CCSBT-ESC/2108/14 (Rev.1)
(ESC Agenda item 10.1)

Oceans & Atmosphere
www.csiro.au

How the Cape Town Procedure deals with missing data

Rich Hillary, Ann Preece, Campbell Davies

CSIRO Oceans & Atmosphere
Battery Point, Hobart 7000, Tasmania, Australia.

Copyright and disclaimer

Important disclaimer

Contents

Abstract

Paper number 14 paper outlines the implications of a range of missing data scenarios for each of the four data inputs to the Cape Town Procedure. For the abundance series - longline CPUE and gene tagging - we recommend having a minimum of at least two data points in any relevant moving average. For the CPUE index, which uses a 4 year moving average, that suggests that missing any more than 2 data points in a given 4 year time window would make it difficult to run the MP. For the gene tagging data, which uses a 5 year weighted moving average, missing data was automatically dealt with in the weighting scheme (by design) and any more than 3 missing data points in a 5 year time window would make it difficult to run the Cape Town Procedure. For both POPs and HSPs, missing data can easily be handled by the CKMR model embedded within the Cape Town Procedure but, as the severity of the missing data increases, the model gradually ceases to update the more recent population dynamics and reverts to the prior values and the influence of the historical abundance and mortality information. In all cases of missing data the meta-rules processes will be used to assess any additional information, or indicators, and relative severity of events on the likely performance of the MP.

1 Background

The Cape Town Procedure (CTP), an updated Management Procedure (MP) designed to replace the Bali Procedure (BP) [1], was adopted in 2019 [2] and implemented for the first time in 2020 [3]. In response to member questions about how missing data either could be, or is already, dealt with in the CTP we have prepared this short paper to outline the general concepts. While no explicit missing data scenarios were explored in the MSE work behind the various updated MP candidates, there are some specific features in the structure of the CTP that relate to being able to accommodate a degree of missing data across the four input data sets: Japanese long-line CPUE; gene tagging estimates of 2 year-old abundance; CKMR Parent-Offspring pair (POP) data; CKMR Half-Sibling pair (HSP) data.

2 How the CTP deals with missing data

The core structure of the CTP is detailed in the Appendix, but we focus only on the features that require attention with respect to missing data scenarios for each of the data input sources.

2.1 Longline CPUE data

In the CTP the HCR requires a 4 year moving average of the most recent longline CPUE data. So it would take, in the extreme case, 4 years of missing CPUE data for the HCR to be rendered fully unusable. One could argue that at least two data points are required to construct an average however, so we might stipulate that a minimum of two CPUE data points are required to run the MP.

2.2 Gene tagging data

The gene tagging part of the CTP HCR uses a *weighted* 5 year average of the gene tagging estimates of absolute 2 year old abundance. The weighting itself is proportional to the number of matches - more matches means higher influence on the overall weighted mean. The reasoning

for doing this was twofold:

1. The precision of the estimates will naturally vary as opposed to our assumptions around the approximate consistency of the precision of the CPUE index. This type of weighting approximates the classical (log-scale) inverse variance weighting used across statistics
2. This weighting also takes care of infrequent missing data (such as will be the case for the 2020 estimate given the COVID-driven problems encountered). Any missing data is given zero weight (there are no matches) and, hence, is automatically left out of the TAC calculations

As with the argument around the CPUE as to whether having a single point versus *at least* two data points this would suggest that if we lost 3 data points in a given 5 year moving average window this would make it difficult to justify running the CTP as it is currently defined.

2.3 CKMR POP data

The POP data is actually two data sets: (i) the adult samples from the Indonesian spawning grounds; (ii) the juvenile samples collected from the surface fishery in the Great Australian Bight.

For a year of missing juvenile samples the outcome is straight foward to explain: we will be missing an entire cohort of juveniles and, hence, we will be lacking (permanently) any information on the overall size/age structure of the adult stock for that particular cohort/juvenile birth year. For a year of missing adult samples the picture is more complex - say we had our juvenile samples but no adults that still leaves a large pool of existing adults we can compare the juveniles with and gain information there. If we were missing both data sets for a number of years then obviously the information content on the most recent years will decrease accordingly. The prior estimates of key time-varying parameters in the CKMR model embedded within the CTP will gradually begin to dominate the data in the most recent years and, while the CTP will run and converge in a statistical sense, the actual information content will slowly decrease to low levels if the data doesn't return. Due to the nature of both the CKMR data and the model within the CTP there is no obvious number of missing data years - for juveniles or adults - that would render the CTP unusable.

