



Data generation & changes to SBT OM

Rich Hillary, Ann Preece, Campbell Davies

12 June 2018

CSIRO Oceans & Atmosphere
Battery Point, Hobart 7004, Tasmania, Australia.

Copyright and disclaimer

© 2018 CSIRO To the extent permitted by law, all rights are reserved and no part of this publication covered by copyright may be reproduced or copied in any form or by any means except with the written permission of CSIRO.

Important disclaimer

CSIRO advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Contents

1 Background	1
2 Gene tagging	1
3 Close-kin mark-recapture	3
3.1 Parent-Offspring pairs	4
3.2 Half-Sibling pairs	4
3.3 Generating indices from CKMR data	5
4 Acknowledgements	6

Abstract

This paper details the structural changes made to the SBT Operating Model required to simulate the new data sources: gene tagging, and close-kin mark-recapture (parent-offspring and half-sibling pairs).

1 Background

Initial methods for generating these new data sources, as well as potentially informative indices that could be derived from them, were explored in [1]. Additionally, the structural changes required to the SBT OM to accommodate these new data sources in the conditioning phase were outlined in [2]. This paper details the actual technical details of the changes made to the SBT OM projection code (`sbtproj.tpl`), and how the new data generation control parameters are defined in the `mycontrol.dat` files.

2 Gene tagging

Details of how we simulate the gene tagging data in the OM ('sbtproj'). Major items covered are:

- The observation error model employed
- Including the "reality" of the gene tagging process
- Defining the estimates of abundance (and CVs thereof)
- Suggestions about what can be used in a candidate MP setting
- What changes are needed in the control and `sbtOMdata` files

The primary goal of the gene tagging (GT) program is to provide an estimate of the absolute abundance of age 2 fish (and some measure of the uncertainty thereof). Moving back to base principles, if you tag a single 2-year old fish in year y , release it "randomly" into the population and attempt to recapture it again (via a random sampling method) in year $y + 1$ then the chance of finding it is $1/N_{y,2}$. If you tag T fish this probability increases to $T/N_{y,2}$. If you genotyped S fish in year $y + 1$ to check for "recaptures" then you would expect to find the following average number of recaptures:

$$\mathbb{E}(R) = \frac{TS}{N_{y,2}}$$

The default appropriate distribution for these kind of data is the binomial distribution. As per previous discussions about how to model these data, and given the distributional features of the 1990s tagging data, we agreed to a more generalised distribution called the beta-binomial distribution. This is an extension of the binomial model that permits additional variability in the probability of recapture. The tagging data currently in the OM clearly show higher variance than the base multinomial distribution would predict. Some of those sources of additional variance (heterogeneity in assumed static parameters) will not be a feature of the GT program - specifically variation in tag loss and reporting rates. As such, the so-called *over-dispersion* coefficient (degree to which variance is inflated) of the tagging data (*ca.* 1.8) would be an upper bound to the GT over-dispersion coefficient, with a value of 1 (no over-dispersion) being an obvious lower bound. We also propose a simple bias factor to be included into the GT observation error

model that is there specifically to deal with recruitment dynamics that result in us systematically sampling a subset of the true age 2 abundance. We would propose the following modification to the expected recapture probability:

$$\tilde{p} = \frac{T}{q^{\text{gt}} N_{y,2}},$$

If we define the over-dispersion coefficient as ϕ . For a given sample size $n = S$, we define the crucial over-dispersion parameter, ω , as follows:

$$\omega = \frac{\phi - 1}{n - \phi},$$

then the parameters of the beta distribution, $p^{\text{gt}} \sim B(\alpha, \beta)$, that underlies the true sampling probability are defined as follows:

$$\alpha = \frac{(n - \phi)\tilde{p}}{(1 - \tilde{p})(\tilde{p} + (1 - \tilde{p})(\phi - 1))}$$

and

$$\beta = \frac{n - \phi}{\tilde{p} + (1 - \tilde{p})(\phi - 1)}.$$

In the practical simulation sense, given the relevant GT control parameters (T , S and ϕ), we first simulate p^{gt} from the underlying beta distribution, and then simulate the number of recaptures, R , from the binomial distribution parameterised by p^{gt} and $n = S$.

The estimate of abundance, given R , T and S is basically the classical Petersen estimator:

$$\hat{N}_{y,2} = \frac{TS}{R},$$

with the approximate CV of this estimate given by $1/\sqrt{R}$.

There are certain adaptive features of the GT program - particularly the post-release resampling program in the farms - that are worth both considering and actively including in the simulation process. The main point of the observation error model is to represent, to the best of our abilities, the actual process of data collection. In the GT program, if we had processed the S samples to find matches and found less than we would prefer (e.g. we have some minimum value, R_{min}), it is the case (as it was this year) that we do have the option of processing an additional number of samples, S_+ , to hopefully obtain additional matches and, as a result, a more accurate estimate of the age 2 abundance. To be clear, this in no way introduces bias to the estimation process: the proportional increase in sample size would be the same as the proportional increase in the expected number of matches. The expected abundance estimate would be the same, but the accuracy would be increased.

