Pacific Ocean Perch State-Space Assessment Model Implementation in Template Model Builder

William H. Aeberhard
Department of Mathematics and Statistics, Dalhousie University
Halifax, Nova Scotia, Canada B3H 4R2

April 17, 2017

Contents

1 Introduction

This report documents the implementation of the Pacific Ocean Perch (POP, *Sebastes alutus*) state-space assessment model (SSAM) of Edwards et al. (2014) in the R package Template Model Builder (TMB; Kristensen et al. 2016). This model is formulated as a Bayesian state-space model, since prior distributions are specified for the model parameters, but is fitted by TMB following a frequentist paradigm. This means that the user will not draw from a posterior density to make inference, but rather will maximize a (Laplace-approximated marginal) log-likelihood function in order to obtain point estimates and standard errors; for general information about TMB, its usage, and contemporary applications, see Cadigan (2015), Kristensen et al. (2016), and Nielsen and Berg (2014).

The POP SSAM has three levels of hierarchy:

1. *prior distributions*: the deepest level specifies prior knowledge about model parameters $\boldsymbol{\theta}$ (in the form of density functions) which, to a non-negligible extent, help the identifiability (or estimability) of the model given the available data;

2. *process equations*: the second level describes the dynamics of unobserved states $\boldsymbol{X}_t$ (also called random effects) and derived quantities which are deterministic function of them, these equations involve process error;

3. *observation equations*: the top level defines how observed variables $\boldsymbol{Y}_t$, such as survey biomass estimates, are linked to the unobserved states, and include observation (or measurement) error.

A Bayesian perspective would consider three sources of error (randomness) in such a model: uncertainty in the parameters, process error and observation error. However, the frequentist approach taken by TMB views model parameters as *fixed parameters* (i.e. not random) and the specified prior distributions act more as loose constraints (or strict box constraints in the case of uniform priors) in the maximization procedure. Hence, only two sources of randomness are considered: the process and the observation error. Also, following a mixed effects terminology, the fixed parameters $\boldsymbol{\theta}$ will be said to be *estimated* while the random effects $\boldsymbol{X}_t$ will be said to be *predicted*.

The reweighting procedure described in Edwards et al. (2014, Appendix F.6.2) is not implemented here, mainly because of the arbitrariness such weights bear. TMB allows the user to estimate both observation and process variances, given identifiability conditions, which naturally balance the various data sources. Identifiability may not be met here given the scarcity of some data sources, so for now observation and process variances are fixed by the user following the original specification of the model.

The report is structured as follows: Section 2 sets the mathematical notation used throughout; Section 3 describes the POP SSAM in detail; Section 4 documents all modifications made to the original model of Edwards et al. (2014) to fit the TMB framework; finally, Section 5 presents the R commands made available with this implementation and describes their usage.

2 Notation

The notation used in this report follows closely that of Edwards et al. (2014), see in particular their Appendix F. Slight differences in the notation include: we consider the time series to end at year 2012 (thus the time series length is $T = 73$) and the commercial catch series is identified by $g = 1$ (rather than $g = 4$) to facilitate the incorporation of additional surveys.

2.1 List of abbreviations

CV	coefficient of variation
paa	proportion-at-age
POP	Pacific Ocean Perch
SD	standard deviation
SRR	stock-recruitment relationship
SSAM	state-space assessment model
TMB	Template Model Builder

2.2 Indices, subscripts and sets of years of available data

- a: age class, ranging from 1 to $A = 30$, where A is an accumulator class ("plus group")
- t: discrete time index representing years, ranging from 1 (year 1940) to $T = 73$ (year 2012)
- g: index identifying the four sources of data:
  - $g = 1$: commercial trawl data
  - $g = 2$: West Coast Vancouver Island synoptic survey series (survey 1)
  - $g = 3$: National Marine Fisheries Service Triennial survey series (survey 2)
  - $g = 4$: GB Reed historical survey series (survey 3)
- s: sex, $s = 1$ for females and $s = 2$ for males (note: s is also used to refer to selectivities, the subscript s exclusively refers to sex so there should be no ambiguity)
- $\mathbf{T}_g$: set of years where data source g has catch/biomass data. These are:
  - $\mathbf{T}_1$: full range of t index, i.e. 1940–2012
  - $\mathbf{T}_2 = \{2004, 2006, 2008, 2010, 2012\}$
  - $\mathbf{T}_3 = \{1980, 1983, 1989, 1992, 1995, 1998, 2001\}$
  - $\mathbf{T}_4 = \{1967, 1968, 1969, 1970\}$
- $\mathbf{U}_g$: set of years where data source g has proportion-at-age (paa) data. These are:
  - $\mathbf{U}_1 = \{1982, 1984, 1991, 1994, 1998, 1999, \ldots, 2006, 2008, 2011\}$

- $\mathbf{U}_2 = \{2004, 2006, 2008, 2010\}$

No paa data for $g = 3, 4$.

2.3 Fixed quantities: covariates and non-estimable parameters

The following are quantities and variables that are either observed or considered fixed. Either way, they are non-random, i.e. without uncertainty or measurement error.

- $n_{t,g}$: number of trips corresponding to the observed proportions-at-age $p_{a,t,g,s}$, $t \in \mathbf{U}_g$
- C_t: commercial catch series ($g = 1$), for $t = 1, 2, \ldots, T$, in tonnes
- $w_{a,s}$: average weight of individual of age-class a and sex s, in kg
- m_a: proportion of females of age a that are mature
- μ_g, Δ_g and υ_g for $g = 3, 4$: parameters specifying the selectivity curves $s_{a,g,s}$ for surveys 2 and 3, fixed respectively to 13.3, 0.22 and 10 for both surveys, these values being the initial values for estimating μ_g, δ_g and υ_g in survey 1. These parameters must be fixed for surveys 2 and 3 because no paa data are available for them.
- $\kappa_{t,g}$: standard deviation (SD) of the log-normal observation error (on the log scale) of the survey biomass index $I_{t,g}$, for $g = 2, 3, 4$ and $t \in \mathbf{T}_g$
- σ_R: SD of the log-normal process error (on the log scale) of the number of recruits R_t

The commercial catches C_t are not defined as a function of other variables and do not appear in catch equations, and hence enter the model as a fixed covariate.

