**Expected Improvements in Precision when Integrating Opportunistic Close-Kin Mark-Recapture Data into Fisheries Stock Assessments**

Nicholas Fisch1

1Pacific Biological Station, Fisheries and Oceans Canada, Nanaimo, British Columbia, Canada

Abstract

Close-Kin Mark-Recapture (CKMR) sampling, by providing information on abundance and survival rates (and potentially other quantities), offers a promising new data source for fisheries stock assessments. To date and to my knowledge, most applications have focused on fitting standalone CKMR models outside of stock assessment models. Fewer studies have directly integrated the CKMR data into the stock assessment model. Sample design in order to achieve a desired precision is somewhat straightforward in simple CKMR models however when integrated within a full stock assessment model with many other data sources, the value of the data (in terms of a reduction in uncertainty of model estimates) is less clear. Herein I demonstrate, using self-test simulations, the expected improvements in accuracy and precision of derived quantities and estimated parameters within statistical catch-at-age models when opportunistic CKMR sampling is conducted and the data integrated within the assessment. By opportunistic CKMR sampling I mean to describe the genetic sampling of individuals that comprise the age composition data, such that increases in CKMR sampling would also increase the age composition samples (and vice versa). I examine the expected improvements (conditional on model being correct representation of the system) across three life history types (Cod-like, Flatfish-like, and Sardine-like) and different amounts of data available to the assessment, including the uncertainty and inclusion of an abundance index, and the sample size and time series length of CKMR and age composition samples.

abstract results

abstract discussion

Introduction

Close-Kin Mark-Recapture (CKMR, Bravington et al. 2016) offers a promising new source of information to fisheries stock assessments, which largely form the basis of commercial fisheries management throughout the world (Dichmont et al., 2016). Traditionally, stock assessments have been developed using routinely collected data from fisheries monitoring programs on total removals, age/length composition of the removals, and catch rates (Maunder and Punt, 2013). CKMR, given it allows for lethal sampling (e.g., sampling of the removals), may provide another data source to be routinely collected from commercial fisheries, contributing additional information on abundance and survival levels within fisheries stock assessments.

The theory of CKMR dates back to the early 2000s (Nielsen et al., 2001; Skaug, 2001) and since the publication of Bravington et al. (2016) has greatly increased in application within fisheries (see Casas and Saborido-Rey (2023) for a suite of examples). The theory operates primarily by making inference on the number of parent-offspring and half-sibling pairs (POPs and HSPs) expected from a sample of a population. At its most basal level the theory posits that the expected number of POPs from a sample of a population is inversely related to the spawning abundance of that population. That is, the more POPs found in samples from a population, the smaller the population likely is, and vice versa. For HSPs, inference is based on an unobserved shared parent of the sibling pair, who must give birth to the older sibling and then survive to give birth to the younger sibling. Thus the expected number of HSPs in a sample is inversely related to both the adult mortality rates in the population and the spawning abundance in the year of each siblings birth. I avoid an in depth presentation of the theory as it is covered extensively in Bravington et al. (2016) and subsequent publications.

The lower limit of the precision of an abundance estimate from a CKMR model can be relatively straightforward to compute for the simplest case. For example, the lower limit of the coefficient of variation (CV) of an abundance estimate for a POP model with a single sampling event can be expressed as (Bravington and Grewe, 2007)

Where *m* denotes the total sample size of potential POPs examined (equally split among juveniles and adults), and the true abundance of adults in the population. Within an integrated fisheries assessment, the expected precision of estimates is less tractable and does not have a simple analytical solution, as it depends on a variety of different data sources and a complex, highly-parameterized non-linear model. However, the tool of self-test simulations can be employed to examine what might be expected in terms of precision of abundance estimates (and other desired quantities) when CKMR data are integrated into the assessment. According to CKMR theory, we expect that the information content of the CKMR data will depend on the size of the spawning abundance of the population and on the sample size collected. However, it is less clear how much specific additional information it will provide when combined with other data sources collected in fisheries assessment, such as age composition data--which, given that CKMR data are aged, will increase commensurate with increases in CKMR data. In addition, within a fisheries assessment one might expect that even for cases where the population is very large and the CKMR sample size relatively small, zeroes in data (no pair matches) may help rule out the probability of a small population size, thereby potentially improving the precision of estimates.

