{(D)}/2$	0.99	0.99	0.98	0.98	0.92	Unif(0.8,0.84)

Table 4.1: The survival parameters used for the simulations. $Unif(a, b)$ indicates a randomly selected value from the uniform distribution between a and b.

The density dependent recruit survival rate in year t is given by

$$S_t^{(D)} = \frac{D^\theta}{D^\theta + R_t^\theta} \tag{4.1}$$

where D is the density dependence half-saturation parameter, θ is the density dependence shape and R_t is the number of of recruits in year t. The density dependence shape parameter was randomly selected to lie between 0.90 and 1.05 and the half-saturation parameter was randomly selected from values between 10,000 and 15,000. The density independent component is a randomly selected value between 0.67 and 0.9. These values were chosen based on the parameters used in Rossi et al. (2021).

Survival for all other ages is set to follow a Bernoulli distribution with the probability of survival from one year to the next differing between males and females as well as between age classes ($\phi_{a,m}$ and $\phi_{a,f}$ respectively). The values chosen for the simulation were derived from Figure 7 of Rossi et al. (2021). When uncertainty intervals were visibly present a random value between the upper and lower bounds was chosen. Table 4.1 shows the breakdown of the probabilities among the sexes and age classes.

4.2.2 Maturity and Reproduction

The probabilities of males and females being mature are configured to follow logistic curves although with separate sets of parameters for each sex. The probability of an individual that is aged a being mature is given by

$$Pr[\text{mature}|a] = \frac{1}{1 + e^{-k_s(a-a_{50,s})}} \tag{4.2}$$

where k_s affects the steepness of the curve and $a_{50,s}$ is the age at which 50% of individuals are mature for sex s. For both males and females k_m and k_f are set to one and $a_{50,f} = 5$ and $a_{50,m} = 8$. Individuals under the age of four for females and seven for males are simulated to be immature.

To allow for skip breeding individuals in the population have a state marking them as a breeder or non-breeder and have an associated probability of transitioning from one state to another or staying in the same state. The probability of a female staying a breeder given they were a breeder last year is set to 0.85, the probability of them staying a non-breeder given they were a non-breeder is 0.30. For males these are set to 0.95 and 0.20 respectively. These values were based on those given in Rossi et al. (2021). Individuals have to be in a breeder state and mature to breed.

Breeding occurs once per year. The number of offspring had by females that are in a breeder state and mature is Bernoulli distributed with the probability controlled by the pregnancy rate which is randomly selected to be between 0.85 and 0.99. Males are allowed to breed with multiple females in a year and are divided into two classes, one representing the intermediate size class mentioned in Lidgard et al. 2005 that are more likely to sire offspring, and a non-intermediate size class. Males are assigned to the intermediate group with a thirty percent chance at birth. Ages 20 through 25 are considered ages of peak performance and greater weight is given to males within that age range. For males at peak performance in the intermediate size class a weight between 5 and 10 is randomly selected for the entire population, for those outside peak performance a value of 5 is used for the entire population. For males not in the intermediate size class in peak performance a weight randomly selected between 1 and 4 is selected for the entire population, and a value of 1 for those not at peak

performance. The weights control how likely it is for a mature male in breeder status to be selected as the father of the pup. Higher weights increase the chance of a male being selected as the father of a pup and thus more likely to be the father.

4.2.3 Sampling

A core part of this feasibility study is examining the practicality of only sampling juveniles and so all sampling presented here is juvenile only and non-lethal. Fisheries and Oceans Canada (DFO) collected genetic samples from the juveniles that were branded in the 1998-2002 and 2014-2016 cohorts (den Heyer 2023) which potentially could be used to kick-start a CKMR program. A similar setup is reflected in the sampling scheme presented below to see how the model might behave when confronted with these gaps in sample coverage. We also test three different sampling schemes, short, medium and long term. Short term examines what we might expect if we only used the samples currently available (e.g., 1998-2002 and 2014-2016). Medium term is the scenario when a few extra years are added to the samples used in the short term. Finally, long term extends the sampling so that roughly one generation of time has passed since the start of sampling. Table 4.2 shows the breakdown of which years of the simulation were sampled and how many juveniles were sampled in a given year. The observations in the long term extend those of the medium term and those in the medium term extend the short term. Year 38 of the post burn-in period was chosen as the start of sampling from the simulation as it roughly reflects the state of the Sable Island grey seal colony in 1998 as suggested by the 2022 stock assessment. For each of the 50 simulated populations, the sampling schemes were repeated 20 times giving a total of 1000 samples. It was assumed that it is possible to distinguish between relationships that are maternal and non-maternal. Size class was considered to be unknown during sampling and sex was assumed to be known.

The main kin pairs of interest and the ones used in the model are POPs, HSPs

Table 4.2: Years sampled and number of juveniles sampled for each of the three sampling schemes. Years are relative to the post-burn in period (i.e., Year 1 is the first year after the burn in period). Long and medium term extend the previous terms.

	Short	Medium	Long
Years	38-42,54-56	62-68	69-84
Annual Juveniles Sampled	500	700	700

and GPGCPs. Kinship between individuals was assumed to be known perfectly and determined directly from the simulated pedigree graph. For simplicity kin pairs from the same cohort were not used. Kinship detection was done by performing a breadth first traversal on the node in the pedigree graph corresponding to each sampled individual to a maximum of 3 relationships (edges) out. When other sampled individuals were encountered during traversal the relationship (path) between them was recorded. This allows for encountering kin pairs with the same degree of relatedness as POP, HSP or GPGCP such as a full uncle/nephew relationship which could cause a problem if they occur with too high frequency and are unaccounted for in the model. Kin pairs outside of POP, HSP and GPGCP are shown to be relatively rare. Across the 1,000 samples taken over the 50 populations two thirds of them did not find a kin pair outside of POP, HSP and GPGCP. Of those that did it was usually one over the entire 31 year sampling period, the highest being 4. The average number of kin pairs found in each of the three sampling schemes across the 1,000 samples are given in Table 4.3. One thing to note is that the number of kin pairs found does not scale linearly with the number of sampling years.

Table 4.3: The average number of kin pairs found in each of the three sampling schemes across the 1,000 samples taken.

Kin Pair	Short	Medium	Long
POP	19.21	62.62	180.49
HSP	334.70	837.96	2880.61
GPGC	5.08	61.07	248.27
Other	0.13	0.21	0.37

HSPs and GPGCPs can not be distinguished by genetics alone using pairwise comparisons (as discussed in Section 1.2). In some scenarios simulated here it is possible to distinguish between the two when the birth years of the two individuals are known (e.g., if they are born one year apart they must be a HSP). However, there are many scenarios where this is not the case and so the number of HSPs and GPGCPs are aggregated together when supplied to the model.

4.3 The Two Sex CKMR Model

In this section we discuss the inner workings of the two sex CKMR model that was fitted to 1,000 samples from each of 50 populations. This model is necessarily simplified from the simulation design presented in Section 4.2. For instance, early testing of the model during development suggested that having multiple survival terms for adults in the model could not be supported by the data necessitating the need for one common term for each sex. It was also found that juvenile density dependent and independent survival terms were not identifiable. The model is implemented in `RTMB`, an `R` package that allows for writing the model in `R` while still enabling accurate calculation of the first and second derivatives through AD (Kristensen 2023). Table 4.4 gives the parameters estimated by the model.

Table 4.4: The parameters estimated by the two sex model.

Parameter	Description
N_{init}	The initial number of non-juveniles
ϕ_f	Female survival of non-juveniles
ϕ_m	Male survival of non-juveniles
p_{rate}	The proportion of females that get pregnant
θ	The density dependent shape parameter
D	The density dependent half-saturation point
$w_{1,2}$	The weight parameter for peak males (ages 20-25) of non-intermediate size class
$w_{2,1}$	The weight parameter for non-peak males of intermediate size class
$w_{2,2}$	The weight parameter for peak males (ages 20-25) of intermediate size class

4.3.1 Population Dynamics

The model tracks the numbers of seals in a three-dimensional array $\boldsymbol{N}$, indexed by age, year and sex. The initial numbers at age a (excepting juveniles) for sex s are given by

$$N_{a,1,s} = N_{init}\left(\frac{\prod_{i=1}^{a}\phi_s}{\sum_{j=1}^{45}(\prod_{i=1}^{j}\phi_f + \prod_{i=1}^{j}\phi_m)}\right). \tag{4.3}$$

The initial numbers of juveniles for all years is the Total Reproductive Output (TRO) of the female seals in a given year, or the sum of all pups from all females in a given year. The average number of pups from a female of age a is

$$\beta_{a,f} = \begin{cases} p_{rate}PM_{a,f} & a \geq 4 \\ 0 & \text{otherwise} \end{cases}, \tag{4.4}$$

where $PM_{a,f}$ is the probability of a female seal being mature as described by Equation 4.2 and p_{rate} is the proportion of mature females that get pregnant in a year. The median age of maturity, a_f, and rate of maturity k_f are assumed to be known and the same values as those used in the simulation. The number of juveniles of sex s in year y is

$$N_{1,y,s} = \frac{\sum_{a=1}^{45}\beta_{a,f}N_{a,y,f}}{2} = \frac{T\hat{R}O_y}{2}. \tag{4.5}$$

The number of individuals of age $a, a \neq 1$ is

$$N_{a,y,s} = N_{a-1,y-1,s}e^{-Z_{a-1,y-1,s}} \tag{4.6}$$

where $Z_{a,y,s}$ is the mortality rate of individuals of sex s and age a in year y. For non-juvenile individuals

$$Z_{a,y,s} = -log(\phi_s), \quad a \neq 1. \tag{4.7}$$

For juvenile individuals is based on a juvenile density Ricker curve similar to the one used in the simulation,

$$Z_{1,y,s} = -log(\phi_y^{(D)}/2) \tag{4.8}$$

where

$$\phi_y^{(D)} = \frac{D^\theta}{D^\theta + T\hat{R}O_y^\theta} \tag{4.9}$$

where D is the half-saturation point and θ is the shape parameter.

The TRO for males is constrained in the model to be equal to the TRO for females. The average fecundity values for males of age a, $\beta_{a,y,m,size}$, is calculated on a yearly basis. Thirty percent of males are set to be in the intermediate size class, $p_{size,2} = 0.3$, and the rest are in the non-intermediate class, $p_{size,1} = 0.7$. Let $W_{i,a}$ be the weights control what proportion of the male TRO corresponds to the size class as well as from individuals that are of peak performance age. Here $W_{i,a} = w_{i,2}$ if $20 \geq a \leq 25$ and $\mathrm{w_{i,1}}$ otherwise. $w_{1,1}$ is fixed in the model to be equal to one, so the other weights are relative to $w_{1,1}$. The proportion of fecundity from male individuals of age a and size i in year y is

$$p_{a,y,i}^{fec} = \frac{PM_{a,m}W_{i,a}p_{size,i}N_{a,y,m}}{\sum_{a=1}^{45}(PM_{a,m}W_{1,a}p_{size,1}N_{a,y,m} + PM_{a,m}W_{2,a}p_{size,2}N_{a,y,m})} \tag{4.10}$$

where $PM_{a,m}$ is the proportion of males aged a that are mature and again follows Equation 4.2 using the same values used in the simulation. This is then used to find $\beta_{a,y,m,i}$,

$$\beta_{a,y,m,i} = \frac{p_{a,y,i}^{fec}T\hat{R}O_y}{p_{size,i}N_{a,y,m}}. \tag{4.11}$$

This also allows finding the amount of TRO in a given year that corresponds to each of the two size classes which is needed below.