$\kappa_{t,g}$ and σ_R are the observation and process SDs, respectively, and govern the overall magnitude of randomness in the SSAM. The series of observation SDs are computed from the coefficient of variation (CV), given on the survey indices' original scale, as follows:

$$\kappa_{t,g} = \sqrt{\log(\mathrm{CV}_{t,g}^2 + 1)}.$$

In Edwards et al. (2014), the process SD is fixed to 0.9 based on some previous work. We found this value to be quite large, in effect creating widely volatile recruits R_t and derived quantities in Monte Carlo simulations. This high variability is even more troublesome when simulating because the commerical catches C_t are fixed and thus the stock biomass B_t can easily drop below them just, rendering the simulated sample useless for model checking purposes. A value around 0.3–0.5 seems more reasonable and still allows for a predicted recruitment process to have sudden jumps (although not dramatically big). Fitting the model to the POP data may require even a smaller value of σ_R for the maximization to converge to a valid solution.

2.4 Response variables

The following are the components of $\boldsymbol{Y}_t$, the response variable measured with observation error (whose magnitude is tuned by $\kappa_{t,g}$).

- $I_{t,g}$: biomass estimates from surveys $g = 2, 3, 4$ with $t \in \mathbf{T}_g$, in tonnes
- $p_{a,t,g,s}$: weighted proportions-at-age (paa) of fish of age-class a and sex s, $t \in \mathbf{U}_g$, only available for commercial catch ($g = 1$) and survey 1 ($g = 2$)

The paa are proportions among fish from the same series (same g) and at the same time point (same t), i.e. $\sum_{a=1}^{A} \sum_{s=1}^{2} p_{a,t,g,s} = 1$ for $g = 1, 2$ and for all $t \in \mathbf{U}_g$.

2.5 Random effects and derived quantities

Strictly speaking, the POP SSAM has only one random effect $\boldsymbol{X}_t$: the number of recruits R_t (in thousands) at each time point t, for $t = 1, 2, \ldots, T$. Note that the initial number of recruits R_0 is estimated as a model parameter, see Section 2.6. All other unobserved features of the stock are deterministic functions (i.e. without additional randomness) of R_t, model parameters defined in Section 2.6 and other fixed quantities defined in Section 2.3. These derived quantities thus span the whole range of $t = 1, 2, \ldots, T$ and are:

- $s_{a,g,s}$: the fishing gear selectivity of series $g = 1, 2, 3, 4$, for fish of sex s and age a; a deterministic function of the selectivity parameters μ_g, Δ_g and υ_g only
- $N_{a,t,s}$: the abundance (in thousands) of fish of sex s and age a at time point t; a deterministic function of C_t, $w_{a,s}$, $s_{a,g,s}$, M_s, R_0, and R_t
- B_t: (spawning) biomass of mature females at the start of year t, in tonnes; a deterministic function of $w_{a,s}$, m_a and predicted $N_{a,t,s}$
- V_t: vulnerable biomass (both males and females) in the middle of year t; a deterministic function of $w_{a,s}$, M_s, the predicted $N_{a,t,s}$ and the catch selectivities $s_{a,1,s}$
- u_t: exploitation rate in the middle of the year t; ratio of C_t to V_t
- $u_{a,t,s}$: proportion of fish of age a and sex s at time t that are caught; product of u_t and catch selectivities $s_{a,1,s}$

To summarize, the randomness in R_t is propagated through the abundance $N_{a,t,s}$, while estimation variability (uncertainty) is introduced through the estimation of the selectivity parameters, and R_0 and M_s. The estimation of other model parameters, such as h entering the Beverton-Holt stock-recruitment relationship (SRR), only have an indirect impact on these derived quantities through the predicted recruits R_t and their possible interactions with other model parameters.

2.6 Model parameters

The model parameter vector $\boldsymbol{\theta}$ is of dimension 13 and consists of:

- R_0: initial number of recruits, in thousands
- M_s: natural mortality rate for $s = 1, 2$, assumed constant across age-classes and over time
- h: steepness parameter in the Beverton-Holt SRR; the parameters of the classical formulation are recovered through $\alpha = (1 - h)B_0/(4hR_0)$ and $\beta = (5h - 1)/(4hR_0)$, where B_0 is the initial spawning biomass set equal to the predicted B_1
- μ_g: age of full selectivity of females, for catches ($g = 1$) and survey 1 ($g = 2$); assumed known for surveys 2 and 3, see Section 2.3 ("full selectivity" is understood here as arbitrarily close to 1, since the logistic curve defined in Section 3.2 cannot be exactly 1)
- Δ_g: age shift defining full selectivity of males (shifted from the females' μ_g), for catches ($g = 1$) and survey 1 ($g = 2$); assumed known for surveys 2 and 3, see Section 2.3
- υ_g: steepness of the logistic selectivity curves (different interpretation than the υ in Edwards et al. (2014), see Section 3.2 and 4) for both females and males, for catches ($g = 1$) and survey 1 ($g = 2$); assumed known for surveys 2 and 3, see Section 2.3
- q_g: catchability coefficients for survey 1 ($g = 2$), survey 2 ($g = 3$) and survey 3 ($g = 4$)

To facilitate the maximization with respect to $\boldsymbol{\theta}$, model parameters that must remain positive are log-transformed within the optimization (not required by the end-user), this is the case of R_0, M_s, h, μ_g, υ_g and q_g.

3 Description of the model

All equations presented hereafter are numbered following Edwards et al. (2014, Appendix F) for ease of comparison and reference. That said, equations which were modified in any way from their original version, including the merging of many original equations into a new one, bear an additional *. Table 1 summarizes the correspondence between original and new equation numbers.