In this study, I simulate a case where the age composition of the removals from populations are simultaneously observed for CKMR sampling. I refer to this opportunistic CKMR as little additional sampling effort is required other than genetic sampling of the age composition data that is already collected. This also has the effect of increasing the amount of composition data when the sample size of CKMR is increased. I therefore explore what one might expect in terms of improvements in model performance if CKMR data is collected and integrated within the stock assessment compared to identical assessments without CKMR data. I examine this comparison across three fish life history types (Cod-like, Flatfish-like, Sardine-like), in addition to across different amounts and quality of data such as the uncertainty in or the inclusion of an abundance index, and the sample size of CKMR and age-composition data.

Methods

The structure of the study was to have identical assessment models that were fit with and without CKMR data, so as to ascertain its value to the assessment. In the following I describe the population simulation operating model and the sampling model, including a separate component for the CKMR section.

Population simulation operating model

I chose to use similar population specifications and life-history scenarios to those used in Ono et al. (2015) and Hurtado-ferro et al. (2015) for simplicity and ease of comparison. The different life history scenarios include cod-like, flatfish-like, and sardine-like models. The main differences are that cod-like life history simulates a slow-growing, large maximum size, low natural mortality, low recruitment variability, old maximum age, and very abundant stock where the flatfish stock is much less abundant but has more recruitment variability, a greater steepness and growth rate. Finally the sardine-like model has the highest recruitment variability, a moderate abundance, rapid growth, high natural mortality, and low steepness. The parameter values and equations governing the model are given in Tables 1 and 2, respectively. The population model runs for 100 years, with a fishing time series beginning in year 26 (years 26-100 experience fishing). The fishing time series was simulated as a two way trip with a linear increase from 0 in year 25 to a fully selected fishing mortality that was >*F*msy () and achieved 85% of maximum sustainable yield (MSY) at equilibrium in year 85 of the time series, followed by a linear decrease to a fully selected fishing mortality that was <*F*msy () in year 100 and also achieved 85% of MSY at equilibrium.

One hundred simulation iterations were run for each population model specification (Table 3), with stochasticity included using log-normal draws for recruitment deviations in each year. A sampling model then sampled from the population and fishery exploitation model to simulate the data collection process.

Sampling model

Annual landings were simulated using a log-normal distribution with a standard deviation (SD) of 0.05. A fishery index was simulated using a lognormal distribution where the SD was specified as a treatment within the simulation design, taking values of 0.25 or 0.5 (Table 3). The catch age-composition available for select years throughout the time series was simulated using a multinomial distribution. Age composition data was simulated from years 26, 36, 46, 51, 56, 61, 66, 71, and annually from year 76 to 100. The number of samples prior to CKMR data collection began at *n*=30 in year 26 of the model (first year with fishing), increasing by 10 each decade until year 51, and then increasing by 10 every 5 years until year 76 where it was fixed at *n*=100 for the rest of the time series. This composition specification was similar to that used in Ono et al., (2015). Given CKMR data are aged samples, when CKMR data collection increased, as did the composition sample size, thus the composition sample size increased commensurate with CKMR data collection treatments (either 100, 1000, or 5000 sampled from the catch from the final 5, 10, or 20 years in the time series).

CKMR data

CKMR data were simulated from the terminal 5, 10, or 20 years of the time series (Table 3). Data were censused from the age composition for those years to make up the CKMR comparisons. That is, the age frequencies of the catch-at-age samples were those for which kinship comparisons were then made. Given the sampling year, the birth year of the older individual, and the birth year of the younger individual, the probability of being a POP or a HSP was calculated. The CKMR pair comparisons were aggregated to pairs with similar covariates (Table 4) and the number of POPs and HSPs for a given set of covariates was then simulated using a multinomial distribution. Note in the following that I do not distinguish “juveniles” and “adults” as the population is age structured with proportional maturity at age. Thus any individual whose age difference with another places them in a potentially mature age (at the year of birth of younger individual) has the potential to be a parent of the individual. I make comparisons between “younger” () and “older” () individuals in reference to both potential POPs and HSPs (rather than say, juveniles and adults). I limit comparisons to omit same-cohort siblings (i.e., age difference must be greater than 0).

