

**SPECIFICATIONS FOR OPERATING MODELS TO EVALUATE BIAS
IN ESTIMATION METHODS IN ACCORDANCE WITH
RECOMMENDATION A.1 OF THE 2014 INTERNATIONAL REVIEW
PANEL**

Penguin Island Closure Task Team¹

Sub-regional biomass surrogate approach

$$\ln(F_{y,i,s}) = K + \alpha_y + \gamma_s + \lambda_i \frac{C_{y,i,p}}{\bar{C}_{i,p}} + \delta_i X_{y,i} + \varepsilon_{y,i,s} \quad (1)$$

for year y , island i , and data series s , where:

K is the intercept,

$F_{y,i,s}$ is the penguin response variable; note that while in most instances F has been taken to be identical to r (the value reported in the tabulations in PENG/DATA1), for the active nest proportion response F was set equal to $r/(1-r)$ to maintain a potential unconstrained positive range for F , and for the foraging track related parameters $F = -r$ so that the sign of the λ fishing effect parameter maintains its same meaning throughout (positive/negative being favourable/unfavourable),

α_y is a year effect reflecting the pelagic fish biomass present in the sub-region (e.g. Dassen and Robben joint vicinity) in year y , with the intercept K chosen so that the α_y have a mean of zero,

γ_s is a series effect (subsuming an island effect),

λ_i is a fishing effect,

$C_{y,i,p}$ is the catch taken in year y in the neighbourhood of island i of pelagic species p ,

$\bar{C}_{i,p}$ is the average catch taken over the years considered, and (excluding years for which fishing was prohibited),

δ_i is the magnitude of the closure effect where $X_{y,i}$ is 0 if island i is closed in year y , and 1 otherwise, and

$\varepsilon_{y,i,s}$ is an error term.

Regarding the neighbourhood of an island, four options will be considered for catches:

- a) block-defined areas reflecting 10, 20 and 30 nm (C10, C20 and C30); and
- b) a circle of radius 18 km around the island (C_{closure}) – note that 18 km has been used in place of the intended 20 km because inaccuracies in catch positions otherwise result in some unduly large catches inside this region when closed to fishing..

For the last option, these data are available directly for the last few years only, as it is only these for which accurate details of catch positions have been kept. To generate C_{closure} values for earlier years,

¹ The Task Team consisted of M.O. Bergh, D.S. Butterworth, K.L. Cochrane (chair), T.L. Morris, R.B. Sherley and H. Winker. A. Ross-Gillespie undertook, on behalf of the Team, all the analyses and tests, under the supervision of D.S. Butterworth.

for each 10 nm square block used to calculate C20, the average over the recent years for which accurate catch positions are available of the proportion of the catch within a 18 km range of the island concerned will be calculated. These average values will then be applied to the C20 information for each earlier year to calculate C_{closure} for that year. (Note that this procedure does not take the uncertainty associated with this extrapolation process into account, but that is considered a factor likely to have minimal impact on results, whose incorporation would anyway be computationally problematic in leading to a very large number of operating models.)

Conditioning

The simulation testing approach put forward here follows that developed at the 2012 ICES Working Group on Methods of Fish Stock Assessment. This was adopted for the SISAM assessment methods testing exercise for which results were reported at the 2013 Boston World Conference on Stock Assessment Methods, and involves conditioning the parameters of the operating model on the data available to better provide results pertinent to the situation being analysed. Specifically here, three forms of the operating model of equation (1) are suggested to be conditioned on the data available for the scenario (response variable/species/catch series) under consideration:

- i) Catch only: $\delta_i = 0$
- ii) Closure only: $\lambda_i = 0$
- iii) Both catch and closure effects.

While parameter estimation has previously been achieved successfully for i) and ii), it is possible that problems may occur for iii) because of difficulties in distinguishing between the effects of catch and of closure given relatively few data, i.e. likelihoods may prove near flat in the associated parameters. This could, for example, result in unrealistic high variance values estimated for λ and δ , with perhaps an unrealistically large positive value for one being compensated by an unrealistically large negative value for the other. If such instances arise, values of δ_i will be fixed at half of their values for ii). Also should the conditioning lead to a result for iii) very close to that for either i) or ii) (hence providing little further information), this previous option with the intermediate values of δ_i will also be added. It is anticipated that fixing the values of δ_i for iii) at half of their values for ii) will result in estimates of λ_i which are about half their values for i); should this prove not to be the case for some scenarios, an alternative based on fixing the λ_i at half their values in i) and then estimating the δ_i will also be considered. In this conditioning, the year effect α_y is treated as a random effect, with REML used for estimation of the parameter values for the equation (1) operating model.