3.1 Prior distributions

Prior distributions represent knowledge about $\boldsymbol{\theta}$ before collecting data and typically formalize the range of most plausible/realistic values. Prior distributions themselves depend on parameters, called hyperparameters, which are specified beforehand as part of the prior knowledge. For parameters for which we do not have any particular prior knowledge, so-called uninformative priors can be formulated and usually do not favor any particular value

Table 1: Correspondence between original equation numbering of Edwards et al. (2014, Appendix F) and equation numbering in the present report, including respective page numbers

Edwards et al. (2014)		Present report	
Eq. number	Page	Eq. number	Page
(F.1)	80	(F.1)	10
(F.2)	80	(F.2)	10
(F.3)	80	(F.3)	10
(F.4)	80	(F.4)	9
(F.5)	80	(F.5)	9
(F.6)	80	(F.6)	9
(F.7)	80	(F.7*)	8
(F.8)	80	(F.8*)	8
(F.9)	81	(F.9)	10
(F.10)	81	(F.10*)	10
(F.11)	81	(F.11)	10
(F.12)	81	(F.12)	10
(F.13)	81	(F.13)	10
(F.14)	81	(F.14*)	10
(F.15)	81	(F.15)	10
(F.17)	82	(F.17*)	10
(F.19)	82	(F.19*)	11
(F.20)	82	(F.20*)	10

within a certain wide range (a uniform density which is constant over a given interval). In a strict Bayesian approach, $\boldsymbol{\theta}$ would be then seen as a random effect in the same way $\boldsymbol{X}_t$ is. This strict approach is not relevant in TMB: $\boldsymbol{\theta}$ remains a *fixed parameter* for which we compute point estimates and corresponding standard errors, following a frequentist approach. However, the priors still play a role as weights in the likelihood function we maximize, since they can favor certain values.

The priors and hyperparameters are taken directly from Edwards et al. (2014, Table F.4, p. 83) and are summarized in Table 2. We note the specification of uniform (uninformative) priors for the initial recruits R_0 and the log-catchabilities $\log q_g$, while some others (e.g. M_s) define quite narrow densities, in effect smoothly constraining the optimization towards certain values. Our experience is that such narrow (highly informative) priors are needed for certain parameters for which the data does not provide much information, this is known as an identifiability/estimability problem (see e.g. Lele et al. 2012). Hints of this phenomenon can already be seen by how close some prior and posterior densities are in Edwards et al. (2014, Figure G.25, p. 118). This seems to affect mostly M_1, M_2, $\Delta_{g=1}$ and $\Delta_{g=2}$.

The last column of Table 2 reports the default initial values used in the likelihood optimization. These are not strictly speaking part of the priors, but are nonetheless specified by the user beforehand and require to be realistic to some extent. They can be changed in the provided R commands, see Section 5. Also, all priors can be disabled by the user (i.e. not favoring any particular value for all parameters); this is done with the `TRUE/FALSE` argument `enable.priors`, again see Section 5.

Table 2: Prior distributions for the components of $\boldsymbol{\theta}$ and default initial values used in the likelihood optimization

Parameter	Prior distribution	Hyperparameters	Initial Value
R_0	uniform	lower bound = 1, upper bound = 10^5	5000
$M_s^\dagger$	normal	mean = 0.07, SD = 0.007	0.07
h	beta	shape $\alpha = 4.573$, shape $\beta = 2.212^\S$	0.674
$\mu_{g=1}$	normal	mean = 10.5, SD = 3.15	10.5
$\Delta_{g=1}$	normal	mean = 0, SD = 0.3	0
$\log \upsilon_{g=1}$	normal	mean = 1.52, SD = 0.456	1.609
$\mu_{g=2}$	normal	mean = 13.3, SD = 4	13.3
$\Delta_{g=2}$	normal	mean = 0.22, SD = 0.066	0.22
$\log \upsilon_{g=2}$	normal	mean = 2.3, SD = 1	2.303
$\log q_g^\ddagger$	uniform	lower bound = -12, upper bound = 5	0

†for $s = 1, 2$

‡for $g = 2, 3, 4$

§these shape parameters yield a mean of 0.674 and a SD of 0.168, as in Edwards et al. (2014)

3.2 Process equations and distributions

The selectivity curves $s_{a,g,s}$ only depend on their respective parameters and are thus independent of the random effect R_t and other quantities. Because of the way TMB treats `if` statements, the Gaussian density selectivity curves of equations (F.7) and (F.8) in Edwards et al. (2014, p. 80) had to be changed, see Section 4. We opted for a logistic function which maps the real line to the $[0, 1]$ interval:

$$s_{a,g,1} = \frac{\exp\big(10(a - (\mu_g - \upsilon_g/2))/\upsilon_g\big)}{1 + \exp\big(10(a - (\mu_g - \upsilon_g/2))/\upsilon_g\big)} \tag{F.7*}$$

$$s_{a,g,2} = \frac{\exp\big(10(a - (\mu_g + \Delta_g - \upsilon_g/2))/\upsilon_g\big)}{1 + \exp\big(10(a - (\mu_g + \Delta_g - \upsilon_g/2))/\upsilon_g\big)}, \tag{F.8*}$$

where the factor 10 is merely a steepness offset helping the logistic and the original curves look alike. μ_g and Δ_g retain their meaning and range of values; υ_g plays the same role of tuning the steepness of the selectivity curve, but its value cannot be compared between the original and new logistic specifications (different scales). The logistic curve is in fact more flexible and can get much flatter than the original one, see Figure 1 for some comparisons. This modification should not have much pratical implications except that υ_g will now take smaller values than before for achieving a similar-looking selectivity curve.