4.3.2 Kinship Probabilities

We only consider POPs, HSPs and GPGCPs, these are kinships that are one quarter or more related and most likely to occur in sampling. While it is possible we may observe other kinship pairs with the same degree of relatedness such as full sibling, full aunt/niece, etc., with random mating and low numbers of sampling relative to the overall population size these are relatively rare as shown in Section 4.2.3 and so are not considered. The kinship probabilities outlined below are needed to find the expected number of kin pairs.

4.3.3 Parent-Offspring Pairs

Suppose we have a pair of individuals, i of sex_i and $size_i$ and j that were born and sampled in years y_i and y_j respectively and that $y_i < y_j$. Let a_{y_j} be the age of individual i in the birth year of individual j. Then the probability that i is the parent is

$$Pr[i \text{ is } j\text{'s parent}|y_i, y_j, sex_i, size_i] = \frac{N_{a_{y_i},y_j,sex_i}}{N_{1,y_i,sex_i}} \times \frac{\beta_{a,y,sex_i,size_i}}{T\hat{R}O^*_{y_j}}. \tag{4.12}$$

Here $T\hat{R}O^*_y$ depends on the size class of individual i. If individual i is female then size class does not matter and $T\hat{R}O^*_y$ is the overall $T\hat{R}O_y$. If i is male then $T\hat{R}O^*_y$ is the amount of $T\hat{R}O^*_y$ that can be contributed to their corresponding size class. Since only non-lethal juvenile sampling is under consideration; we must consider the probability that individual i survived from birth to the birth year of individual j. This is given by the leading fraction of the probability.

4.3.4 Half-Sibling Pairs

Suppose again we have two individuals i and j that are born in years y_i and y_j to a potential unobserved parent of sex sex_p and size $size_p$, where again $y_i < y_j$. Let d be the age difference between i and j. Then the probability that i and j share a parent

in common is

$$Pr[i \text{ and } j \text{ share a parent}|y_i, y_j, sex_p, size_p] = \sum_{a=d}^{A-d} N^*_{a,y_i,sex_p} \times \left(\frac{\beta_{a,y_i,sex_p,size_p}}{T\hat{R}O^*_{y_i}}\right)\left(\frac{\beta_{a+d,y_j,sex_p,size_p}}{T\hat{R}O^*_{y_j}}\right). \tag{4.13}$$

With HSPs we have to sum over all of the possible parents of both individuals which is why we multiply by N^*_{a,y_i,sex_p}. If the sex of the potential parent is male then we must take into account the differences caused by the two size classes, and so N^* and TRO^* correspond to the values with relation to the size class of the potential parent. Again, if potential parent is female then those are simply the overall values for all females.

4.3.5 Grandparent-Grandchild Pairs

GPGCPs are similar to HSPs in concept where there is an unobserved parent that needs to be considered. Again, assume we have a pair of individuals i and j and there is a potential parent of j and potential child of i with sex sex_p and size $size_p$. In addition, assume we also know the sex and size class of i then the probability that i and j are a GPGCP is

$$Pr[i \text{ is } j\text{'s grandparent}|y_i, y_j, sex_i, size_i, sex_p, size_p] = \sum_{k=y_i+1}^{y_j-1} N^*_{y_j-k,y_j,sex_p} \times \frac{N^*_{y_j-k,y_j,sex_p}}{N^*_{1,k,sex_p}} \times \frac{N^*_{k-y_i,k,sex_i}}{N^*_{1,y_i,sex_i}} \times \left(\frac{\beta_{k-y_i,k,sex_i,size_i}}{T\hat{R}O^*_{k}}\right)\left(\frac{\beta_{y_j-k,y_j,sex_p,size_p}}{T\hat{R}O^*_{y_j}}\right) \tag{4.14}$$

Like with HSPs it's necessary to sum over all of the potential parents of j which is reflected in the leading $N^*_{y_j-k,y_j,sex_p}$ term and survival from birth to the birth of their child needs to be considered for both the possible parent and grandparent.

4.3.6 Likelihood

Similar to Bravington et al. (2016) we take a pseudo-likelihood approach and only consider pairwise comparisons between individuals given their set of covariates, $\boldsymbol{C}$ and the set of parameters $\boldsymbol{\gamma}$. The number of kin pairs for a relationship status k for a set of covariates, $X_{\boldsymbol{C}}$, is assumed to follow a binomial distribution,

$$X_{\boldsymbol{C}} \sim \text{Binomial}(n_{\boldsymbol{C}}, Pr[K = k|\boldsymbol{C}]) \tag{4.15}$$

where $n_{\boldsymbol{C}}$ is the number of pairwise comparisons for a set of covariates and $Pr[K = k|\boldsymbol{C}]$ is the probability of relationship status k occurring as defined above. For cases where the covariates can not distinguish between a pair of individuals being a HSP or GPGCP then the two probabilities are added together to find the expected number. The model is provided with whether or not a set of comparisons was looking for a maternal or non-maternal relationship. In the latter case for GPGCPs it is then also necessary to sum over both potential sexes for the unobserved parent. In addition for the two size classes of males, size is considered unknown. It is assumed that the fraction of observed kin pairs from each size class corresponds to the fraction of the TRO from each size class. The pseudo-likelihood used is

$$l_{Pse}(\boldsymbol{\gamma}) = \sum_{1 \leq i < j < \leq n} \log Pr(X_{\boldsymbol{C}} = x|\boldsymbol{C}; \boldsymbol{\gamma}). \tag{4.16}$$

4.3.7 Results

The two sex model was fitted on each of the 1,000 long, medium and short term samples from across the 50 simulated populations. All of the two sex models fitted on the long term successfully reached model convergence, 0.6% of medium term model fits led to false convergence and 56.7% of the short term model fits ran into issues like

false convergence and in a few instances resulted in the optimizer completely failing. Even among short term models that resulted in successful convergence there were still issues using `RTMB` to estimate the standard errors of the parameters of the model with over 80% of the short term model fits running into issues on at least one parameter. This only occurred 3.6% of the time on the medium term samples and never in the long term sample fits.

Overall the two sex model tends to underestimate the true total abundance of the population across all the samples tested. On average the models fit to the long and medium term samples underestimated the true abundance by 25% and the short term models by 23%. Figure 4.2 gives a summary of the 20 model fits for the three different sampling schemes by providing the median, 5th and 95th percentiles of the model estimates from the second population simulated. One clear issue that is present in the figure and in the other populations is that the models fit to the long term samples tends to think the population has hit peak population and has already started to trend downwards earlier than the true populations.

The non-juvenile survival for males was generally estimated to be lower than that of females. The average male survival across all 1,000 long term samples were found to be 0.926 and the average female survival was found to be 0.982. The true average survival across the fifty populations ranged from 0.979 to 0.955 for males with a mean of 0.971 and for females it ranged from 0.985 to 0.965 with a mean of 0.97. Only 27.1% of the 95% confidence intervals from the long term models contained the true average female survival and 0% of them for male survival. A similar story occurs for the medium term sample fits where 32.8% of the 95% confidence intervals containing the true average female survival and 0.5% of the time for male survival.

When looking at fecundity, models fit on the long term sample tended to estimate the pregnancy rate, p_{rate} at around 0.5 with the average being 0.529. This is generally much lower than the true average pregnancy rates for the simulated populations when

you factor in non-breeding females which varied between 0.713 to 0.827. Interestingly the medium and short term model fits tended to find values close to or exactly one. For male fecundity the model fits had issues distinguishing between the peak and non-peak performance period and lowered the weight given to individuals from the peak performance ages of 20-25 resulting in a lower mean fecundity from that age class than others. For example, on the first simulated population the average number of pups from males in the 20-25 age range is 0.435 and 0.556 for the non-intermediate and intermediate size classes respectively and again 0.11 and 0.22 for mature males outside those age ranges, yet on one of the long term model fits for this population the means were 1.29 and 1.44 for peak ages for the non-intermediate and intermediate size classes and 1.82 and 1.74 for non-peak ages. This occurred most often on the medium term model fits, happening 79.5% of the time, then the long term 54% of the time. Although it was not something that happened often on the short term model fits, only happening 1.1% of the time. The model also typically estimated the weights for the peak and non-peak age classes very similarly for both size classes under all sampling schemes as seen in Table 4.5 despite in some of the populations they should be twice as large or more.

Table 4.5: The difference in peak and non-peak performance estimated male fecundity weights between size classes.

Sampling Scheme	Peak Performance	Non-peak
Short	0.0036	0.027
Medium	-0.010	0.0055
Long	-0.1614	0.074

Clearly the two sex model presented here has issues and would not make a good replacement for the current assessment used. However, the model presented here seems to be picking up some sort of signal reliably on the medium and long term samples and perhaps with further refinement could be improved.