CKMR Simulator/Estimator

*Parent-offspring pairs*

Given the birth year of an older individual (), the birth year of a younger individual (), and the year the older individual was sampled (), the probability of the older individual being a parent-offspring pair with the younger individual is calculated as

which describes the expected reproductive output of the potential parent in the year of younger individuals birth ( denotes age difference) divided by the total reproductive output in that year. A potential parent has to have been sampled after or on the year of younger individuals birth, because sampling is lethal in this case (as it is made up of the catch). Note that sex is not modeled and the total reproductive output is assumed to be made up of both males and females who contribute equally to spawning biomass. The 2 in the numerator reflects the fact that an individual has two parents. An alternative way of thinking about this probability, assuming a 50:50 sex ratio, is the probability that the older sampled individual is female is 0.5 and the probability that the female is a maternal-offspring pair with the younger individual is Thus, adding the probability of a maternal-offspring pair () and a paternal-offspring pair gives

*Half-sibling pairs*

For HSPs, given the parent is not directly observed, a sum is taken over the probabilities of all the potential parents of the pair alive at the year of *i*’s birth (hence ). The first term within the summation describes the expected reproductive output of a potential parent (age ) over the total reproductive output in the year of the *i*’s birth. The addition of is to account for the abundance alive at *i*’s birth, and assuming a 50:50 sex ratio, 0.5 in the numerator and denominator cancel out. The second term is the probability that a potential parent survives from the time of *i*’s (older individual) birth to the time of *j*’s (younger individual) birth (i.e., through the year prior to *j*’s birth), assuming that spawning happens at the very start of the year. So a potential parent has to survive the year of *i*’s birth up to the age difference of the potential pair. Once the survival of the potential parent is accounted for, the third term in the equation is the probability of a parent-offspring pair between that potential parent and the younger individual (), where the 4 in the numerator in this case accounts for the fact that the parent is unobserved and could be a mother or a father (, Hillary et al. 2018).

Treatment Overview

Table 3 depicts the levels of each treatment type, which were implemented in a full factorial design to examine all combinations of treatments. Treatments included the species life history type, the sample size of CKMR and composition data, the length of time CKMR or increased composition data was collected, and the inclusion or sampling variance of the abundance index.

Estimation Model

The estimation/assessment models were identical in structure to the population operating model. They began in year 26 of the time series (first year with data). The fixed-effect parameters estimated within the model included the unfished recruitment, the recruitment SD, a natural mortality parameter, two logistic selectivity parameters, catchability of the fishery index (if the index is included in assessment), and annual fully selected fishing mortality levels in each year (Table 1). Recruitment deviations in each year and those that make up the initial unfished abundance were treated as random effects. All other parameters were fixed at their true values from the operating model. The assessment models also fit the data using the same likelihoods used to generate data in the sampling model. In real-world situations, grandparent-grandchild pairs are difficult to distinguish genetically from HSPs (Bravington et al. 2016). In order to simulate a case where the chance of analyzing grandparent-grandchild pairs is minimal, HSP comparisons within the multinomial likelihood were restricted to comparisons whose age difference was <2 times the age at 50% maturity for each life history (3.02, 4.8, and 3.4 for Cod-like, Flatfish-like, and Sardine-like, respectively).

Model Fitting

Models were fit in a mixed-effects maximum marginal likelihood context in Template Model Builder (TMB, Kristensen et al., 2016) with recruitment deviations specified as random effects and the remainder of the fixed effect parameters identified in Table 1. TMB calculates the marginal negative log-likelihood given the fixed effect parameters using the Laplace approximation to integrate over random effects. Fixed effect parameters are estimated via minimizing the marginal negative log-likelihood within the program R using the nlminb function. Random effects and derived quantities are then predicted using empirical Bayes (Kristensen et al., [2016](javascript:;)). Starting parameter values for each simulation iteration were chosen by sampling from a uniform distribution with the bounds specified as 50% below and 50% above the true parameter value. Standard errors of parameter estimates were extracted in addition to those of derived quantities using the generalized delta method built into TMB.