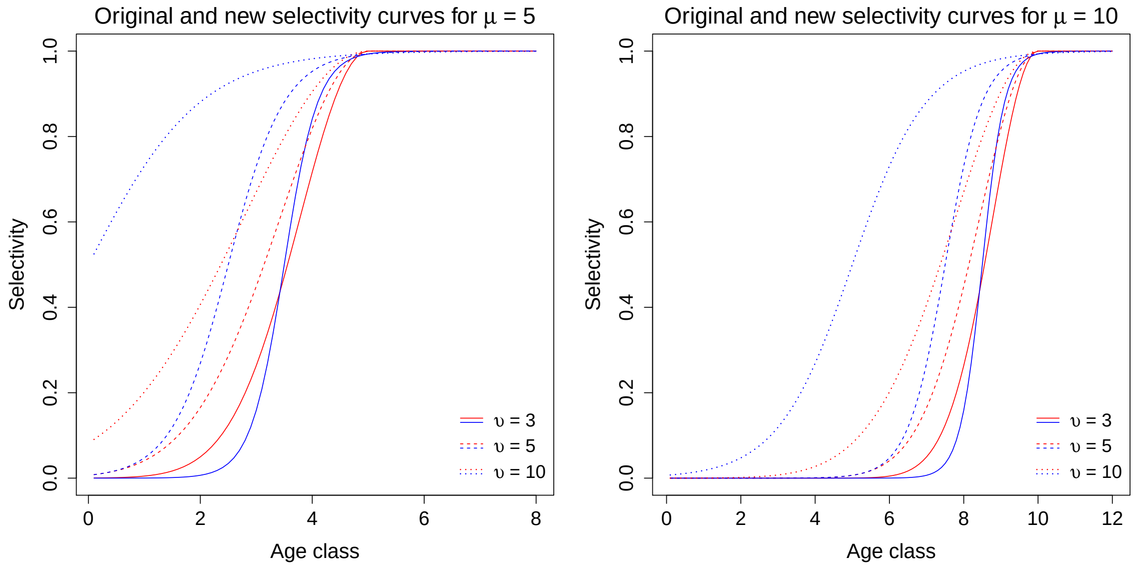

Figure 1: Original (in red) and new logistic (in blue) selectivity curves, for $\upsilon_g = 3$ (solid), 5 (dashed) and 10 (dotted), and $\mu_g = 5$ (left panel) and 10 (right panel).

The random effect R_t and the derived quantities defined in Section 2.5 depend upon another in intricate ways. The following initializations are necessary:

$$N_{a,1,s} = 0.5R_0 \exp[-M_s(a-1)], \quad 1 \le a \le (A-1),\ s = 1,2 \tag{F.4}$$

$$N_{A,1,s} = 0.5R_0 \frac{\exp[-M_s(A-1)]}{1-\exp(-M_s)}, \quad s = 1,2 \tag{F.5}$$

$$B_0 = B_1 = \sum_{a=1}^{A} w_{a,1} m_a N_{a,1,1}. \tag{F.6}$$

Next, the dynamics for $1 \le t \le T$ and $s = 1,2$ are specified through the following (recursive) equations:

$$\bar{R}_t = \frac{4hR_0B_{t-1}}{(1-h)B_0 + (5h-1)B_{t-1}} \tag{F.10*}$$

$$R_t = \bar{R}_t\epsilon_t \tag{F.17*}$$

$$N_{1,t,s} = 0.5R_t \tag{F.1}$$

$$N_{a,t,s} = \exp(-M_s)(1-u_{a-1,t-1,s})N_{a-1,t-1,s}, \quad 2 \le a \le (A-1) \tag{F.2}$$

$$N_{A,t,s} = \exp(-M_s)(1-u_{A-1,t-1,s})N_{A-1,t-1,s} + \exp(-M_s)(1-u_{A,t-1,s})N_{A,t-1,s} \tag{F.3}$$

$$B_t = \sum_{a=1}^{A} w_{a,1}m_aN_{a,t,1}. \tag{F.9}$$

$$V_t = \sum_{s=1}^{2}\sum_{a=1}^{A} \exp(-M_s/2)w_{a,s}s_{a,1,s}N_{a,t,s} \tag{F.11}$$

$$u_t = C_t/V_t \tag{F.12}$$

$$u_{a,t,s} = s_{a,1,s}u_t, \quad 1 \le a \le A, \tag{F.13}$$

where $\bar{R}_t$ refers to the deterministic prediction for R_t, and ϵ_t is an independent process error distributed as log-normal with a mean on the natural logarithm scale of $-\sigma_R^2/2$ and standard deviation of σ_R on the log scale. The mean of $-\sigma_R^2/2$ on the log scale implies a mean of one for the (multiplicative) error ϵ_t; throughout, all log-normal error terms are parameterized in this fashion so that their mean on the original scale is unity.

3.3 Observation equations and distributions

The three survey series ($g = 2, 3, 4$) are related to the stock's underlying features through the following observation equations, for $t \in \mathbf{T}_g$:

$$\hat{I}_{t,g} = q_g \sum_{s=1}^{2}\sum_{a=1}^{A} \exp(-M_s/2)(1-u_{a,t,s}/2)w_{a,s}s_{a,g,s}N_{a,t,s} \tag{F.14*}$$

$$I_{t,g} = \hat{I}_{t,g}\eta_{t,g}, \tag{F.20*}$$

where $\hat{I}_{t,g}$ is the fitted value for $I_{t,g}$, and $\eta_{t,g}$ is an independent log-normal observation error term with mean $-\kappa_{t,g}^2/2$ and SD $\kappa_{t,g}$ on the log scale. Note that equation (F.20*) only differs from the original (F.20) because of $\eta_{t,g}$ being parameterized here to have a mean of one (on the exponential scale).