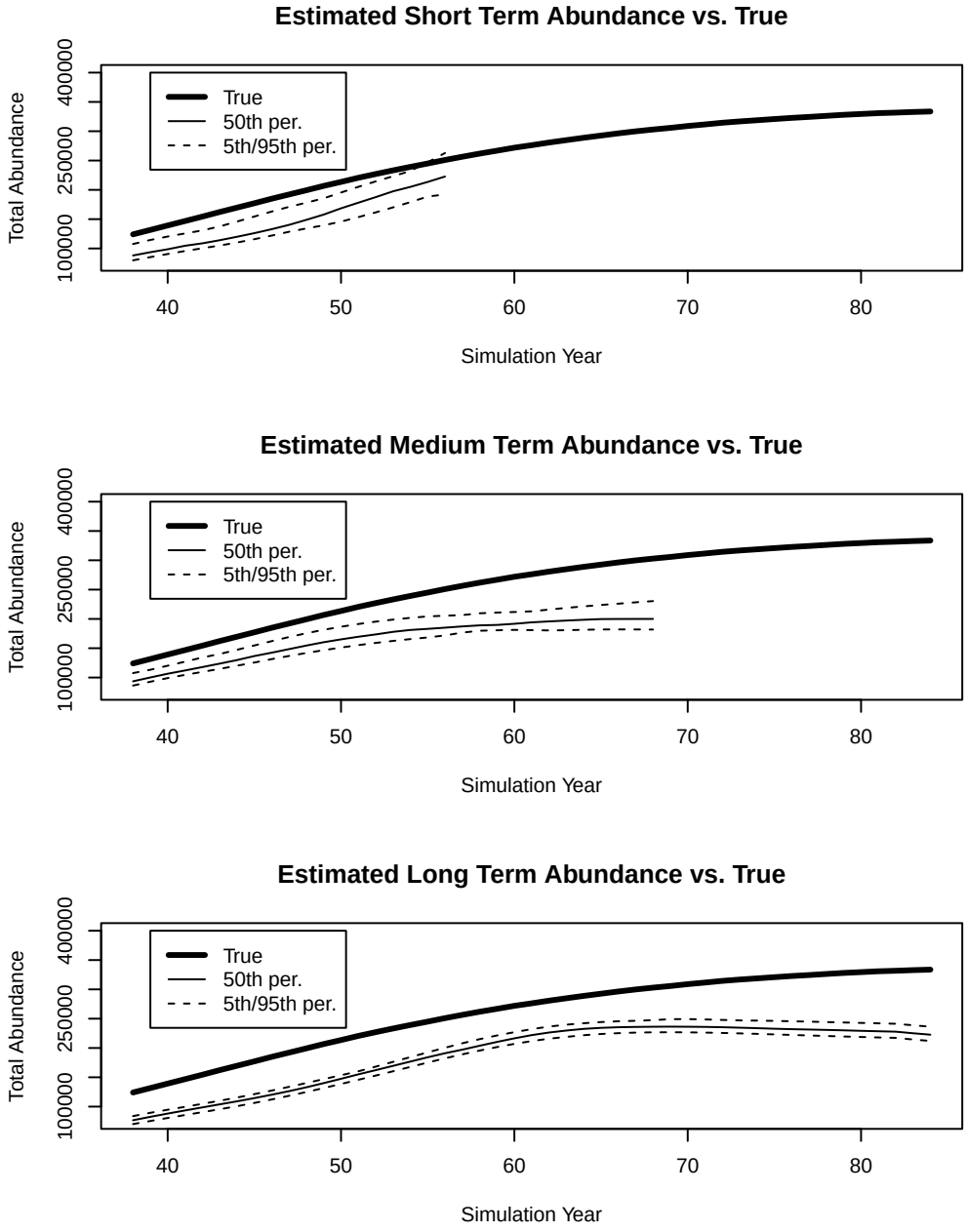


Figure 4.2: The 5th, 50th and 95th percentile abundance estimates from the 20 samples for the second population simulated.

4.4 Female Only Abundance Model

The difficulties of the two sex model in distinguishing the fecundity between males of the two different size classes likely contributed to the clear negative bias present when estimating abundance. We also considered a model which only tracks female abundance as a way to try and steer clear of the problems associated with male fecundity. The FOAM works in the same fashion as the two sex model described in Section 4.3 but with the male and size components removed and so for brevity details are omitted. It also only uses kin pair relationships that are maternal which does have the impact of reducing the number of observations available to the model. The parameters estimated by FOAM are given in Table 4.6.

Table 4.6: The parameters estimated by the female only model.

Parameter	Description
N_{init}	The initial number of non-juvenile females
ϕ_f	Female survival of non-juveniles
p_{rate}	The proportion of females that get pregnant
θ	The density dependent shape parameter
D	The density dependent half-saturation point

4.4.1 Results

The FOAM was fitted to the same long, medium and short term samples as the two sex model presented above. However, it only uses observations where the kinship relationship is known to be maternally linked (i.e, the mother was a female or the mother & grandparent are female). This results in about only 46% of the total observations being used in comparison to the two sex model. As with the two sex model the model convergence failure rates decreased as the sampling term increased, from 33.5% with the short term sample to 4.1% in the medium term to 0.3% in the long term. The FOAM also encountered similar rates of problems generating the standard errors as with the two sex model.

FOAM underestimates the true female total abundance. On average it underestimates total female abundance by 24.17%, 18.40% and 15.3% in the short, medium and long term sampling schemes respectively. In Figure 4.3 we show the 5th, 50th and 95th percentiles from the model estimates on the samples from the second population simulated which is representative of the other populations. It shows that the bias tends to be relatively consistent across the entire sampling period tending to come closer to the true values nearer to the edges. The overall shape of the trend tends to be closer to the true female abundance than the two sex model.

FOAM always estimates the pregnancy rate to be one or extremely close to one on all three of the different sampling lengths. The average of the estimates of non-juvenile survival was 0.982, 0.980 and 0.963 from the short, medium and long term samples respectively. The true mean female non-juvenile survival parameter was within the 95% confidence intervals 56.9% of the time for the long term samples, 95.6% of the time for medium term samples and 90.2% of the time for model fits where confidence intervals could be created.

We also compared the performance of the FOAM to the two sex model by comparing the female abundance estimated from the two sex model to what was estimated by FOAM. Comparing the Root Mean Squared Error (RMSE) of the total female abundance between the two models found that 96.2%, 47.0% and 79.0% of the time the RMSE was lower for the two sex model than the FOAM for the long, medium and short term sampling schemes. Figure 4.4 shows an example of the estimated female abundance as the median of the 20 samples for the second population for both models and highlights the differences in overall trend that tends to occur between the two.

If potential GPGCP pairs are removed from the data set then the underestimate of abundance appears to vanish in all fifty of the simulated populations. Figure 4.5 shows the 5th, 50th and 95th percentiles of female abundance estimates from the FOAM applied to second population simulated when potential GPGCP pairs are

removed giving a demonstration of this. When the potential GPGCP kin pairs are removed the model still has tendency to estimate the fecundity as one.

While the FOAM avoids some of the issues present in the two sex model it still has issues of its own. It may be easier to correct problems with the FOAM than the two sex model. If this is the case then one avenue for using CKMR on the Sable Island grey seals may be to adjust the population size by an external estimate of the total population sex ratio.

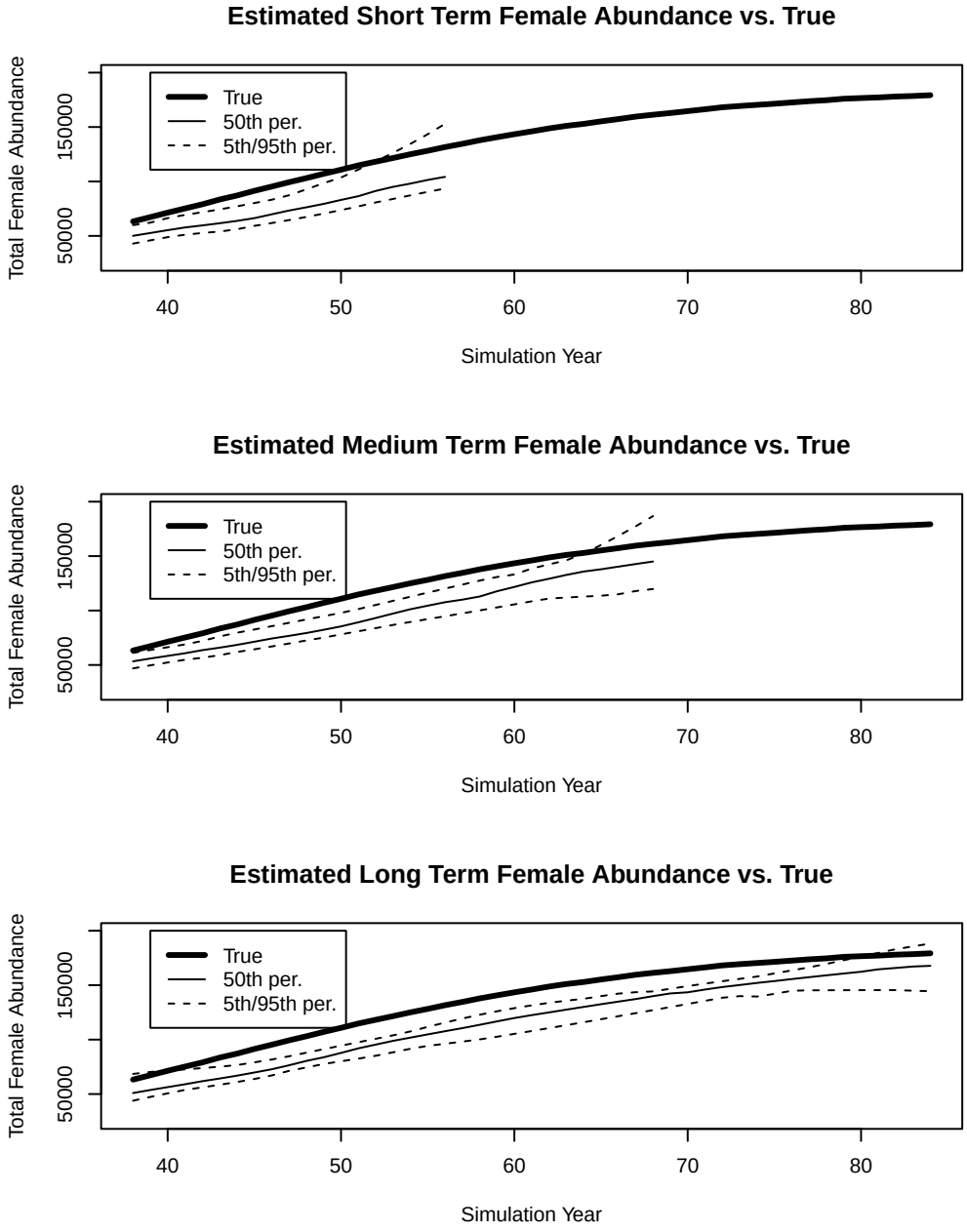


Figure 4.3: The 5th, 50th and 95th percentile female abundance estimates from the 20 samples for the second population simulated using FOAM.

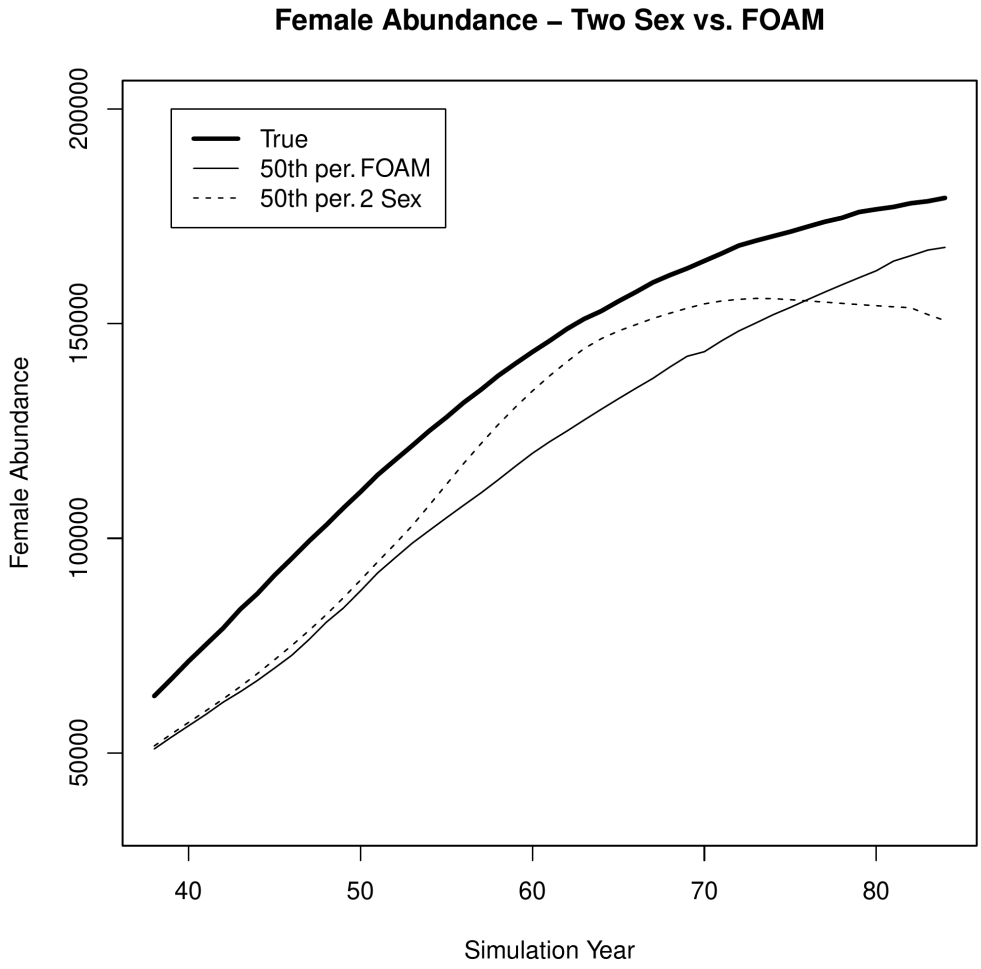


Figure 4.4: 50th percentiles of the female abundance estimates from the 20 samples for the second population simulated from both the two sex model and FOAM.