2.4 CKMR HSP data

The juveniles are the only data set used in generating the HSP data. For a missing year the effect is more nuanced than for the simple POP scenario described in the previous section. As with the POPs we obviously lose that particular cohort's information on absolute adult abundance in the HSP sense. So a missing year of juvenile samples means that, across both the POPs and HSPs, we will be missing the adult abundance information for that particular juvenile birth year. In the HSP sense we will also be missing the compound information on adult abundance trend and mortality that we obtain from comparing the missing juvenile cohort to all the other existing ones. So a missing year of juvenile data has a more widespread effect on the HSP data, relative to the POPs, given we compare each juvenile cohort to all of the other ones. The same argument applies to the one given in the previous section about actually running the CTP for increasing missing HSP data: eventually it will revert to prior/historical information as we increasingly get less and less recent data.

3 Discussion

This paper outlines the implications of a range of missing data scenarios for each of the four data inputs to the CTP. For the abundance series - longline CPUE and gene tagging - we recommended that having a minimum of at least two data points in any relevant moving average window would be a base level requirement. For the CPUE index, which uses a 4 year moving average, that suggested that missing any more than 2 data points in a given 4 year time window would make it difficult to run the CTP. For the gene tagging data, which uses a 5 year weighted moving average, missing data was automatically dealt with in the weighting scheme (by design) and that anymore than 3 missing data points in a 5 year time window would make it difficult to run the CTP.

For the CKMR data sets it is a more nuanced problem. For the POP data a missing juvenile sampling year basically means we are missing the adult abundance and age structure information for the particular juvenile birth year. For a missing year of adult sampling it just reduces the overall information content over the recent past but does not automatically remove the information for a given year. Missing both data sets obviously means both effects come into play. For the HSPs, using only juvenile samples, a missing year removes the absolute information for the particular juvenile birth year (as with the POPs) but also removes the information we gain on abundance trend and mortality over time by comparing that cohort with all the other ones. For both POPs and HSPs, missing data can easily be handled by the CKMR model embedded within the CTP but, as the severity of the missing data increases, the model gradually ceases to update the more recent population dynamics and reverts to the prior values and the influence of the historical abundance and mortality information.

In all cases of missing data the meta-rules processes will be used to assess any additional information, or indicators, and relative severity of events on the likely performance of the MP.

4 Acknowledgements

This work was funded by the Australian Fisheries Management Authority and CSIRO.

References

[1] Hillary, R. M., Preece, A. L., Davies, C. R., Kurota, H., Sakai, O., Itoh, T., Parma, A. N., Butterworth, D. S., Ianelli, J., and Branch, T. A. (2016) A scientific alternative to mortaoria for rebuilding depleted international tuna stocks. *Fish and Fisheries* **17**: 469–482.

[2] Anonymous. (2019) Report of the 24th Meeting of the Extended Scientific Committee. *CCSBT*.

[3] Anonymous. (2020) Report of the 25th Meeting of the Extended Scientific Committee. *CCSBT*.

Appendix

Adult population model

The adult population model is defined as follows:

$$
\begin{aligned}
N_{y_{\min},a_{\min}} &= \bar{R}\exp\left(\xi_{y_{\min}} - \sigma_R^2/2\right), \\
N_{y,a_{\min}} &= \bar{R}\exp\left(\epsilon_y - \sigma_R^2/2\right), \\
\epsilon_y &= \rho\epsilon_{y-1} + \sqrt{1-\rho^2}\xi_y, \\
\xi_y &\sim N(0,\sigma_R^2), \\
N_{y+1,a+1} &= N_{y,a}\exp\left(-Z_{y,a}\right) \qquad a \in (a_{\min}, a_{\max}), \\
N_{y+1,a_{\max}} &= N_{y,a_{\max}-1}\exp\left(-Z_{y,a_{\max}-1}\right) + N_{y,a_{\max}}\exp\left(-Z_{y,a_{\max}}\right), \\
Z_{y,a} &= Z_y \qquad a \leq 25, \\
Z_{y,a} &= Z_y + \frac{a-25}{a_{\max}-25}\left(Z_{a_{\max}} - Z_y\right) \qquad a \in [26, a_{\max}], \\
Z_y &= \frac{Z_{\max}e^{\chi_y} + Z_{\min}}{1+e^{\chi_y}}, \\
\chi_{\text{init}} &\sim N(\mu_{\chi_{\text{init}}}, \sigma^2_{\chi_{\text{init}}}), \\
\chi_{y+1} &= \chi_y + \zeta_y, \\
\zeta_y &\sim N(0,\sigma_\chi^2), \\
TRO_y &= \sum_a N_{y,a}\varphi_a
\end{aligned}
$$