The fitted values of the paa are computed as:

$$\hat{p}_{a,t,g,s} = \frac{\exp(-M_s/2)(1-u_{a,t,s}/2)s_{a,g,s}N_{a,t,s}}{\sum_{s=1}^{2}\sum_{a=1}^{A}\exp(-M_s/2)(1-u_{a,t,s}/2)s_{a,g,s}N_{a,t,s}}, \tag{F.15}$$

for $1 \le a \le A$, $t \in \mathbf{U}_g$, $g = 1, 2$ and $s = 1, 2$. They can enter the model in two different ways: the original specification of Edwards et al. (2014) assumes Gaussian observation noise directly added to $\hat{p}_{a,t,g,s}$ (set by the argument `lkhd.paa = "normal"`), while we propose an alternative binomial specification following Cadigan et al. (2014) which may be more statistically sound and arguably simpler (`lkhd.paa = "binomial"`). The Gaussian specification implies:

$$p_{a,t,g,s} = \hat{p}_{a,t,g,s} + \xi_{a,t,g,s}, \tag{F.19*}$$

where $\xi_{a,t,g,s}$ is a normally distributed observation error with mean zero and variance

$$\mathrm{Var}[\xi_{a,t,g,s}] = \frac{p_{a,t,g,s}(1 - p_{a,t,g,s}) + (10A)^{-1}}{n_{t,g}}.$$

The $(10A)^{-1}$ constant guarantees that the variance is always strictly positive, even when the observed $p_{a,t,g,s}$ is exactly 0 or 1; this constant can be disabled with the user-level option `var.paa.add` (`TRUE`/`FALSE`). In addition, in the `POPsim` command the observed $p_{a,t,g,s}$ are to be generated and thus cannot be used to compute the variance; for simulating data we therefore use $\hat{p}_{a,t,g,s}$ in $\mathrm{Var}[\xi_{a,t,g,s}]$.

The binomial specification assumes that the observed counts $n_{t,g}p_{a,t,g,s}$ are realizations of independent binomial random variables parameterized by $\hat{p}_{a,t,g,s}$ (seen as a probability) and the number of trips $n_{t,g}$. This implies a mean of $n_{t,g}\hat{p}_{a,t,g,s}$ and a variance of $n_{t,g}\hat{p}_{a,t,g,s}(1 - \hat{p}_{a,t,g,s})$. We note here that the Gaussian likelihood can be viewed as an approximation (by the Central Limit Theorem) to the binomial one, with the exception of using the observed $p_{a,t,g,s}$ in $\mathrm{Var}[\xi_{a,t,g,s}]$ instead of the fitted $\hat{p}_{a,t,g,s}$ (see Edwards et al. 2014, Section F.5.3, p. 87). The two specifications are expected to yield similar estimates whenever $p_{a,t,g,s}$ is far from both 0 and 1 and $n_{t,g}$ is large.

3.4 Joint and marginal likelihoods and TMB optimizations

The prior, process and observation distributions define the SSAM, which can be represented by the *joint* likelihood of $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T)$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)$:

$$L_{\text{joint}}(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{X}) = \pi(\boldsymbol{\theta}) \prod_{t=1}^{T} L_{\boldsymbol{\theta}}(\boldsymbol{X}_t | \boldsymbol{X}_{t-1}) L_{\boldsymbol{\theta}}(\boldsymbol{Y}_t | \boldsymbol{X}_t),$$

where $\pi(\boldsymbol{\theta})$ is the product of all prior densities defined in Section 3.1, $L_{\boldsymbol{\theta}}(\boldsymbol{X}_t | \boldsymbol{X}_{t-1})$ is the log-normal likelihood of ϵ_t defined in Section 3.2, and $L_{\boldsymbol{\theta}}(\boldsymbol{Y}_t | \boldsymbol{X}_t)$ is the product (over a, g and s) of the log-normal densities of $\eta_{t,g}$ and the normal/binomial densities of $p_{a,t,g,s}$ from the previous section. Since this joint likelihood involves unobserved variables (the unknown states $\boldsymbol{X}_t$), the function used for estimation and inference is the *marginal* likelihood

$$L_{\text{marginal}}(\boldsymbol{\theta}, \boldsymbol{Y}) = \int L_{\text{joint}}(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{x})\, d\boldsymbol{x}.$$

Due to its high dimensionality, and the non-linearity of the process and observation equations, this integral cannot be computed explicitly and needs to be approximated. TMB uses the Laplace approximation and returns (within the R environment) an approximated marginal log-likelihood along with its gradient (which is exact thanks to automatic differentiation). It is this Laplace-approximated marginal log-likelihood that is maximized with respect to $\boldsymbol{\theta}$ for obtaining parameter estimates. Standard errors of these estimates are computed by a generalized delta method based on the numerical Hessian matrix evaluated at the attained solution; see Kristensen et al. (2016) for more details. Once estimates for $\boldsymbol{\theta}$ are obtained, predicted random effects are computed by maximizing the joint log-likelihood $\log L_{\text{joint}}(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{x})$ with respect to $\boldsymbol{x}$. Finally, derived quantities are to be computed according to their formulae given estimates and predictions.

4 Modifications to the original model

The following modifications were made to the original model specification of Edwards et al. (2014).

- We consider the time series length T to be 73, with the full range of the t index identifying the years for which we have commercial catch data, i.e. 1940–2012. Our year $T+1$ is thus 2013 and implies for instance that the (predicted) biomass at the beginning of year 2013 is an out-of-sample prediction.

- The selectivity curves defined in equations (F.7) and (F.8) of Edwards et al. (2014, p. 80) cannot be easily implemented in TMB as they are. This is because these functions involve `if` statements which are by default ignored by the TMB compiler to guarantee the differentiability of the objective function (regardless of whether the function defined with `if` statements is actually differentiable or not). We defined alternative logistic selectivity curves in our equations (F.7*) and (F.8*) which mimic well the original ones for a certain range of parameters. The interpretation and values of μ_g and Δ_g remain unchanged, they still represent the age of full selectivity and the shift of this age from females to males, respectively. υ_g retains the same general meaning, it tunes the steepness of the selectivity curve, but it has a broader range of impact on the shape of the logistic curve. Hence its value cannot be directly compared between specifications. See Figure 1 for the comparison of the two selectivity curves: a small value of υ_g can make the logistic curve much flatter than the original one, and this is can be exagerated when μ_g is itself small. For $\upsilon_g < 5$ and $\mu_g > 5$ the two curves tend to be rather similar (although a fair comparison would not use the same value for υ_g with both curves).