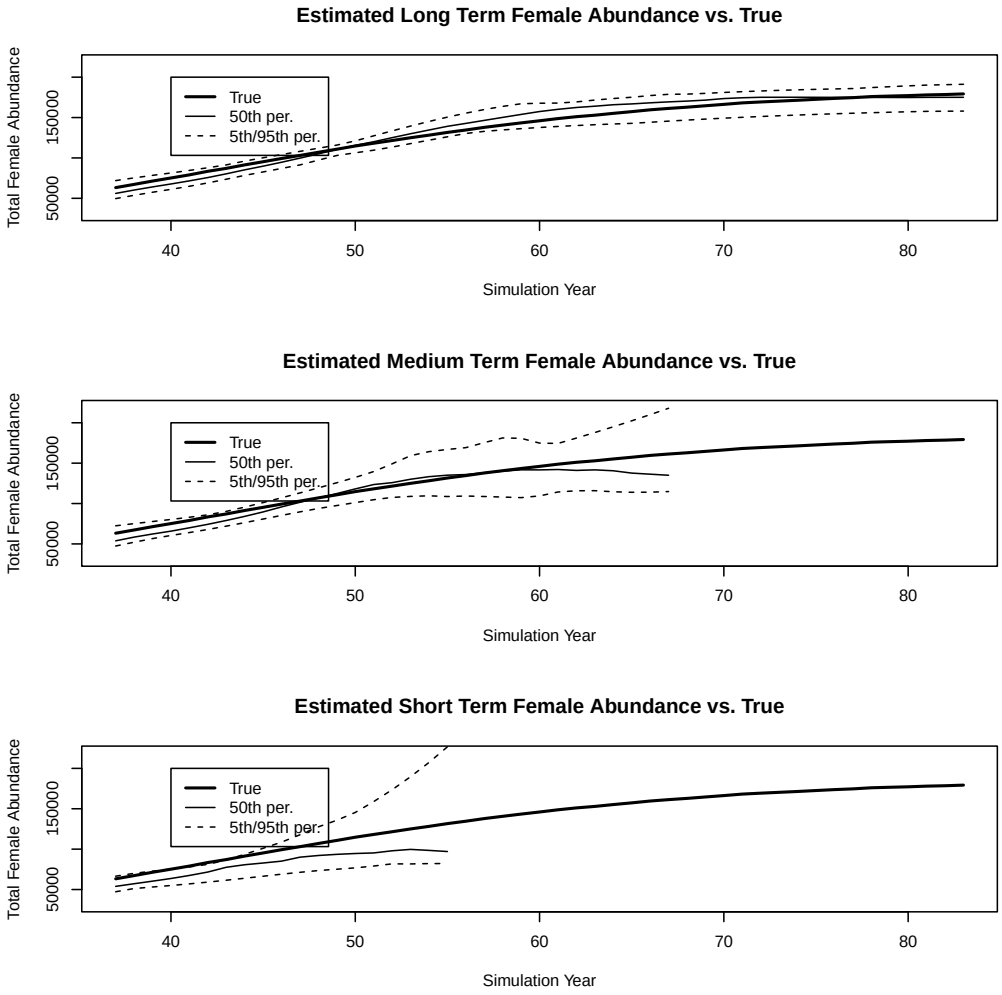


Figure 4.5: The 5th, 50th and 95th percentile female abundance estimates from the 20 samples for the second population simulated using FOAM when potential GPGCP pairs are removed.

4.5 Conclusion

In this chapter we constructed an individual based simulation that reflects our current understanding of the Sable Island grey seal colony. We then considered three different juvenile only sampling schemes varying in length (based on sampling effort already conducted by DFO that could be repurposed for CKMR). By applying the sampling schemes to the simulated populations we were able to see how many kin pairs we might expect and how common kin pairs outside of POP, HSP and GPGCP might be in the real population. We tested the sampling schemes on two models, a two sex model that tracks both males and females and accounts for differences in survival and fecundity between them, and a female only abundance model.

Both the two sex model and the FOAM result in estimates of abundance that tends to underestimate the true population values by 10 to 30 percent. For the two sex model a known source of bias is it's difficulty in handling male fecundity due to some males having increased probability of siring an offspring not being correctly reflected. Estimates of abundance for males and females end up being reduced to compensate. The model also had more trouble dealing with the short term sampling scheme which resulted in issues with convergence and obtaining standard errors for parameters. Fitting the CKMR model to data from the short term may require more compromise and simplification. Both models could reasonably well estimate the average survival values of non-juveniles. Some insight into juvenile survival could also be obtained from the model.

If potential GPGCP kin pairs are removed from the FOAM then the bias issues present in estimating abundance disappear. Since ages are known perfectly and there is an understanding of when grey seals begin to mature, it would be feasible in practice to omit any potential GPGCP kin pairs. Even if the problems the two sex model has in estimating male fecundity can not be resolved, it should still be possible to get an

estimate of the total abundance in the by scaling by the total sex ratio.

Further work could be done to try and find the source of the biases in the two sex model. Effort could also be extended to determining the optimal number of samples required for future sampling efforts. Here we only considered continuing the amount of genetic sampling that had previously been done. Another possible avenue to investigate is the addition of a small amount adult only sampling that could potentially be performed when adult samples are needed for other reasons (e.g., diet sampling). This should make it possible for the model to better distinguish between survival rates for different age classes rather than having a single non-juvenile survival term for each sex. The models presented here are also relatively simple and assume that kin finding can be done perfectly and that all samples can be used. This may not be the case in practice and it's something that may need to be further tested.

While the models presented here can not perfectly capture the simulated populations they were tested against, they at least identify the overall trend, general magnitude of the population, and get in the ballpark of average survival of males and females. With further refinement and investigation there is some hope of juvenile only CKMR being a viable replacement for the current MR and aerial surveying based methods.

Chapter 5

Conclusion

This thesis worked towards the goal of furthering methods for fisheries science by extending existing techniques and incorporating new ones. In Chapter 2, a spatial ALK model is proposed with support for physical barriers through the use of an approximation to a GF. Spatial ALKs yield better estimates of age distribution than non-spatial ALKs and the added support for physical barriers was shown to improve estimates when there is a large landmass present. In addition to the new method, an `R` package was developed to make it easy to apply and compare a number of different ALK models, including spatial and non-spatial approaches.

In Chapter 3, it was shown that CKMR can be used to estimate effective population size through the relationship between the variance in number of offspring and the number of observed sibling pairs born in the same cohort year. This enables not only estimating the effective population size but also the variance in number of offspring born each year by age class. Chapter 4 presents a feasibility study of applying CKMR to the grey seals colony of Sable Island with emphasis on a juvenile only based sampling scheme. The two sex model presented highlights the importance of properly dealing with varying fecundity among individuals and the bias it can cause. It revealed limitations in terms of what parameters could be reliably estimated. Chapter 4 also demonstrated that it may be more worthwhile to omit potential GPGCP kin pairs than to include them as doing so resulted in biased estimates of abundance.

The methods presented in this thesis relied heavily on simulation. Simulation

methods allow for knowing the true parameters and values in the population. Knowing the truth allows us to easily check model performance and compare against alternative methods. We are also able to test what happens if our assumptions happen to be wrong as was done in Chapter 3. Simulation has become a powerful and common tool in statistics, however often those presented in the literature are simulating from the model proposed. This assumes that the model is truly representative of the process under are study. If this is not the case then simulation may suggest better performance than is the case. Many of the simulations performed in this thesis instead worked from a different direction. Rather than simulating from the model being presented the simulations were based on other processes. In Chapter 2 fish were given lengths at age through a stochastic von Bertalanffy growth model instead of from the CRL ordinal regression model. In Chapters 3 and 4 individual based simulations were used rather than simulating directly from the models. This allowed for more easily reflecting reality than may have been possible with a strictly model based approach. For instance, in Chapter 4 it was possible to make it so some males in the simulation had an increased chance of breeding. It also made it possible to see the impact of kinship relationships outside of the considered POP, HSP and GPGCP which would have been more difficult to model. Having separate implementations of the simulations and models also helped to reveal issues during development. Simulations also allowed for exploring ideas and scenarios such as assessing the feasibility of CKMR on grey seals as was done in Chapter 4 that otherwise would have required years of sampling work and very expensive lab work.

Chapters 3 and 4 were centred around CKMR. Originally the work described in Chapter 3 was anticipated to be applied to the same set of data used in Ruzzante, McCracken, Førland, et al. (2019). However, the existing microsatellite based genotyping does not provide enough information to reliably identify the HSPs required to be able to estimate the variance in number of offspring. Ideally one outcome of this

thesis would be to take what was developed in 3 and apply it to a real world dataset to see how it performs and determine any potential limitations that might exist.

Another possible extension of Chapter 3 would be to investigate if it is possible to estimate the total variance in reproductive success directly by considering the number of sibling pairs that occur over an individual's entire lifetime rather than a single year. This would avoid having to make an assumption about the mean-variance relationship between age classes as well potentially avoid complications arising from non-independence in reproduction between years. However, this method is complicated by the fact that for sibling pairs the ages of parents is unknown and perhaps this alone makes infeasible.

Currently there is an ongoing project between DFO and Dalhousie to apply CKMR to the Atlantic halibut population managed by DFO. The lessons and experience gained over the development of this thesis are useful to that project and future work involves applying those lessons there. Similarly, the work presented in Chapter 4 could be extended to further investigate what the optimal sampling strategy and further improvements to the model might be.

The most important contribution of this thesis is the work described in Chapter 3. We showed explicitly the link between N_e and CKMR by using within-cohort comparisons. This link had previously been hinted at in Bravington et al. (2016) and in Waples and Feutry (2022) but was not fully explored. Not only did we uncover the link but we also showed that it was also possible to use just CKMR to estimate N_e as well provide estimates of the variance in number of offspring. Something that was not possible beforehand.

This thesis furthers the development of statistical methods and analysis for fisheries stock assessments. With climate change seemingly becoming a more visible reality with each passing day and the consequences of its impact on ocean ecosystems not fully realized there is more pressing need for more accurate and capable methods.

To do so future stock assessment models will likely need to incorporate new sources of data and techniques such as spatial data and CKMR as was discussed here. While incorporating them may be a challenge hopefully the work presented here gives some indication that it is worthwhile to do so.