Comparison

I focused the comparison of assessment models with and without CKMR data on the accuracy achieved between simulations and on the precision achieved within and between simulations. In addition, although the models were correctly specified, I make note of the bias in relative error for a subset of assessment models. The accuracy of point estimates between simulations was assessed by examining the median absolute relative error over simulations, with absolute relative error defined as

Where denotes a point estimate from the assessment and the true value for that quantity in simulation iteration *s*. The precision within simulations was assessed by examining the estimated standard errors for quantities within an individual simulation iteration, summarized using the median of the distribution of standard errors over simulations. The precision between simulations was assessed by examining the interquartile range (IQR) of the distribution of relative error over simulations, with relative error calculated as

I chose to focus the comparison on spawning stock biomass over the time series (*SSBy*), unfished recruitment (*R0*), and natural mortality (*M*) as they are of general interest in fisheries assessment and the CKMR data is thought to primarily inform abundance and survival (Eqs. 2-3). I summarize performance metric results by presenting the values obtained from assessment models fit with CKMR data divided by those obtained from the same assessment models fit without CKMR data (e.g., including CKMR data led to median AREs that were 80% of those from models without CKMR [20% improvement in accuracy]).

Results

*Bias*

Estimation models were effectively unbiased in *SSBy*, however were negatively biased in *R0* and *M* for each life history type. The level of bias increased from Cod to Flatfish to Sardine. For example, in one of the more data-poor scenarios – max composition sample size 100, index SD=0.5, no CKMR data – the Cod models had negligible bias levels with median relative errors of -0.01 and -0.02 for *R0* and *M*, respectively. This bias increased to -0.06 and -0.04 for Flatfish and further to -0.13 and -0.08 for Sardine (Supplemental Figure 6). This trend was present in all models and appeared to be due to the sparseness or lack of composition data at the beginning of the time series, as the models became unbiased when composition data was simulated in every year of the time series with sample sizes of at least 100 (Supplemental Figure 7).

*POPs and HSPs from operating and sampling model*

The number of pairs found differed greatly between the life-history types (Table 5). For each combination of life-history, sample size, and length of CKMR data collection, more HSPs were found than POPs. The cod-like model, largely due to the size of the spawning stock at the time data collection, resulted in very few POPs and HSPs found for most sampling models. For sample sizes less than 5,000, the mean (over simulations) number of pairs found was nearly always <1 (exception being HSPs at *n*=1000 for 20 years). The flatfish models resulted in the greatest number of pairs found, often in the thousands and tens of thousands. Even at the lowest sample sizes (100 collected for 5 years), the flatfish models produced >1 POP and >6 HSPs on average. The sardine models resulted in moderate numbers of pairs found, with few (<1 on average) found when sample sizes were 100 annually and hundreds found when samples sizes were 5,000 annually.

In terms of simply finding kin pairs, the sample size collected did appear more important than the length of data collection given 5,000 individuals sampled for 5 years resulted in considerably more pairs than the treatments that sampled 1,000 for 20 years in each scenario (and likewise for 1,000 sampled for 5 years compared to 100 sampled for 20 years). Although I note that the total sample sizes in these cases is not necessarily equivalent (e.g., 25,000 total fish vs 20,000), this result persisted when the sample sizes were equivalent in additional exploratory runs.