- In (F.20*) the log-normal error term $\eta_{t,g}$ has a mean of $-\kappa_{t,g}^2/2$ on the log scale that guarantees a mean of one on the original scale of the survey series. This is done for ease of interpretation and for consistency with the process error of equation (F.17*).

- The process error $\xi_{a,t,g,s}$ in (F.19*) does not have the 1/100 term that appeared in the original equation (F.19) which was meant to "downweight large residuals". Such a goal was not achieved in this way since this constant term was added to all observations regardless of the difference $(p_{a,t,g,s} - \hat{p}_{a,t,g,s})$, i.e. regardless of whether an observation was badly fitted (outlying) or not.

In addition, some additional modifications were necessary in the R script that simulates data (`POPsim`):

- In (F.19*), the SD of the Gaussian likelihood cannot be based on the observed $p_{a,t,g,s}$, because these are the quantities that are precisely to be simulated. Instead, the SD is based on the mean $\hat{p}_{a,t,g,s}$. See Edwards et al. (2014, Section F.5.3, p. 87).

5 Usage of R commands

5.1 Format of input files

All input files should be loaded in the R environment using standard R commands such as `read.table` and `read.csv`. For ease of interaction with TMB, all input objects are vectors and matrices (i.e. two dimensions at most); this means that variables with more than two indices (e.g. the observed $p_{a,t,g,s}$) must be split in separate objects. We chose to split such variables by sex ($s = 1, 2$ distinguishes two separate objects in the R environment) and by data source ($g = 1, 2, 3, 4$ distinguishes separate objects too). This way, we retain the age-class (a) and time (t) dimensions within each object. Any variable/object indexed by t must have the years as first column, as this is necessary for matching indices based on the $\mathbf{T}_g$ and $\mathbf{U}_g$ sets of years of available data.

The inputs are the following:

- $I_{t,g}$: three matrices corresponding to the three surveys ($g = 2, 3, 4$). The rows represent years, so their number is the size of $\mathbf{T}_g$, e.g. five rows for the survey 1 because $\mathbf{T}_2 = \{2004, 2006, 2008, 2010, 2012\}$. Each matrix has three columns, the first one is the year label (the elements of $\mathbf{T}_g$), the second column is the survey biomass estimate, and the third column reports the observation error SD values $\kappa_{t,g}$.

- $p_{a,t,g,s}$: four matrices, resulting from separating the two sexes ($s = 1, 2$) and the two data sources ($g = 1, 2$). For each matrix, the rows represent years and their number is the size of $\mathbf{U}_g$, e.g. four rows for survey 1 for both sexes because $\mathbf{U}_2 = \{2004, 2006, 2008, 2010\}$. Each matrix has 31 columns, the first column is the year label (the elements of $\mathbf{U}_g$) and the next 30 columns represent the age classes ($A = 30$). Age classes for which no fish were observed at a given year are to be reported as a 0 paa (rather than a missing value).

- $n_{t,g}$: two matrices corresponding to the catches ($g = 1$) and survey 1 ($g = 2$). The rows represent years and their number is the size of $\mathbf{U}_g$. Each matrix has two columns, the first one is the year label (same first column as the corresponding matrices of paa), the second column reports the number of trips.

- C_t: a single $(T \times 2)$ matrix. Rows represent years, the first column is the year label (i.e. 1940–2012) and the second column is the corresponding catch.
- $w_{a,s}$: two vectors of dimension $A = 30$ each, split by sex $(s = 1, 2)$. Each value is the average weight at age a.
- m_a: a single vector of dimension $A = 30$. Each value is the proportion of mature females of age a.

In addition, optional inputs include the fixed parameters for the selectivities of surveys 2 and 3 for which no paa data is available (μ_g, Δ_g and υ_g, for $g = 3, 4$) and the fixed process error SD σ_R. These values are to be supplied as a named vector/list or a data frame, where the names must match `muS2`, `deltaS2`, `upsilonS2`, `muS3`, `deltaS3`, `upsilonS3` and `sigmaR`, respectively.

Finally, starting values for the model parameters, to be used in TMB's maximization of the Laplace-approximated marginal log-likelihood, can be supplied by the user as a named vector/list or a data frame, where the names must match `R0`, `M1`, `M2`, `muC`, `deltaC`, `upsilonC`, `muS1`, `deltaS1`, `upsilonS1`, `h`, `qS1`, `qS2` and `qS3`, respectively. The supplied values should be on the original scale, parameter transformation (e.g. log transformation for ensuring the positivity of a parameter within the optimization) is applied within the build and simulation procedures.

5.2 Compiling the TMB C++ template

The main file implementing the POP SSAM in TMB is the `POP.cpp` file. It is a TMB template (with a C++ syntax) defining the negative joint log-likelihood $-\log L_{\text{joint}}(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{X})$. The integration of the random effect $\boldsymbol{X}$ (by the Laplace approximation) and the computation of the gradient are done by TMB in the background. This TMB template needs to be compiled from within R by running:

```r
library(TMB)
compile("POP.cpp")
```

This will create different files in the current working directory depending on the operating system (dynamic-link libraries `.dll` on Microsoft Windows, shared object libraries `.so` on Unix systems). This compilation operation is only needed once, as the created libraries are saved on the disk and will be then loaded upon reopening R. Loading the TMB library however will be needed every time the R session is restarted (just like any other R package).

5.3 Building a `POPobj` object: the `POPbuild` command

The `POPbuild` command takes data and fixed parameters as inputs and returns a list of class `POPobj` to be then fed to the `POPfit` command.

The usage is the following:

```
POPbuild(survey1, survey2, survey3,
         paa.catch.female, paa.catch.male, n.trips.paa.catch,
         paa.survey1.female, paa.survey1.male, n.trips.paa.survey1,
         catch, paa.mature, weight.female, weight.male,
         misc.fixed.param = NULL, theta.ini = NULL,
         lkhd.paa = "normal", var.paa.add = TRUE, enable.priors = TRUE)
```

The arguments are:

- `survey1`: matrix (or data frame) of biomass estimates from survey 1 ($g = 2$); each row corresponds to a year of available data, the first column reports the year ($\mathbf{T}_2$), the second column is the survey biomass estimate $I_{t,2}$, and the third column is the observation error SD values $\kappa_{t,2}$, for $t \in \mathbf{T}_2$.