Bibliography

[1] S. Aanes and J. H. Vølstad. "Efficient statistical estimators and sampling strategies for estimating the age composition of fish". In: *Canadian Journal of Fisheries and Aquatic Sciences* 72.6 (2015), pages 938–953. DOI: 10.1139/cjfas-2014-0408.

[2] W. H. Aeberhard, J. Mills Flemming, and A. Nielsen. "Review of state-space models for fisheries science". In: *Annual Review of Statistics and Its Application* 5 (2018), pages 215–235. DOI: 10.1146/annurev-statistics-031017-100427.

[3] A. Agresti. *Categorical data analysis*. Volume 482. John Wiley & Sons, 2003. DOI: 10.1002/0471249688.

[4] *Atlantic Mackerel Commercial Fishery and Bait Closure*. https://www.nfl.dfo-mpo.gc.ca/en/node/1071. Accessed: 2023-09-12. 2022.

[5] J. Babyn et al. "A Gaussian field approach to generating spatial age length keys". In: *Fisheries Research* 240 (2021), page 105956. ISSN: 0165-7836. DOI: https://doi.org/10.1016/j.fishres.2021.105956. URL: https://www.sciencedirect.com/science/article/pii/S0165783621000849.

[6] H. Bakka. "How to solve the stochastic partial differential equation that gives a Matérn random field using the finite element method". Mar. 2018.

[7] H. Bakka. *Mesh Creation including Coastlines*. URL: https://haakonbakkagit.github.io/btopic104.html.

[8] H. Bakka, H. Rue, et al. "Spatial modelling with R-INLA: A review". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 10.6 (2018).

[9] H. Bakka, J. Vanhatalo, J. Illian, et al. "Accounting for physical barriers in species distribution modeling with non-stationary spatial random effects". In: *arXiv preprint arXiv:1608.03787* (2016).

[10] H. Bakka, J. Vanhatalo, J. B. Illian, et al. "Non-stationary Gaussian models with physical barriers". In: *Spatial Statistics* (2019). DOI: 10.1016/j.spasta.2019.01.002.

[11] C. W. Berg and K. Kristensen. "Spatial age-length key modelling using continuation ratio logits". In: *Fisheries Research* 129 (2012), pages 119–126. DOI: 10.1016/j.fishres.2012.06.016.

[12] R. J. Beverton and S. J. Holt. *On the dynamics of exploited fish populations*. Volume 11. Springer Science & Business Media, 2012.

[13] W. Bowen, S. J. Iverson, et al. "Reproductive performance in grey seals: age-related improvement and senescence in a capital breeder". In: *Journal of Animal Ecology* 75.6 (2006), pages 1340–1351.

[14] W. Bowen, J. McMillan, and W. Blanchard. "Reduced population growth of gray seals at Sable Island: evidence from pup production and age of primiparity". In: *Marine Mammal Science* 23.1 (2007), pages 48–64.

[15] J. Brattey, D. Porter, and C. George. "Exploitation rates and movements of Atlantic cod (Gadus morhua) in NAFO Subdiv. 3Ps based on tagging experiments conducted during 1997". In: (2002).

[16] M. V. Bravington, H. J. Skaug, and E. C. Anderson. "Close-Kin Mark-Recapture". In: *Statistical Science* 31.2 (May 2016), pages 259–274. ISSN: 0883-4237. DOI: 10.1214/16-sts552. URL: http://dx.doi.org/10.1214/16-STS552.

[17] N. G. Cadigan. "A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates". In: *Canadian*

Journal of Fisheries and Aquatic Sciences 73.2 (2015), pages 296–308. DOI: 10.1139/cjfas-2015-0047.

[18] H. Caswell. *Matrix population models*. Volume 1. Sinauer Sunderland, MA, USA, 2000.

[19] D. G. Chapman. "Some properties of the hypergeometric distribution with applications to zoological censuses". In: *Univ. Calif. Stat.* 1 (1951), pages 60–131.

[20] W. G. Cochran. *Sampling Techniques*. John Wiley& Sons Inc., 1977.

[21] J. F. Crow, M. Kimura, et al. "introduction to population genetics theory". In: (1970).

[22] E. Dalley and J. Anderson. "Age-dependent distribution of demersal juvenile Atlantic cod (Gadus morhua) in inshore/offshore northeast Newfoundland". In: *Canadian Journal of Fisheries and Aquatic Sciences* 54.S1 (1997), pages 168–176. DOI: 10.1139/f96-171.

[23] R. Dalton. "Animal-rights group sues over'disturbing'work on sea lions". In: *Nature* 436.7049 (2005), pages 315–316.

[24] J. N. Darroch. "The multiple-recapture census: II. Estimation when there is immigration or death". In: *Biometrika* 46.3/4 (1959), pages 336–351.

[25] J. N. Darroch. "The multiple-recapture census: I. Estimation of a closed population". In: *Biometrika* 45.3/4 (1958), pages 343–359.

[26] C. den Heyer. personal communication. 2023.

[27] C. den Heyer, W. D. Bowen, et al. "Contrasting trends in gray seal (Halichoerus grypus) pup production throughout the increasing northwest Atlantic metapopulation". In: *Marine Mammal Science* 37.2 (2021), pages 611–630.

[28] C. den Heyer and W. Bowen. *Estimating changes in vital rates of Sable Island grey seals using mark-recapture analysis*. Canadian Science Advisory Secretariat (CSAS), 2017.

[29] S. Engen, R. Lande, and B.-E. Saether. "Effective size of a fluctuating age-structured population". In: *Genetics* 170.2 (2005), pages 941–954.

[30] S. Engen, R. Lande, and B.-E. Saether. "Effective Size of a Fluctuating Age-Structured Population: Figure 1.—". In: *Genetics* 170.2 (Apr. 2005), pages 941–954. ISSN: 1943-2631. DOI: 10.1534/genetics.104.028233. URL: http://dx.doi.org/10.1534/genetics.104.028233.

[31] M. P. Fahay et al. "Essential fish habitat source document. Atlantic cod, Gadus morhua, life history and habitat characteristics". In: (1999).

[32] J. Felsenstein. "Inbreeding and variance effective numbers in populations with overlapping generations". In: *Genetics* 68.4 (1971), page 581.

[33] J. Felsenstein. "Theoretical evolutionary genetics". In: *University of Washington, Seattle* (2005).

[34] R. Fisher. "Stage of development as a factor influencing the variance in the number of offspring, frequency of mutants and related quantities". In: *Annals of Eugenics* 9.4 (1939), pages 406–408.

[35] D. Fournier and C. P. Archibald. "A general theory for analyzing catch at age data". In: *Canadian Journal of Fisheries and Aquatic Sciences* 39.8 (1982), pages 1195–1207.

[36] A. Fridriksson. "On the calculation of age-distribution within a stock of cod by means of relatively few age-determinations as a key to measurements on a large scale". In: *Rapports Et Proces-Verbaux Des Reunions, Conseil International Pour l'Exploration De La Mer* 86 (1934), pages 1–5.

[37] R. J. Gray. "Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis". In: *Journal of the American Statistical Association* 87.420 (1992), pages 942–951. DOI: 10.1080/01621459.1992.10476248.

[38] G. Gudmundsson. "Time series analysis of catch-at-age observations". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 43.1 (1994), pages 117–126.

[39] J. Gulland. *Estimation of Mortality Rates*. Technical report. International Council for the Exploration of the Sea, 1965.

[40] M. Hammill et al. "Grey Seal Abundance in Canadian Waters and Harvest Advice". In: (2023).

[41] F. E. Harrell. "Regression modeling strategies". In: *as implemented in R package 'rms' version* 3.3 (2014). DOI: 10.1007/978-1-4757-3462-1.

[42] R. Hilborn and D. Ovando. "Reflections on the success of traditional fisheries management". In: *ICES journal of Marine Science* 71.5 (2014), pages 1040–1046. DOI: 10.1093/icesjms/fsu034.

[43] W. G. Hill. "A note on effective population size with overlapping generations". In: *Genetics* 92.1 (1979), pages 317–322.

[44] R. Hillary et al. "Genetic relatedness reveals total population size of white sharks in eastern Australia and New Zealand". In: *Scientific reports* 8.1 (2018), pages 1–9.

[45] T. Huxley. *Inaugural Meeting of the Fishery Congress: Address Delivered June 18, 1883*. 1883. URL: https://books.google.ca/books?id=6RUtAAAAYAAJ.

[46] D. Ings et al. *Assessing the status of the cod (Gadus morhua) stock in NAFO Subdivision 3Ps in 2018*. Fisheries & Oceans Canada, Science, Canadian Science Advisory Secretariat, 2019.

[47] D. Johnson. "American plaice, Hippoglossoides platessoides, life history and habitat characteristics". In: *NOAA Tech. Mem. NMFS-NE* 187 (2004).

[48] G. M. Jolly. "Explicit estimates from capture-recapture data with both death and immigration-stochastic model". In: *Biometrika* 52.1/2 (1965), pages 225–247.

[49] H. Julius Skaug. "Allele-sharing methods for estimation of population size". In: *Biometrics* 57.3 (2001), pages 750–756.

[50] K. Kristensen. *"R" bindings for "TMB" [R package RTMB version 1.2]*. Aug. 2023. URL: https://cran.r-project.org/package=RTMB.

[51] K. Kristensen et al. "TMB: Automatic differentiation and laplace approximation". In: *Journal of Statistical Software* 70.5 (2016), pages 1–21. DOI: 10.18637/jss.v070.i05.

[52] T. Kvist, H. Gislason, and P. Thyregod. "Using continuation-ratio logits to analyze the variation of the age composition of fish catches". In: *Journal of applied statistics* 27.3 (2000), pages 303–319. DOI: 10.1080/02664760021628.

[53] D. C. Lidgard et al. "State-dependent male mating tactics in the grey seal: the importance of body size". In: *Behavioral Ecology* 16.3 (2005), pages 541–549.

[54] F. Lindgren and H. Rue. "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". In: *Journal of the Royal Statistical Society. Series B* (2011). DOI: 10.1111/j.1467-9868.2011.00777.x.

[55] L. R. Lloyd-Jones et al. "Close-kin mark-recapture informs critically endangered terrestrial mammal status". In: *Scientific reports* 13.1 (2023), page 12512.

[56] S. L. Lohr. *Sampling: design and analysis*. Nelson Education, 2009. DOI: 10.1201/9780429296284.

[57] T. J. Miller and B. C. Stock. *The Woods Hole Assessment Model (WHAM)*. Version 1.0. 2020. URL: https://timjmiller.github.io/wham/.

[58] M. J. Morgan et al. *Assessing the status of the American Plaice (Hippoglossoides platessoides) stock in NAFO Subdivision 3Ps in 2019*. Fisheries & Oceans Canada, Science, Canadian Science Advisory Secretariat, 2020.

[59] M. Morgan. "Interactions between substrate and temperature preference in adult American plaice (Hippoglossoides platessoides)". In: *Marine & Freshwater Behaviour & Phy* 33.4 (2000), pages 249–259. DOI: 10.1080/10236240009387096.

[60] R. A. Myers, J. A. Hutchings, and N. J. Barrowman. "Why do fish stocks collapse? The example of cod in Atlantic Canada". In: *Ecological applications* 7.1 (1997), pages 91–106.