*General patterns*

As expected given the pair information, in terms of improvements in performance metrics, the CKMR data was most informative for the Flatfish models followed by the Sardine models and finally the Cod models. For *SSBy* performance metrics, the improvements (if present) manifested in terms of reductions in metrics during the latter portion of the time series (depending on the length of CKMR data collection) to a minimum (i.e., maximum improvement) followed by an increase in metrics to the terminal year. In what follows I make reference the minimum level as the level for a specific year where a performance metric *SSBy* exhibited the maximum improvement compared to a model with no CKMR. At their most informative (No index, flatfish life history, 20yrs of CKMR), integrating CKMR data into the stock assessment improved accuracy, precision within, and precision between simulations to ~10% of the levels achieved from estimation models fit without CKMR data for *SSBy* (the minimum level achieved in the time series, Figures 1-3). For *M* and *R0*, the maximum improvement was realized in the same treatment scenario and was ~30 and ~60% of the levels from estimation models fit without CKMR data, respectively (Figure 4). A pattern evident across most models was that CKMR data appeared to improve performance metrics on *M* more so than on *R0.*

*Inclusion and uncertainty of abundance index*

Compared to other treatment options, there appeared to be a more minor effect regarding the inclusion or standard deviation of the abundance index data. The inclusion of CKMR data did provide greater improvements in performance metrics to the models that did not include an abundance index time series compared to those that were fit with an index and an index SD of 0.25 (Figures 1-4). As an example, for the flatfish model fit without an index, at CKMR sample sizes of 1,000 fish collected for 5 years, the median ARE of SSB was 22% and 46% the level of models without CKMR data at the minimum and terminal year in the time series, respectively (year 70 and 76). For the same scenario where the models included an index with a SD of 0.25, the median ARE of SSB was 38% and 63% the level of models fit without CKMR data at the minimum and terminal year in the time series (same years). In the interest of brevity I focus the remainder of the results on the moderate case where the abundance index SD=0.5.

*Performance metrics*

*Accuracy*

For each life history type and duration of data collection, the CKMR data collection scenarios that collected 100 fish per year did not appear to improve the accuracy of *SSBy* point estimates (Figure 1). Although in these scenarios, at least for the flatfish models, the accuracy of *M* and *R0* estimates did appear to improve to 85-90% of the levels from models fit without CKMR. For the Cod models, even the largest CKMR sample sizes collected for 20 years did not appear to notably improve the accuracy of *SSBy*, *M*, or *R0* point estimates compared to models fit without CKMR.

For the Sardine models, improvements in accuracy were evident for scenarios which sampled 1,000 fish for 20 years or 5,000 fish for 5, 10, or 20 years. Improvements manifested in terms of reductions in median AREs for *SSBy* in the second half of the time series, and in *M* and *R0*. At their most informative (*n*=5,000 for 20 years), inclusion of CKMR data led to median ARE for *SSBy* that was 32% at its minimum (year 69) and 87% in the terminal year compared to models fit without CKMR data. For the same CKMR treatment, the accuracy of point estimates of *M* and *R0* in improved to 24% and 69% of the median ARE for models fit without CKMR data (Figure 4).

For the flatfish models, across all treatments where *n*>100, the accuracy of point estimates markedly improved for *M*, *R0*, and *SSBy* near the end of the time series compared to models fit with no CKMR data. As an example of a moderate treatment where *n*=1,000 for 10 years, median ARE in *SSBy* improved to 27% at its minimum (year 70) and 55% in the terminal year compared to models with no CKMR and *M* and *R0* improved to 55% and 60%, respectively. There did appear to be an effect specific to the flatfish models where the accuracy in *SSBy* for the first half of the time series actually worsened compared to models fit with no CKMR, before markedly improving in the second half of the time series.

*Precision within simulations*

The precision within simulations showed similar results to those from the accuracy performance metric. One notable exception was for the cod life history, where for all data collection treatments accuracy did not appear to improve, for treatments where *n*=5,000 the precision within simulations instead did improve for *SSBy* at the end of the time series (~80-90% at minimum, Figure 2) and for *M* and *R0* (~90-95%, Figure 4) relative to models with no CKMR.

For the sardine models, precision within simulations notably appeared to improve only in cases where *n*=5,000. For the case where *n*=5,000 for 20 years, integration of CKMR data into the assessment led to SEs that were 46% and 77% of those from models without CKMR for *M* and *R0*, respectively. For the same treatment, improvement in SEs of *SSBy* was 46% at its minimum (year 68) and 86% in the terminal year of the time series.