- `survey2`: matrix (or data frame) of biomass estimates from survey 2 ($g = 3$); each row corresponds to a year of available data, the first column reports the year ($\mathbf{T}_3$), the second column is the survey biomass estimate $I_{t,3}$, and the third column is the observation error SD values $\kappa_{t,3}$, for $t \in \mathbf{T}_3$.

- `survey3`: matrix (or data frame) of biomass estimates from survey 1 ($g = 4$); each row corresponds to a year of available data, the first column reports the year ($\mathbf{T}_4$), the second column is the survey biomass estimate $I_{t,4}$, and the third column is the observation error SD values $\kappa_{t,4}$, for $t \in \mathbf{T}_4$.

- `paa.catch.female`: matrix (or data frame) of paa for the females ($s = 1$) from the commercial catch ($g = 1$); each row corresponds to a year of available data, the first column reports the year ($\mathbf{U}_1$) and the next 30 columns correspond to the paa per age class and year $p_{a,t,1,1}$, for $t \in \mathbf{U}_1$.

- `paa.catch.male`: matrix (or data frame) of paa for the males ($s = 2$) from the commercial catch ($g = 1$); each row corresponds to a year of available data, the first column is the year ($\mathbf{U}_1$) and the next 30 columns correspond to the paa per age class and year $p_{a,t,1,2}$, for $t \in \mathbf{U}_1$.

- `n.trips.paa.catch`: matrix (or data frame) of the number of trips per year of available paa for the commercial catch ($g = 1$); each row corresponds to a year, the first column reports the year ($\mathbf{U}_1$), the second column is the number of trips $n_{t,1}$, for $t \in \mathbf{U}_1$.

- `paa.survey1.female`: matrix (or data frame) of paa for the females ($s = 1$) from survey 1 ($g = 2$); each row corresponds to a year of available data, the first column reports the year ($\mathbf{U}_2$) and the next 30 columns correspond to the paa per age class and year $p_{a,t,2,1}$, for $t \in \mathbf{U}_2$.

- `paa.survey1.male`: matrix (or data frame) of paa for the males ($s = 2$) from survey 1 ($g = 2$); each row corresponds to a year of available data, the first column is the year

($\mathbf{U}_2$) and the next 30 columns correspond to the paa per age class and year $p_{a,t,2,2}$, for $t \in \mathbf{U}_2$.

- `n.trips.paa.survey1`: matrix (or data frame) of the number of trips per year of available paa for survey 1 ($g = 2$); each row corresponds to a year, the first column reports the year ($\mathbf{U}_2$), the second column is the number of trips $n_{t,2}$, for $t \in \mathbf{U}_2$.
- `catch`: matrix (or data frame) of commercial catch ($g = 1$); each row corresponds to a year of available data, the first column reports the year ($\mathbf{T}_1$, i.e the whole range 1940–2012), the second column is the catch C_t.
- `paa.mature`: vector of $A = 30$ proportions of mature females.
- `weight.female`: vectors of $A = 30$ average weight at age of females ($s = 1$).
- `weight.male`: vectors of $A = 30$ average weight at age of males ($s = 2$).
- `misc.fixed.param`: optional named vector/list (or data frame) of fixed values for the selectivity parameters of surveys 2 and 3 and for the process error SD σ_R; the names must match `muS2`, `deltaS2`, `upsilonS2`, `muS3`, `deltaS3`, `upsilonS3` and `sigmaR`. If unused (left to `NULL`), then the following default values are used: `muS2 = muS3 = 13.3`, `deltaS2 = deltaS3 = 0.22`, `upsilonS2 = upsilonS3 = 10` and `sigmaR = 0.05`.
- `theta.ini`: optional named vector/list (or data frame) of starting values for the model parameters to be estimated; the names must match `R0`, `M1`, `M2`, `muC`, `deltaC`, `upsilonC`, `muS1`, `deltaS1`, `upsilonS1`, `h`, `qS1`, `qS2` and `qS3`. If unused (left to `NULL`), then the default values given in Table 2 are used.
- `lkhd.paa`: either `"normal"` or `"binomial"`; specifies a binomial likelihood for the paa, following Cadigan et al. (2014), which may be more statistically sound (and simpler) or a Gaussian likelihood as originally specified in Edwards et al. (2014). Defaults to `"normal"`.
- `var.paa.add`: either `TRUE` or `FALSE`; if `TRUE` (the default), it enables the addition of the $1/(10A)$ term in $\mathrm{Var}[\xi_{a,t,g,s}]$ related to the Gaussian observation error of (F.19*). If `FALSE`, this additional term is 0 (disabled). If some observed $p_{a,t,g,s}$ are exactly 0 or 1, as in the POP data, then it must be set to `TRUE` to avoid zero variances with `lkhd.paa = "normal"`.
- `enable.priors`: either `TRUE` or `FALSE`; if `TRUE` (the default), it enables all the prior distributions defined in Section 3.1. If `FALSE`, the priors are disabled and the maximization over $\boldsymbol{\theta}$ is carried without any constraints (apart from constraints on their range of possible values enforced by means of transformations, such as strictly positive variances).

The `POPbuild` command returns a list (of class `POPobj`) which is is to be fed to the `POPfit` command for fitting the POP model. The output list has the following elements:

- `parlist`: a list of named arguments corresponding to the starting values for model parameters to be estimated and the vector of random effects (the recruits R_t), all appropriately transformed for the optimizations.

- `datalist`: a list of named arguments corresponding to the data, fixed values of additional parameters and options.

- `length.theta`: integer representing the dimension of $\boldsymbol{\theta}$ (13).

- `years`: vector of years, 1940–2012.