The settings for the GT sampling settings are contained in an augmented `mycontrol.dat` file used for the projections on previous occasions. An example file (`mycontrolGTMP.dat`) is now included in the git repository in the develop branch. The GT control variables are:

1. *Trel*: number of initial "releases", T , from which sample is taken
2. *Srec*: main number of samples, S , taken from harvested fish the following year

3. *Sadd*: "back-up" samples, S_+ that can be processed if needed
4. *Rmin*: minimum number of matches, R_{min} , if adaptive sampling required
5. *gtadsw*: adaptive sampling switch (0 off; 1 on) - if number of matches less than minimum amount, process the additional back-up samples to get more
6. *qgt*: value of the bias, q_{gt} , in the GT estimator (1, no bias; alternatives will be less than 1 probably...)
7. *phigt*: value of the over-dispersion, ϕ , for the GT simulations (1 means no OD basically)

The text in the augmented `mycontrol.dat` where the GT variables are set is found immediately below where the CPUE and aerial survey control variables are found:

```
# switches for robustness tests highCPUECV, highaerCV and updownq
# if highCPUECV=1 (set sigmaq=0.30 and sigmaaerial=0.30)
# if highAerialCV=1 (set sigmaq=0.20 and sigmaaerial=0.50)
# if updownq = 1 (increase q by 50% for first 5 years)
0 0 0
# control parameters for GT program
# Trel (releases)
# Srec (recapture samples scanned)
# Sadd (additional samples if needed)
# Rmin (minimum number of recaptures)
# gtadsw (switch for adaptive sampling given Rmin; 0 off,1 on)
# qgt (bias level for GAB N2)
# phigt (over-dispersion level)
5000 10000 5000 5 0 1.0 1.0
```

These settings are probably about right (in terms of sampling specifics) given the previous two GT field programs but can and probably will be adjusted as we get more into the work. In terms of files it is the `.s-` named files that contain the summaries of various quantities from the projections. For the future GT data a new file has been created (`.s11`) which contains the key variables of interest:

- Grid element relating to that sample
- Estimated abundance of age 2 fish from GT
- True value of age 2 fish in that projection sample
- The approximate CV of the GT estimate of of age 2 fish
- The observed number of matches given the GT program settings

3 Close-kin mark-recapture

Alongside the gene-tagging (GT) data, the OM will now also simulate close-kin mark-recapture (CKMR) data - both parent offspring (POP) and half-sibling (HSP) pairs. This document covers both the simulation of the data using the SBT OM, as well as some empirical and model-based options for using these data in an MP context.

Cohort	Adult year	Adult age	nK	nC
c	y	a	$\text{Bin}(p, nC)$	$\text{MN}(p_a, M_j)$

3.1 Parent-Offspring pairs

The key covariates of importance in the SBT OM in relation to POPs are:

- The year of sampling, y , of the adult
- The age at sampling, a , of the adult
- The birth year/cohort, c , of the juvenile

while the key parameters and derived variables of importance are:

- Abundance-at-age, $N_{y,a}$
- Relative reproductive output-at-age, φ_a

so that for an juvenile-adult pair $\{i, j\}$, so $z_i = \{c\}$ and $z_j = \{y, a\}$, then the probability of that pair being a POP is given by

$$\mathbb{P}(K_{ij} = POP | z_i, z_j) = \mathbb{I}(c < y < c + a) \frac{2\varphi_{a-(y-c)}}{\sum_i N_{c,i}\varphi_i} \quad (3.1)$$

where $\mathbb{I}()$ is the indicator function. The assumed distribution of the POPs is binomial, given the sampling probability in (1), and all juveniles are assumed to be sampled at 3 years old, and that each year there are M_i samples taken. For the adults, a random draw from the multinomial distribution of likely adults is taken (for a given adult sample size, M_j). Both the sample size control parameters are defined in the `mycontrol.dat` file as with the GT control parameters. The data are organised in the same way as the historical POP data are: a 5-d data frame:

where p_a is the distribution of adults in a given year and $\text{Bin}()$ and $\text{MN}()$ are the binomial and multinomial distributions, respectively.

A quirk of the CKMR data is that future simulated data collection (i.e. from projections) actually alters the data in the past - future adult samples are compared to pre-projection juvenile samples to look for POPs. To be clear, this does not invalidate the use of the actual historical CKMR data in the OM, it just outlines the complex nature of the temporal accumulation of comparisons and matches in the CKMR data.