[61] A. Nielsen and C. W. Berg. "Estimation of time-varying selectivity in stock assessments using state-space models". In: *Fisheries Research* 158 (2014), pages 96–101.

[62] S. R. Noren et al. "Body condition at weaning affects the duration of the post-weaning fast in gray seal pups (Halichoerus grypus)". In: *Physiological and Biochemical Zoology* 81.3 (2008), pages 269–277.

[63] J. K. Parrish. "Using behavior and ecology to exploit schooling fishes". In: *Environmental Biology of Fishes* 55.1-2 (1999), pages 157–181.

[64] T. A. Patterson et al. "Rapid assessment of adult abundance and demographic connectivity from juvenile kin pairs in a critically endangered species". In: *Science Advances* 8.51 (Dec. 2022). DOI: 10.1126/sciadv.add1679. URL: https://doi.org/10.1126%2Fsciadv.add1679.

[65] K. H. Pollock. "A capture-recapture design robust to unequal probability of capture". In: *The Journal of Wildlife Management* 46.3 (1982), pages 752–757.

[66] S. Prystupa et al. "Population abundance in arctic grayling using genetics and close-kin mark-recapture". In: *Ecology and evolution* 11.9 (2021), pages 4763–4773.

[67] A. E. Punt. "Modelling recruitment in a spatial context: A review of current approaches, simulation evaluation of options, and suggestions for best practices". In: *Fisheries Research* 217 (2019), pages 140–155. DOI: 10.1016/j.fishres.2017.08.021.

[68] A. E. Punt, A. Dunn, et al. "Essential features of the next-generation integrated fisheries stock assessment package: a perspective". In: *Fisheries Research* 229 (2020), page 105617.

[69] A. E. Punt, M. Haddon, and G. N. Tuck. "Which assessment configurations perform best in the face of spatial heterogeneity in fishing mortality, growth and recruitment? A case study based on pink ling in Australia". In: *Fisheries research* 168 (2015), pages 85–99. DOI: 10.1016/j.fishres.2015.04.002.

[70] T. J. Quinn. "Ruminations on the development and future of population dynamics models in fisheries". In: *Natural Resource Modeling* 16.4 (2003), pages 341–392.

[71] P. M. Regular et al. "SimSurvey: An R package for comparing the design and analysis of surveys by simulating spatially-correlated populations". In: *PloS one* 15.5 (2020), e0232822. DOI: 10.1371/journal.pone.0232822.

[72] A. Rindorf and P. Lewy. "Analyses of length and age distributions using continuation-ratio logits". In: *Canadian Journal of Fisheries and Aquatic Sciences* 58.6 (2001), pages 1141–1152. DOI: 10.1139/f01-062.

[73] S. M. Ross. *Introduction to probability models*. Academic press, 2014.

[74] S. P. Rossi et al. "Forecasting the response of a recovered pinniped population to sustainable harvest strategies that reduce their impact as predators". In: *ICES Journal of Marine Science* 78.5 (2021), pages 1804–1814.

[75] H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005. DOI: 10.1201/9780203492024.

[76] D. E. Ruzzante, G. R. McCracken, B. Førland, et al. "Validation of close-kin mark–recapture (CKMR) methods for estimating population abundance". In: *Methods in Ecology and Evolution* 10.9 (July 2019). Edited by R. Altwegg, pages 1445–1453. ISSN: 2041-210X. DOI: 10.1111/2041-210x.13243. URL: http://dx.doi.org/10.1111/2041-210X.13243.

[77] D. E. Ruzzante, G. R. McCracken, S. Parmelee, et al. "Effective number of breeders, effective population size and their relationship with census size in an iteroparous species, Salvelinus fontinalis". In: *Proceedings of the Royal Society B: Biological Sciences* 283.1823 (Jan. 2016), page 20152601. ISSN: 1471-2954. DOI: 10.1098/rspb.2015.2601. URL: http://dx.doi.org/10.1098/rspb.2015.2601.

[78] G. A. Seber. "A note on the multiple-recapture census". In: *Biometrika* 52.1/2 (1965), pages 249–259.

[79] Y. Sharma et al. "Close-kin mark-recapture methods to estimate demographic parameters of mosquitoes". In: *PLOS Computational Biology* 18.12 (Dec. 2022). Edited by P. A. Hancock, e1010755. DOI: 10.1371/journal.pcbi.1010755. URL: https://doi.org/10.1371%2Fjournal.pcbi.1010755.

[80] S. Smith and G. Somerton. *STRAP: A User-Oriented Computer Analysis System for Groundfish Research Trawl Survey Data*. Technical report. Department of Fisheries and Oceans, Sept. 1981.

[81] T. D. Smith. *Scaling fisheries: the science of measuring the effects of fishing, 1855-1955*. Cambridge University Press, 1994.

[82] T. Stari et al. "Smooth age length keys: observations and implications for data collection on North Sea haddock". In: *Fisheries Research* 105.1 (2010), pages 2–12. DOI: 10.1016/j.fishres.2010.02.004.

[83] P. J. Sullivan. "A Kalman filter approach to catch-at-length analysis". In: *Biometrics* (1992), pages 237–257.

[84] E. A. Thompson. "The estimation of pairwise relationships." In: *Annals of human genetics* 39.2 (1975), pages 173–188.

[85] R. Thomson et al. "Close kin mark recapture for School Shark in the SESSF". In: *FRDC report for project* 2014/024 (2020), page 108.

[86] J. T. Thorson and L. A. Barnett. "Comparing estimates of abundance trends and distribution shifts using single-and multispecies models of fishes and biogenic habitat". In: *ICES Journal of Marine Science* 74.5 (2017), pages 1311–1321. DOI: 10.1093/icesjms/fsw193.

[87] J. T. Thorson, A. O. Shelton, et al. "Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes". In: *ICES Journal of Marine Science* 72.5 (2015), pages 1297–1310. DOI: 10.1093/icesjms/fsu243.

[88] J. T. Verma. *30 years after the moratorium, what have we really learned about cod and science?* https://www.cbc.ca/news/canada/newfoundland-

labrador/verma-fisheries-science-moratorium-history-1.6513310. Accessed: 2023-09-13. 2022.

[89] P. J. Verweij and H. C. Van Houwelingen. "Penalized likelihood in Cox regression". In: *Statistics in medicine* 13.23-24 (1994), pages 2427–2436. DOI: 10.1002/sim.4780132307.

[90] L. Von Bertalanffy. "A quantitative theory of organic growth (inquiries on growth laws. II)". In: *Human biology* 10.2 (1938), pages 181–213.

[91] S. Wacker et al. "Considering sampling bias in close-kin mark–recapture abundance estimates of Atlantic salmon". In: *Ecology and Evolution* 11.9 (2021), pages 3917–3932.

[92] J. Wang. "Sibship reconstruction from genetic data with typing errors". In: *Genetics* 166.4 (2004), pages 1963–1979.

[93] R. S. Waples, T. Antao, and G. Luikart. "Effects of overlapping generations on linkage disequilibrium estimates of effective population size". In: *Genetics* 197.2 (2014), pages 769–780.

[94] R. S. Waples, C. Do, and J. Chopelet. "Calculating N e and N e /N in age-structured populations: a hybrid Felsenstein-Hill approach". In: *Ecology* 92.7 (2011), pages 1513–1522. ISSN: 00129658, 19399170. URL: http://www.jstor.org/stable/23035104 (visited on 05/08/2023).

[95] R. S. Waples and P. Feutry. "Close-kin methods to estimate census size and effective population size". In: *Fish and Fisheries* 23.2 (Mar. 2022), pages 273–293. DOI: 10.1111/faf.12615. URL: https://doi.org/10.1111%2Ffaf.12615.

[96] S. N. Wood, M. V. Bravington, and S. L. Hedley. "Soap film smoothing". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.5 (2008), pages 931–955. DOI: 10.1111/j.1467-9868.2008.00665.x.

[97] B. Worm et al. "Rebuilding global fisheries". In: *science* 325.5940 (2009), pages 578–585. DOI: 10.1126/science.1173146.

Appendix A

Simulation Code for Spatial ALKs

The simulation study uses largely the same process for `SimSurvey` as was outlined and described in Regular et al. (2020). To assess the effectiveness of each aging method spatially (as described in Figure 2.3) the lengths and ages of every single fish needs to be simulated. `SimSurvey` only applies individual lengths as fish are sampled, only storing the number of fish of each age in the simulation grid cells which greatly reduces storage costs. To have the lengths available for every single fish `SimSurvey` was modified to generate fish lengths from the growth curve at the same time as the population is generated. These lengths are then kept as fish are distributed spatially into the simulation grid cells. The modified functions allowing all individual fish lengths to be known were created in a private fork of the `SimSurvey` package.

In addition, the Age-year-space covariance discussed in Appendix S3 of Regular et al. (2020) was instead obtained using a GMRF approximation with support for physical barriers as described in Section 2.2.2.1. A precision matrix $\boldsymbol{Q}$ is generated from a mesh and a specified set of hyperparameters. This is then approximated for each grid point in the `SimSurvey` simulation. This has the benefit over the default `SimSurvey` method of constraining the simulation to also have to abide by any physical barriers present and also provides a speed boost by using a GMRF approximation when the mesh has less nodes than there are cells in the grid. An example of a mesh used in the simulation is shown in Figure A.1 and example of the true abundance at age for the simulation referred to in Figures 2.3a and 2.2 is shown in Figure A.2.

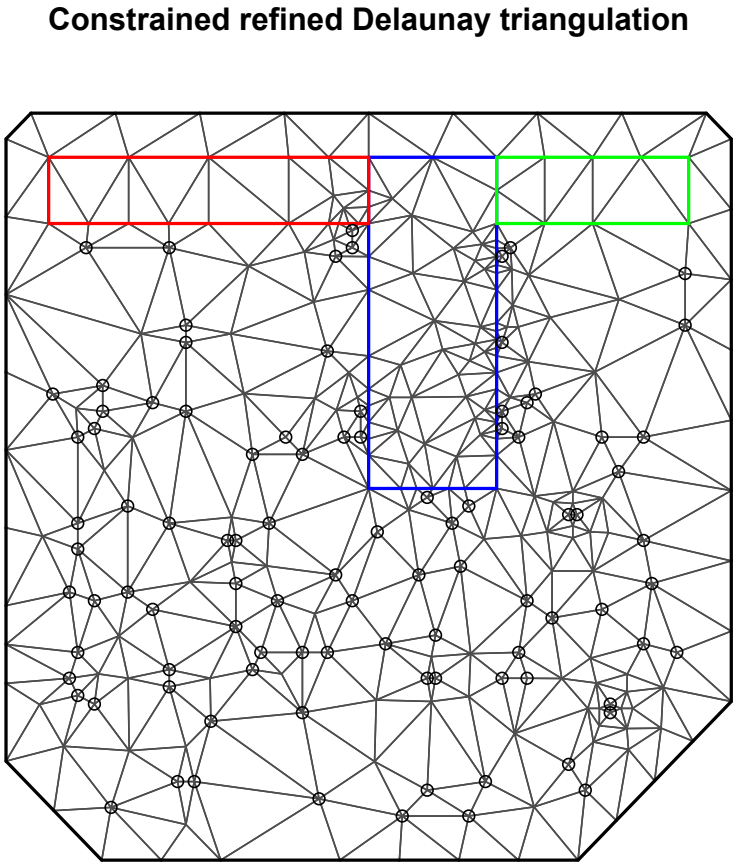


Figure A.1: An example of a mesh of the simulated survey area for one simulation. The coloured rectangles make up the boundary area, circles are sampled locations. Triangles form the mesh and each vertex is an element in the covariance matrix.