The fixed parameters and settings of this model are given by the following table:

Parameter	Value
$a_{\min}$	6
$a_{\max}$	30
σ_r	0.25
ρ	0.5
σ_χ	0.15
$Z_{\min}$	0.05
$Z_{\max}$	0.4
$Z_{a_{\max}}$	0.5
$\mu_{\chi_{\text{init}}}$	-1.38
$\sigma_{\chi_{\text{init}}}$	0.2
q_{hsp}	1

Table 4.1: *Fixed parameters and priors of the adult population model.*

The estimated parameters of this model are:

1. The mean adult recruitment, $\bar{R}$
2. The adult recruitment deviations, ϵ_y

3. The initial value, χ_{init}, that "starts" the random walk for Z_y (with an associated normal prior mean and SD)
4. The random walk deviations ζ_y

The likelihood for the POP data is similar to that used in the OM. The total reproductive output is calculated as follows:

$$TRO_y = \sum_{a=a_{\text{amin}}}^{a_{\max}} N_{y,a}\varphi_a$$

and consider a juvenile-adult pair $\{i, j\}$, where $z_i = \{c\}$ is the juvenile covariate and c is it's cohort (year of birth) and $z_j = \{y, a\}$ is the adult covariate and y and a are the year and age at sampling, respectively. The probability of that pair being a POP is given by

$$\mathbb{P}\left(K_{ij} = POP \,|\, z_i, z_j\right) = \mathbb{I}\left(c < y < c + a\right) \frac{2\varphi_{a-(y-c)}}{TRO_c}$$

This probability is used to create the binomial likelihood for the POP data. For the HSP data the comparison is of a juvenile-juvenile pair i and i', where the key covariates are their respective years of birth - or cohorts - c. The probability of finding an HSP is defined as follows:

$$\begin{aligned}
\mathbb{P}\left(K_{ii'} = HSP \,|\, z_i, z_{i'}\right) &= \frac{4\pi^{\eta} q_{\text{hsp}}}{TRO_{c_{\max}}} \left(\sum_a \gamma_{c_{\min},a} \left(\prod_{k=0}^{\delta-1} \exp\left(-Z_{c_{\min}+k,a+k}\right) \right) \varphi_{a+\delta} \right), \\
\gamma_{y,a} &= \frac{N_{y,a}\varphi_a}{TRO_y}, \\
\{z_i, z_{i'}\} &= \{c_i, c_{i'}\}, \\
c_{\min} &= \min\{c_i, c_{i'}\}, \\
c_{\max} &= \max\{c_i, c_{i'}\}
\end{aligned}$$

and this probability forms the basis of the binomial likelihood for the HSP data.

Harvest Control Rule

The general structure of the revised MP is as follows:

$$TAC_{y+1} = TAC_y \left(1 + \Delta_y^{\text{cpue}} + \Delta_y^{\text{ck}}\right) \times \Delta_y^{\text{gt}}, \tag{4.1}$$

Before detailing the functional form of the HCR we recap some useful variables:

- I_y^{ck}: moving average (of length τ^{ck}) of the estimated TRO from the MP population model (projected forward to the current year using the model to project forward for 4 years to avoid too much intertia in the signal when you need it)
- $\tilde{I}$: average estimated TRO from 2003 to 2014 (reference period w.r.t. relative rebuilding criterion)
- γ: proportional amount of TRO rebuilding we wish to achieve

- $\eta = I_y^{\rm ck}/(\gamma\tilde{I}) - 1$: the variable at which passing from negative to positive indicates the point at which the TRO rebuilding has been achieved and the transition in the reactivity of the MP occurs (i.e. it goes from reactive to passive w.r.t. CPUE and CKMR signals *only*)

For the CPUE part of the HCR we used a density-dependent gain parameter:

$$k^{\rm cpue}(\eta) = w_1^{\rm cpue}\left(1 - \left(1 + e^{-2\kappa\eta}\right)^{-1}\right) + w_2^{\rm cpue}\left(1 + e^{-2\kappa\eta}\right)^{-1}$$