3.2 Half-Sibling pairs

For the HSP data we compare to juvenile samples and ask: what is the probability that they share a mother or a father? For juvenile samples i and i' , the key covariate is their year of birth, or cohort c . The additional derived variables needed for constructing the HSP probabilities are natural mortality and fleet-summed harvest rates:

$$\mathbb{P}(K_{ii'} = HSP | z_i, z_{i'}) = \frac{4\pi^n q_{\text{hsp}}}{S_{c_{\text{max}}}} \left(\sum_a \gamma_{c_{\text{min}},a} \left(\prod_{k=0}^{\delta-1} \phi_{c_{\text{min}}+k,a+k} \right) \varphi_{a+\delta} \right) \quad (3.2)$$

where

$$\phi_{y,a} = \exp(-M_a) \prod_{s=1}^2 \left(1 - \sum_f h_{s,f,y,a} \right) \quad (3.3)$$

is the annual survival probability, and:

- $\{z_i, z_{i'}\} = \{c_i, c_{i'}\}$
- $c_{\min} = \min\{c_i, c_{i'}\}$
- $c_{\max} = \max\{c_i, c_{i'}\}$
- $\delta = c_{\max} - c_{\min}$
- q_{hsp} is a parameter to cover potentially length-driven or recruitment dynamic differences between POP and HSP absolute abundance information
- π^η is the false-negative retention probability (1 minus the false negative exclusion rate defined by the false-positive cut-off)
- $\gamma_{y,a}$ is the true age distribution of adults in year y

Both the sampling distributions of the POP and HSP data are assumed to be binomial. Indeed, extensive work has been done over the years to assess whether more complex distributions (such as the beta-binomial) were needed, and found that the predictive properties of the binomial likelihood assumed in the OM were satisfactory as they are (Hillary *et al.*, 2017).

There are only three additional entries required in the `mycontrol.dat` file to accommodate the variables needed to switch on and control the CKMR sampling part of the projection code. Just below the GT settings is the following text:

```
# control parameters for CK program
# Mjuv (juvenile samples)
# Madu (adult samples)
# cksw (0/1 off/on switch to simulate data)
1500 3000 1
```

The first parameter controls the number of juvenile age 3 samples (M_i), the second controls the total number of adult samples (M_j), and the third is a switch to turn the CKMR simulation on (value of 1) or off (value of 0). The simulated future POP data are stored in the `.s12` file (for each grid sample), and the future HSP data are stored in in the `.s13` file.

3.3 Generating indices from CKMR data

A detailed exploration of options were explored in [1]. The key information content in both the POP and HSP data is in the **ratio** of comparisons to matches. The general idea is this: if, over time, the average ratio of comparisons to matches is increasing/decreasing it is highly likely that the total adult abundance (TRO really) is decreasing/increasing. Another complication with using the CKMR data in the MP setting is that the projection and historical data need to be merged together.

Empirical indices can be derived from both the POP and HSP data, and can obviously be combined in a weighted average index that correlates well with the true trend in TRO. As an example, with the default sample size controls in the supplied `mycontrol.dat` file ($M_i = M_j = 3,000$) and a constant catch projection at the current TAC (17,647t) - where the true TRO increases

steadily to a median of around 34% by 2035 - we see that:

- The base POP index (no moving average just as defined in [1]) correlates with true TRO at around 60%
- The base HSP index (no moving average as with the POP index) correlates with true TRO around the 65% level
- A combined index of POP and HSP indices (weighted evenly) correlated at just above 60%
- By using say a simple 3 or 4 year moving average of the indices this correlation can be increase to the 70% and above range

The satisfactory level of correlation (better or at least as good as say an actual survey of the TRO with a 25% CV) suggests these indices could indeed be used empirically and in a target-type HCR. As there is a reasonably strong linear relationship between the indices and the true TRO it stands to reason that there is some target value of the index (relative to the current value) that would match reasonably well with the tuning target of future TRO depletion. These indices can be used in much the same form (trend, target or even limit) as were the CPUE and aerial survey indices in the Bali Procedure [3].

4 Acknowledgements

This work was funded by CSIRO and the Department of Agriculture and Water Resources.

References

- [1] R. M Hillary, A. Preece, and C. R. Davies (2016) Methods for data generation in projections. *CCSBT-OMMP/1609/07*
- [2] R. M Hillary, A. Preece, and C. R. Davies (2017) Updates required for new data sources and reconditioning of the CCSBT OM. *CCSBT-OMMP/1706/04*.
- [3] R M. Hillary *et al.* (2016) A scientific alternative to moratoria for rebuilding depleted international tuna stocks. *Fish and Fisheries*. **17**:469–482.

CONTACT US

t 1300 363 400

+61 3 9545 2176

e csiroenquiries@csiro.au

w www.csiro.au

WE DO THE EXTRAORDINARY EVERY DAY

We innovate for tomorrow and help improve today for our customers, all Australians and the world.

Our innovations contribute billions of dollars to the Australian economy every year. As the largest patent holder in the nation, our vast wealth of intellectual property has led to more than 150 spin-off companies.

With more than 5,000 experts and a burning desire to get things done, we are Australia's catalyst for innovation. WE IMAGINE. WE COLLABORATE. WE INNOVATE.