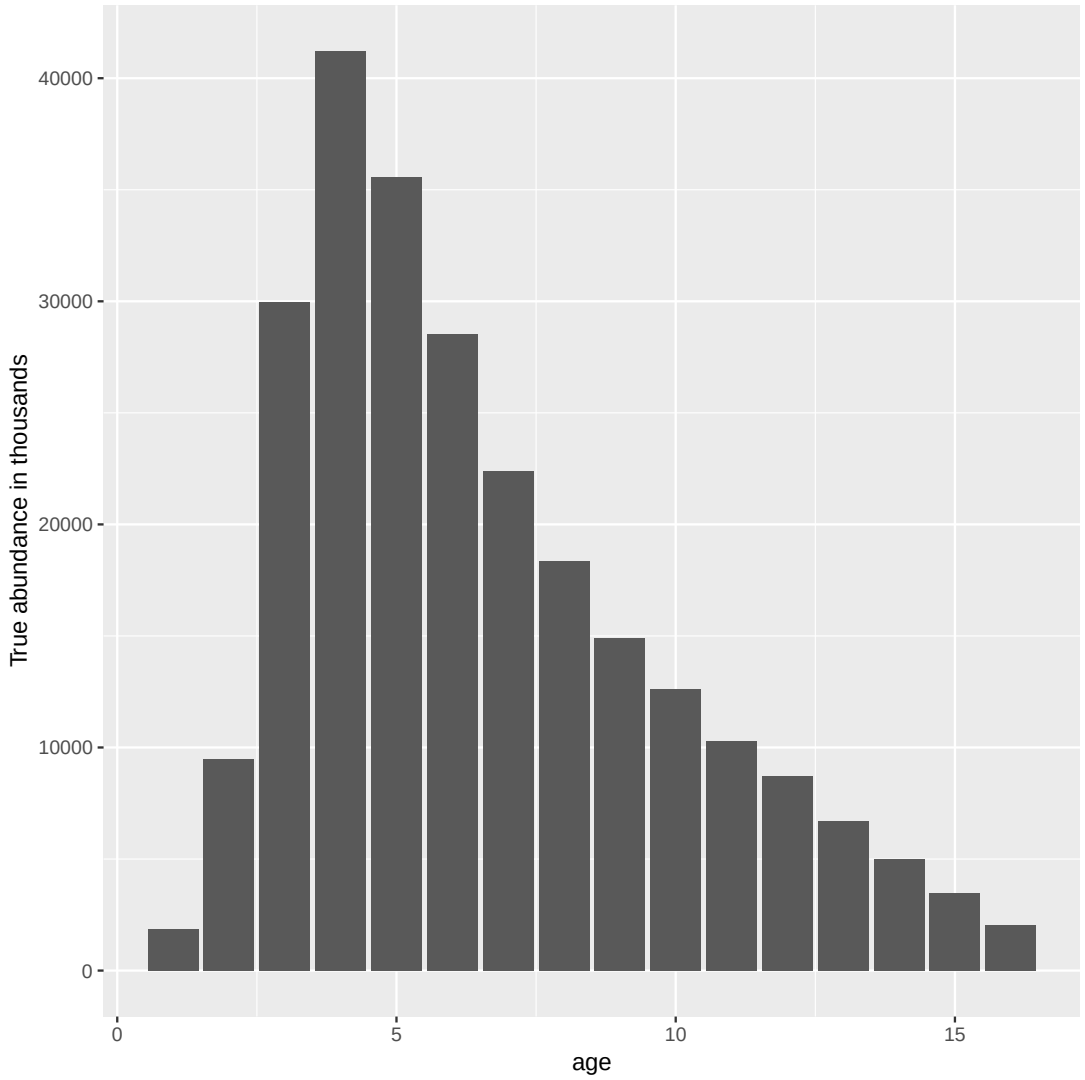


Figure A.2: The true abundance at age (in thousands) for the simulation used in Figures 2.3 & 2.2.

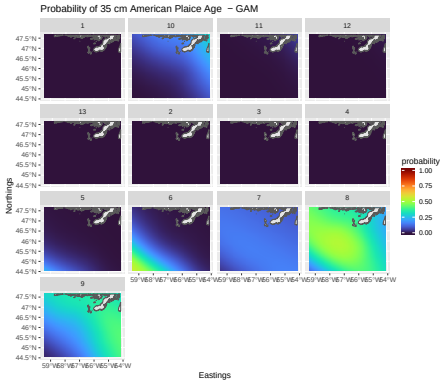


Figure A.3: The Probability of an American Plaice being each age class in the study area given a length of 35 cm as predicted by the GAM model.

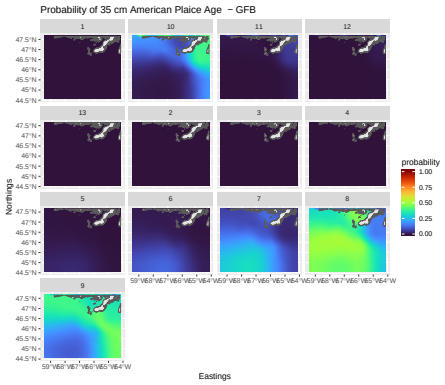


Figure A.4: The Probability of an American Plaice being each age class in the study area given a length of 35 cm as predicted by the GFB model.

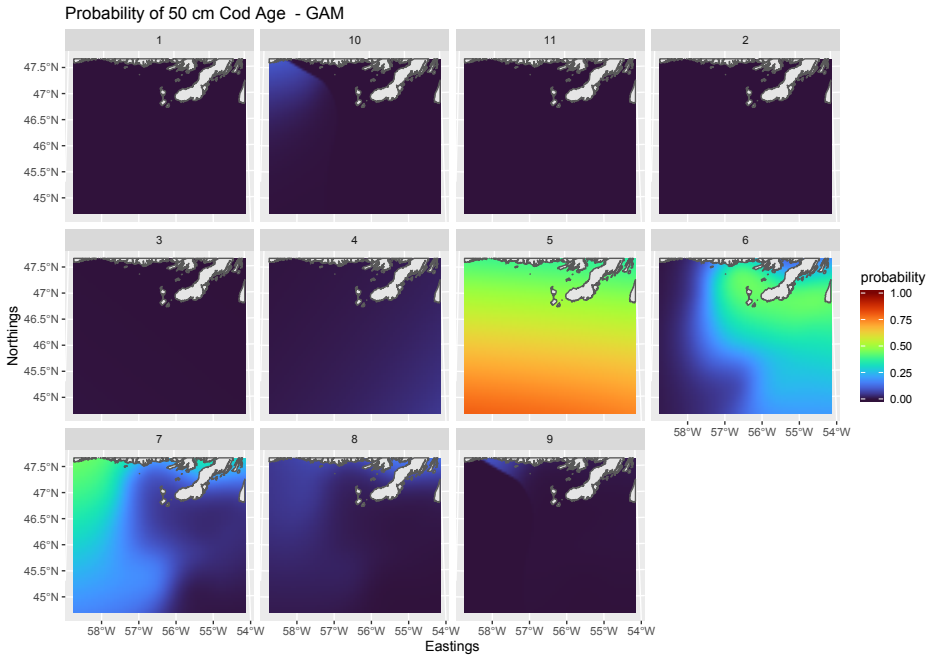


Figure A.5: The Probability of a Cod being each age class in the study area given a length of 50 cm as predicted by the GAM model.

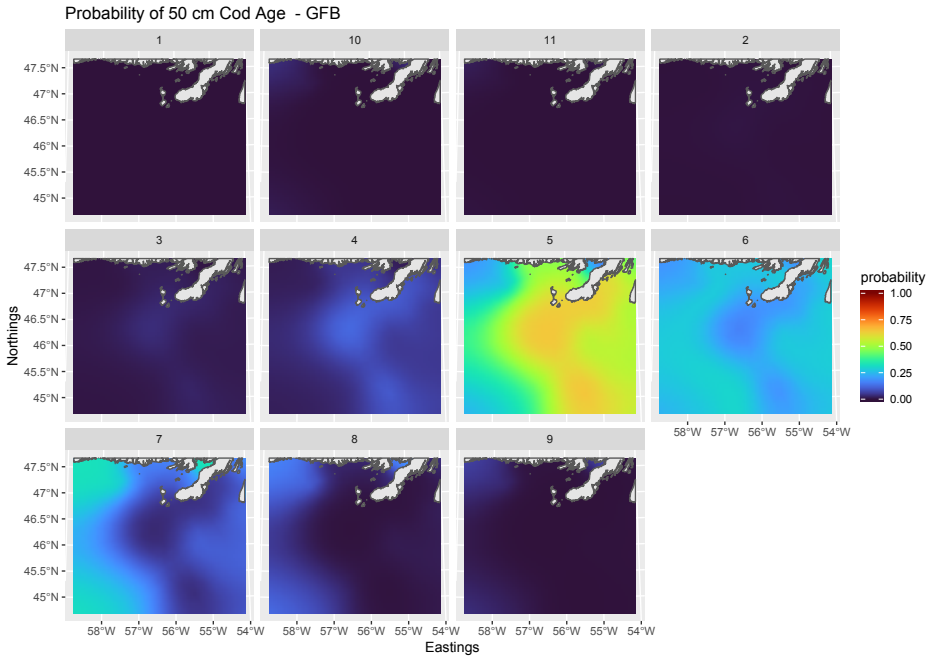
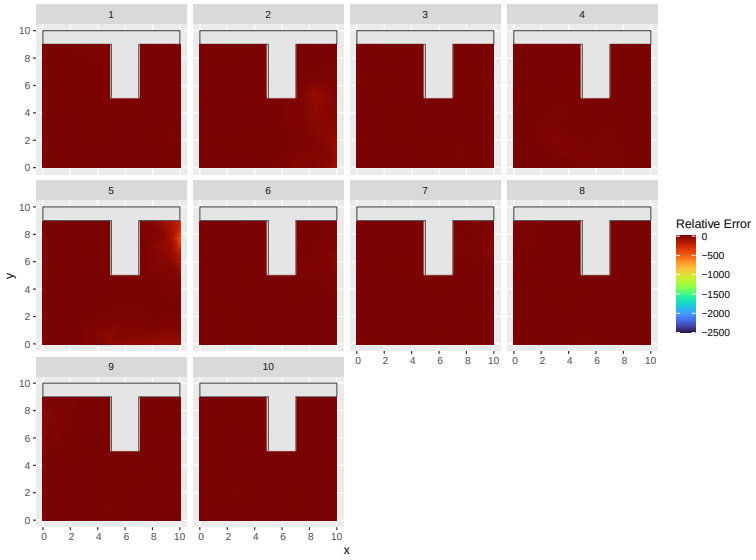
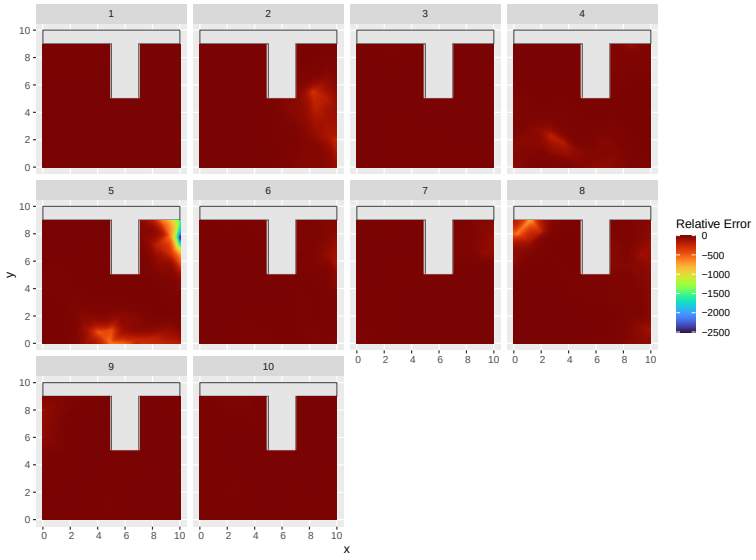


Figure A.6: The Probability of a Cod being each age class in the study area given a length of 50 cm as predicted by the GFB model.