This is using the logistic function approximation to the Heaviside step function $H[\eta]$ ($H[\eta < 0] = 0$, $H[\eta \geq 0] = 1$). We set $\kappa = 20$ so the transition between the two gain parameters, given η, happens within $\pm 5\%$ of $\delta = 1$. The CPUE multiplier is then just defined as follows:

$$\Delta_y^{\rm cpue} = k^{\rm cpue}(\eta)\left(\delta_y^{\rm cpue} - 1\right)$$

and $\delta_y^{\rm cpue}$ is actually very similar in form to the gene tagging part of the HCR

$$\begin{aligned}
\delta_y^{\rm cpue} &= \left(\frac{\bar{I}_{\rm cpue}}{I_{\rm low}}\right)^{\alpha_1} && \forall \bar{I}_{\rm cpue} \leq I_{\rm low}, \\
\delta_y^{\rm cpue} &= 1 && \forall \bar{I}_{\rm cpue} \in (I_{\rm low}, I_{\rm high}), \\
\delta_y^{\rm cpue} &= \left(\frac{\bar{I}_{\rm cpue}}{I_{\rm low}}\right)^{\beta_1} && \forall \bar{I}_{\rm cpue} \geq I_{\rm high},
\end{aligned}$$

where $\bar{I}_{\rm cpue}$ is the (4 year) moving average LL1 CPUE, $\bar{I}_{\rm low}$ and $\bar{I}_{\rm high}$ are upper and lower threshold CPUE values, and α_1 and β_1 allow for an asymmetric response above or below the threshold zone.

For the CKMR part of the HCR we try to ensure a minimum rate of increase in the TRO *beneath* the target level, and once it is achieved we would like to maintain the TRO at that level. To include this kind of behaviour in the HCR we also include some density-dependence in the log-linear growth rate at which the HCR moves from a TAC increase to a TAC decrease:

$$\begin{aligned}
\Delta_y^{\rm ck} &= k^{\rm ck}(\eta)\left(\lambda^{\rm ck} - \tilde{\lambda}(\eta)\right), \\
k^{\rm ck}(\eta) &= k_1^{\rm ck}\left(1 - \left(1 + e^{-2\kappa\eta}\right)^{-1}\right) + k_2^{\rm ck}\left(1 + e^{-2\kappa\eta}\right)^{-1}, \\
\tilde{\lambda}(\eta) &= \lambda_{\rm min}\left(1 - \left(1 + e^{-2\kappa\eta}\right)^{-1}\right)
\end{aligned}$$

The threshold level at which the log-linear trend, $\lambda^{\rm ck}$, goes from supporting a TAC decrease to an increase essentially begins at $\lambda_{\rm min} > 0$ and, as the estimated TRO approaches the target level, rapidly decreases to zero (in a similar way to the CPUE trend term). This is to ensure that a minimum level of rebuilding is encouraged for **all** trajectories below the target, and where above the target the *status quo* is preferred.

To calculate the recent mean age 2 abundance from the gene tagging data consider a weighted moving average approach:

$$\bar{N}_{y,2} = \sum_{i=y-1-\tau^{\mathrm{gt}}}^{y-2} \omega_i \widehat{N}_{i,2}$$

where ω_i is a weighting proportional to the number of matches used to produce the GT estimate $\widehat{N}_{i,2}$ (basically inverse variance weighting). The 2 year delay between having the estimate and what year it actually refers to is factored into the calculation. The multiplier for the GT part of the HCR is as follows:

$$\begin{aligned}
\Delta_y^{\mathrm{gt}} &= \left(\frac{\bar{N}_{y,2}}{N_{\mathrm{low}}}\right)^{\alpha} \quad \text{if} \quad \bar{N}_{y,2} \leq N_{\mathrm{low}}, \\
\Delta_y^{\mathrm{gt}} &= 1 \qquad\qquad\quad \text{if} \quad \bar{N}_{y,2} \in (N_{\mathrm{low}}, N_{\mathrm{high}}), \\
\Delta_y^{\mathrm{gt}} &= \left(\frac{\bar{N}_{y,2}}{N_{\mathrm{high}}}\right)^{\beta} \quad \text{if} \quad \bar{N}_{y,2} \geq N_{\mathrm{high}}
\end{aligned}$$

with N_{low} the limit level and N_{high} the upper level at where TAC increases are permitted. Table 2 details the parameter values for the HCR in the adqpted MP.

Parameter	Value
τ^{cpue}	4
w_1^{cpue}	0.9
w_2^{cpue}	0.005
I_{low}	0.45
I_{high}	1.42
α_1	1
β_1	1
τ^{gt}	5
N_{low}	1e+6
N_{high}	2.6e+6
α	1.5
β	0.25
τ^{ck}	3
k_1^{ck}	1.25
k_2^{ck}	0.05
γ	1.5
λ_{min}	0.001
κ	20

Table 4.2: *Fixed values of parameters of the HCR in the CTP.*

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 Australias catalyst for innovation.

WE IMAGINE. WE COLLABORATE. WE INNOVATE.