For the flatfish models, precision within simulations did improve for *n*=100 however this improvement was minimal (>95%, Figure 2). At *n*>100, precision within simulations improved markedly compared to models fit with no CKMR. For example, in a moderate case where *n*=1,000 collected for 10 years, the median SE of *SSBy* improved to 30% at its minimum (year 69) and 58% in the terminal year of the time series relative to models with no CKMR (Figure 2). In the same treatment for *M* and *R0*, the median SEs for each improved to 55% and 60% of the levels for models fit without CKMR data, respectively.

*Precision between simulations*

No marked improvements in precision of estimates between simulations were observed in the cod models. For the sardine models, precision between simulations for estimates of *M* and *R0* improved for all treatments where *n*>100. Metrics regarding *SSBy* however only appeared to markedly improve upon the models without CKMR data when either *n*=1,000 for 20 years or when n=5,000 for 5, 10, or 20 years. For example, in the case where *n*=1,000 for 20 years, the IQR of *SSBy* improved to 66% of the IQR for models without CKMR at the minimum (year 68) and 94% in the terminal year of the time series. In the same scenario for *M* and *R0*, IQR improved to 76% and 93% of the IQR for models with no CKMR, respectively.

For the flatfish models, precision between simulations improved for all scenarios where *n*>100. For example, in the case where *n*=1,000 for 20 years, IQR of *SSBy* improved to 24% of the IQR for models without CKMR at the minimum (year 69) and 48% in the terminal year of the time series. In the same scenario for *M* and *R0*, IQR improved to 52% and 66% of the IQR for models with no CKMR, respectively. Similar to the accuracy results, there did appear to be an effect specific to the flatfish models where the IQR for *SSBy* for the first half of the time series actually worsened compared to models fit with no CKMR, before markedly improving in the second half of the time series when *n*=5,000.

Discussion

The main factor in information content remains the spawning abundance of the stock coupled with the sample size taken (Look at percentage of stock sampled). Although zeroes did provide some information as evidenced by the improvements in precision within simulations for the cod models.

CKMR is a backwards looking inference (citation), backdated to the year of birth of the younger individual for POPs and the years of birth for each sibling in HSPs, hence the maximum improvements in each performance metric were observed some years before the terminal year of the time series.

This simple exercise could theoretically be undertaken by any given assessment to evaluate whether collecting and integrating CKMR data would be of value, although would be highly dependent on the scale of the operating model being reasonably close to the real-world stock.

Which is difficult if uncertain, especially since the CKMR data can help inform you on scale

The CKMR data improved the accuracy and precision of *M* estimates apparently more so than those of *R0*. This was perhaps a function of CKMR informing estimates in the terminal end of the time series much more so than at the beginning, and *R0* is primarily used to obtain initial scale of the population and recruitment in each year.

The integration of CKMR data may therefore be of great importance for helping to distinguish between time-varying selectivity and natural mortality.

Often in commercial fishery and research survey sampling programs fish are able to be sexed. In addition, many contemporary assessments are sex-specific (citation). The ability to sex biological samples would further refine CKMR analysis and change Equations 2 and 3 in perhaps obvious ways. Such refinement could ideally lead to further improvements in estimating mortality and abundance when integrating opportunistic CKMR data into stock assessments.

I have presented a simple case with no pair assignment error, no ageing error, no spatial dynamics ///, each of which should likely be expected in the real world and could negatively affect the expected improvements from opportunistic CKMR integration within fisheries assessments (Conn et al. 2020)

Data accessibility

All code associated with analysis is available on *https://github.com/nickcfisch/CKMR\_Project*.

Acknowledgements

I would like to acknowledge and give thanks to Eric Anderson and Paul Conn, whose workshop on close-kin mark-recapture at the 2022 annual meeting of the Wildlife Society introduced me to modeling associated with CKMR theory. Paul Conn also provided comments on an earlier draft. I would also like to acknowledge that the National Marine Fisheries Service funded my attendance at the mentioned workshop, and part of this analysis was conducted while under their employment.

References

Casas, L. and Saborido-Rey, F., 2023. A review of an emerging tool to estimate population parameters: the close-kin mark-recapture method.