(a) Relative error (true minus predicted over true) for the GFB model. A new spatial ALK is generated and applied to each simulation grid cell.



(b) Relative error (true minus predicted over true) from the traditional ALK. The same global traditional ALK was applied to each simulation grid cell. The non-spatial CRL model results in a very similar plot to the traditional ALK.

Figure A.7: The relative error between the predicted proportions and the true proportions in each simulation cell. The median relative error for the GFB model is -0.1334 and -0.4038 for the traditional ALK. Both methods have the smallest relative error occur on the right-hand side for age 5 fish which is -467.9 for the GFB model and -2443.91 for the traditional ALK.

Appendix B

Probability of within-cohort siblings

Suppose we have two individuals i and j born in the same cohort that are sampled at ages x and y where $y > x \geq a$ with a being the age at which survival between individuals become independent of one another. Let R_{xm} and R_{ym} be the number of individuals from mother m surviving to ages x and y, R_{x+} and R_{y+} be the total size of the cohort at ages x and y. The probability of the two individuals sharing the same mother is

$$Pr(i \text{ and } j \text{ share a mother}|g = h) = \sum_{m}^{N_f} \frac{R_{ym}}{R_{y+}} \times \frac{R_{xm} - 1}{R_{x+} - 1} \tag{B.1}$$

Since survival between individuals older than a are independent we can rewrite the above solely in terms of R_{xm} and R_{x+}, if ϕ_y is the probability of an individual surviving from age x to y then

$$Pr(i \text{ and } j \text{ share a mother}|g = h) = \sum_{m}^{N_f} \frac{\phi_y R_{xm}}{\phi_y R_{x+}} \times \frac{R_{xm} - 1}{R_{x+} - 1} = \frac{1}{R_{x+}(R_{x+} - 1)} \sum_{m} (R_{xm}^2 - R_{xm}). \tag{B.2}$$

If abundance is large then $R_{x+}(R_{x+} - 1) \approx (N_f E[R_{xm}])^2$ where N_f is the number of mothers in that cohort year and we can use that to replace the denominator. Then if we take expectations

$$Pr(i \text{ and } j \text{ share a mother}|g = h) \propto E[R_{xm}^2] - E[R_{xm}] = (R_{xm}) + E[R_{xm}](E[R_{xm} - 1]) \tag{B.3}$$

Let ϕ be the probability of survival from age a to x. Since survival is independent among ages greater than or equal to a then $R_{xm} \sim \text{Bin}(R_{am}, \phi)$ and combined with the law of total variance

$$Var(R_{xm}) = E_{R_{am}}[Var(R_{xm}|R_{am})] + Var_{R_{am}}(E[R_{xm}|R_{am}]) \tag{B.4}$$

$$= E_{R_{am}}[R_{am}\phi(1-\phi)] + Var_{R_{am}}(R_{am}\phi) \tag{B.5}$$

$$= \mu_a\phi(1-\phi) + \sigma_a^2\phi^2 \tag{B.6}$$

where μ_a and σ_a^2 are the mean and variance of the number of maternal offspring surviving to age a. Then

$$Pr(i \text{ and } j \text{ share a mother}|g = h) = \frac{\mu_a\phi(1-\phi) + \sigma_a^2\phi^2 + \phi\mu_a(\phi\mu_a - 1)}{N_f\phi^2\mu_a^2} \tag{B.7}$$

$$= \frac{\phi^2\sigma_a^2 + \phi^2\mu_a(\mu_a - 1)}{N_f\phi^2\mu_a^2} \tag{B.8}$$

$$= \frac{1}{N_f} \times \left(1 + \frac{\sigma_a^2 - \mu_a}{\mu_a^2}\right). \tag{B.9}$$

The above does not depend on x or y so the probability that two individuals are a sibling pair from the same cohort does not depend on their age at capture so long as those ages are greater than or equal to the age at which survival becomes independent between individuals. Thus we are free to select any suitable reference age where survival between individuals is independent.

Appendix C

The link between reproductive success and N_e

Here we show an intuitive heuristic derivation of variance N_e to illustrate the link between it and the mean and variance of lifetime reproductive success. Variance N_e is typically defined as the number of individuals in an ideal population that would generate the same variance for the change in allele frequency between two generations as in the original population. Here we instead consider the number of individuals in an ideal population with the same allele frequency in the next generation rather than the variance in allele frequency. Measuring the average allele frequency in one generation followed by the next still informs us about how the allele frequency changes from generation to generation.

Suppose we have an ancestral cohort of C aged year olds and there are N_C of them. Further suppose there is a rare allele in the population with allele frequency p. This results in there being pN_C individuals with the rare allele and since the probability two individuals with the rare allele breed together is negligible, we can ignore the possibility of it occurring. Suppose that X_i is the lifetime reproductive output of C year old individual i with mean μ_c and variance σ_c^2. Since only heterozygotes exist in the population with the rare allele the number of descendants that carry it from individual i is given by a binomial random variable with probability of success $1/2$ and size X_i, which we denote by Y_i. Then the frequency of individuals with the rare allele in the next generation is given by

$$P = \frac{\sum_{pN_c} Y_i}{\sum_{N_c} X_i}$$

since individuals without the rare allele contribute nothing. Similar to Hill (1979) we ignore terms of $O(p^2)$ and replace the denominator of P with it's expectation $N_c\mu_c$. The variance of P is approximately,

$$Var(P) \approx (N_c\mu_c)^{-2} p N_c Var(Y_i)$$

and the variance of Y_i are

$$Var(Y_i) = E_{X_i}[Var(Y_i|X_i)] + Var_{X_i}(E[Y_i|X_i])$$

$$= E_{X_i}\left[\frac{X_i}{4}\right] + Var_{X_i}\left(\frac{X_i}{2}\right) = \frac{\mu_c}{4} + \frac{\sigma_c^2}{4}$$

Then

$$Var(P) \approx (N_c\mu_c)^{-2} p N_c \frac{\mu_c + \sigma_c^2}{4} = \frac{p(\mu_c + \sigma_c^2)}{4N_c\mu_c^2}$$

Under the Wright-Fisher (WF) ideal case $\mu_c = \sigma_c^2 = 2$ and

$$Var_{WF}(P) = \frac{4p}{16N_{WF}} = \frac{p}{4N_{WF}}$$

and the effective population size ignoring overlapping generations is

$$N_e \approx \frac{p}{4Var(P)} = \frac{N_c\mu_c^2}{\mu_c + \sigma_c^2} = \frac{N_c\mu_c}{1 + \frac{\sigma_c^2}{\mu_c}}.$$

We use the same logic as Hill (1979) to account for overlapping generations and adjust by the generation length L to give

$$N_e \approx \frac{N_c\mu_c L}{1 + \frac{\sigma_c^2}{\mu_c}} \tag{C.1}$$

which only differs from the form given in Equation 3.1 by $1 - 1/N_c$ caused by terms of $O(p^2)$.

Appendix D

Leslie Matrix Approach to Population Dynamics

One way of representing a population's structure is to consider the demographics such as fecundity and survival by age. We can represent the number of individuals at age in year y by the vector $\boldsymbol{N}_y$ where each element is the number of individuals in each of the age classes. Individuals born in the same year are said to belong to the same cohort. In the deterministic case we can find the number of individuals at age in year $y+1$ using a Leslie matrix,

$$\boldsymbol{N}_{y+1} = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 & \dots & \beta_{A-1} & \beta_A \\ \phi_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \phi_2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \phi_3 & \dots & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & & \dots & \phi_{A-1} & 0 \end{bmatrix} \begin{bmatrix} N_{1,y} \\ N_{2,y} \\ N_{3,y} \\ \vdots \\ N_{A-1,y} \\ N_{A,y} \end{bmatrix}$$

where the β_i represent the per-captia fecundity of individuals in age class i and ϕ_i is the proportion of individuals in the ith age class that survive to the $i+1$ age class. Here A represents the oldest possible age class in the population. The growth rate of the population λ is the largest eigenvalue of the Leslie matrix (Caswell 2000).

Sometimes we may not be able to observe all of the age classes in a population or we may not want to avoid problems like non-independence of sampling of age classes or if survival of certain age classes are not independent. Instead of working with the first age class we can instead choose another age to use as our reference age. Suppose

we have a population where the age of maturity happens at the third age class (so $\beta_1 = \beta_2 = 0$) and we only start observing the population at the third age class. This means that we cannot find the values of ϕ_1 and ϕ_2 and the Leslie matrix of what we can observe is equivalent to

$$\boldsymbol{Les_3} = \begin{bmatrix} 0 & 0 & \beta'_1 = \phi_1\phi_2\beta_3 & \dots & \beta'_{A-1} = \phi_1\phi_2\beta_{A-1} & \beta'_A = \phi_1\phi_2\beta_A \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \phi_3 & \dots & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & & \dots & \phi_{A-1} & 0 \end{bmatrix}$$

as we only observe the number of individuals that have survived to the third age class. Here β'_a is the average number of offspring from individuals aged a that survive to the chosen reference age of three.

Even though we can not estimate age-specific survival for ages before our chosen reference age the two Leslie matrices here have identical Eigenvalues and hence growth rates. For the ages equal to and beyond the reference age the two Leslie matrices will result in the same numbers at age given the corresponding numbers in the year before.

Appendix E

Copyright Release

The following chapter was published under the following citation:

Chapter 2: A Gaussian Field Approach to Generating Spatial Age Length Keys

J. Babyn et al. "A Gaussian field approach to generating spatial age length keys". In: *Fisheries Research* 240 (2021), page 105956. ISSN: 0165-7836. DOI: https://doi.org/10.1016/j.fishres.2021.105956. URL: https://www.sciencedirect.com/science/article/pii/S0165783621000849

The author of published Elsevier articles retains the right to publish their articles within their dissertation. In this case, letters of permission are not required. (https://www.elsevier.com/about/policies/copyright).