- `A, TC, TS1, TS2, TS3, UC, US1`: integers representing the number of age classes, the length of the catch and surveys series and the length of the paa series for catch and survey 1, respectively.

5.4 Simulating data based on a user-supplied design: the `POPsim` command

The `POPsim` command takes the exact same inputs as `POPbuild` and returns a similar `POPobj` list as an output, with a few additional elements. The main difference is that, given the fixed quantities (see Section 2.3) and values for the model parameters (`theta.ini` or the defaults), this command simulates data according the POP model. It does so by generating process error ϵ_t and observation errors $\eta_{t,g}$ and $\xi_{a,t,g,s}$ and by using all equations in Section 3.2 and 3.3 to reconstruct series of survey estimates and paa. The output list can then be fed to the `POPfit` command so that sample estimates and predictions can be compared to the true values that generated the data (Monte Carlo simulation study).

The usage is the following, see details of `POPbuild` for the description of each argument:

```
POPsim(survey1, survey2, survey3,
       paa.catch.female, paa.catch.male, n.trips.paa.catch,
       paa.survey1.female, paa.survey1.male, n.trips.paa.survey1,
       catch, paa.mature, weight.female, weight.male,
       misc.fixed.param = NULL, theta.ini = NULL,
       lkhd.paa = "normal", var.paa.add = TRUE, enable.priors = TRUE)
```

The `POPsim` command returns a list (of class `POPobj`) with the same elements as the output of `POPbuild` but with the addition of:

- `true.R`: vector of simulated recruits R_t.

- `true.B`, `true.V`, `true.u`: vectors of derived quantities B_t, V_t and u_t, respectively.

- `true.uaa.female`, `true.uaa.male`: matrices corresponding to the derived $u_{a,t,1}$ and $u_{a,t,2}$.

- `true.select.female`, `true.select.male`: matrices corresponding to the derived $s_{a,g,1}$ and $s_{a,g,2}$.

- `true.abund.female`, `true.abund.male`: matrices corresponding to the derived $N_{a,t,1}$ and $N_{a,t,2}$.

5.5 Fitting the model to data: the `POPfit` command

The `POPfit` command takes a `POPobj` list as input (coming from either `POPbuild` or `POPsim`), fits the POP model to the data (either supplied or simulated) and returns parameter estimates, random effect predictions, derived quantities and corresponding standard errors.

The usage is the following:

```
POPfit(POPobj, trace = TRUE,
       optim.control = list(eval.max = 5000, iter.max = 5000))
```

The arguments are:

- `POPobj`: a list of class `POPobj` coming from either `POPbuild` or `POPsim`.
- `trace`: either `TRUE` or `FALSE`; if `TRUE` (the default), some information relevant to the optimizations is output in the console, including timings. If `FALSE`, nothing is output (useful in simulations to keep the console clean).
- `optim.control`: a list of named arguments directly passed to `nlminb`'s control argument, see `help(nlminb)`.

The `POPfit` outputs a list with the following elements:

- `theta`, `se.theta`: vectors of the estimates of $\boldsymbol{\theta}$ and corresponding standard errors.
- `R`, `se.R`: vectors of predicted recruits R_t (random effect) and corresponding standard errors.
- `B`, `se.B`: vectors of derived spawning biomass B_t and corresponding standard errors.
- `V`, `se.V`: vectors of derived vulnerable biomass V_t and corresponding standard errors.
- `u`, `se.u`: vectors of derived exploitation rates u_t and corresponding standard errors.
- `uaa.female`, `se.uaa.female`: matrices of derived proportions of females caught $u_{a,t,1}$ and corresponding standard errors.
- `uaa.male`, `se.uaa.male`: matrices of derived proportions of males caught $u_{a,t,2}$ and corresponding standard errors.
- `select.female`, `se.select.female`: matrices of derived selectivities for females $s_{a,g,1}$ and corresponding standard errors.
- `select.male`, `se.select.male`: matrices of derived selectivities for males $s_{a,g,2}$ and corresponding standard errors.

- `abund.female`, `se.abund.female`: matrices of derived abundances of females $N_{a,t,1}$ and corresponding standard errors.
- `abund.male`, `se.abund.male`: matrices of derived abundances of males $N_{a,t,2}$ and corresponding standard errors.
- `mean.survey1`: vector of fitted values for survey 1 $\hat{I}_{t,1}$.
- `mean.survey2`: vector of fitted values for survey 2 $\hat{I}_{t,2}$.
- `mean.survey3`: vector of fitted values for survey 3 $\hat{I}_{t,3}$.
- `mean.paa.catch.female`: matrix of fitted values for catch paa for females $\hat{p}_{a,t,1,1}$.
- `mean.paa.catch.male`: matrix of fitted values for catch paa for males $\hat{p}_{a,t,1,2}$.
- `mean.paa.survey1.female`: matrix of fitted values for survey 1 paa for females $\hat{p}_{a,t,2,1}$.
- `mean.paa.survey1.male`: matrix of fitted values for survey 1 paa for males $\hat{p}_{a,t,2,2}$.

References

Cadigan, N. G. (2015). A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. *Canadian Journal of Fisheries and Aquatic Sciences* 73 (2), 296–308.

Cadigan, N. G., M. J. Morgan, and J. Brattey (2014). Improved estimation and forecasts of stock maturities using generalised linear mixed models with auto-correlated random effects. *Fisheries Management and Ecology* 21 (5), 343–356.

Edwards, A. M., R. Haigh, and P. J. Starr (2014). *Pacific Ocean Perch (Sebastes alutus) stock assessment for the west coast of Vancouver Island, British Columbia.* Canadian Science Advisory Secretariat Research Document 2013/093. Fisheries and Oceans Canada.

Kristensen, K., A. Nielsen, C. W. Berg, H. Skaug, and B. M. Bell (2016). TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software* 70 (5), 1–21.

Lele, S. R., K. Nadeem, and B. Schmuland (2012). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association* 105 (492), 1617–1625.

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