The residuals $\varepsilon_{y,i,s}$ are taken to be normally distributed: $N(0, \sigma_\varepsilon^2(y))$, where the variance $\sigma_\varepsilon^2(y)$ is taken to have the form:

$$\sigma_\varepsilon^2(y) = \sigma_0^2 + \sigma_1^2 / N(y) \quad (2)$$

where $N(y)$ is the sample size for the year concerned, and σ_1^2 is determined by a linear regression forced the origin of the squared standard error of the $\ln(F_{y,i,s})$ observation against $1/N(y)$.

Generating pseudo-data

These are generated as follows:

$$\ln(F_{y,i,s}) = \hat{K} + \hat{\alpha}_y + \hat{\gamma}_s + \hat{\lambda}_i \frac{C_{y,i,p}}{C_{i,p}} + \hat{\delta}_i X_{y,i} + \varepsilon_{y,i,s} \quad (3)$$

where

\hat{K}_y is the best estimate of K ,

$\hat{\alpha}_y$ are generated from a normal distribution with a mean of zero and a variance equal to that of the year-effect series in question (i.e. the variance estimated for this random effect) which is provided by the conditioning process, but truncating any values generated outside the range covered by plus/minus twice the standard deviation. This truncation is achieved by setting any value generated outside these boundaries to the value at the boundary.

$\hat{\gamma}_s$ are the best estimates of γ_s ,

$\hat{\lambda}_i$ are the best estimates of λ_i ,

$\hat{\delta}_i$ are the best estimates of δ_i ,

$X_{y,i}$ are maintained at their historical values for the scenario under consideration,

$$\begin{aligned} C_y \text{ values are generated from } \quad C_{y,i,p} &= \bar{C} + m(B_y - \bar{B}) + \eta_i & \eta_i &\sim N(0, \sigma_\eta^2) \\ &= \bar{C} + \frac{m}{\kappa} \hat{\alpha}_y + \eta_i & & \end{aligned} \quad (4)$$

where \bar{C} is the average of the historical C_y series in question,

m is the slope of the linear regression of catch in the vicinity of the islands against regional biomass (see more details below),

$$\hat{\alpha}_y = \kappa(B_y - \bar{B}),$$

$\kappa = \sigma_\alpha / \sigma_B$, i.e. $\hat{\alpha}_y$ is scaled to biomass B_y by use of the ratio of the standard deviations of the two distributions.

$$\text{Further } \sigma_B^2 = \sigma_{Bobs}^2 - \left(\sum_y \sigma_{B_y} / \sum_y 1 \right)^2$$

where σ_{Bobs} is the standard deviation of the B_y series and $\sigma_{B_y} = CV_{B_y} * B_y$ is the standard error of B_y (from Table 12 of MARAM/IWS/DEC15/PengD/BG1), and

σ_{Bobs} estimates the standard deviation of the true B_y values by subtracting the effect of survey sampling error for the values available.

Now from Equation (4):

$$\sigma_C^2 = \left(\frac{m}{\kappa} \right)^2 \sigma_\alpha^2 + \sigma_\eta^2 \quad (5)$$

where σ_C^2 is the standard deviation of the distribution of the catch series in question.

Further if ϕ is the correlation between catch and biomass, then:

$$\begin{aligned} \phi^2 &= 1 - \text{var}[(C_y - \bar{C}) - m(B_y - \bar{B})] / \sigma_C^2 \\ &= \left(\frac{m}{\kappa} \right)^2 \frac{\sigma_\alpha^2}{\sigma_C^2} \\ &= m^2 \frac{\sigma_B^2}{\sigma_C^2} \end{aligned} \quad (6)$$

From this it follows that:

$$\sigma_{\eta}^2 = (1 - \phi^2) \sigma_C^2 \text{ or } \sigma_{\eta} = \sqrt{1 - \phi^2} \sigma_C \quad (7)$$

In the generation process σ_B and σ_C are fixed by the data, and ϕ is fixed by input selection. This means that m is fixed by equation (6), and may differ from the m value obtained in the original regression and listed in Table 1. This difference is inevitable, as the data used for the corresponding regression have their own correlation which may differ from the value input, necessitating a change in the value of m used for the data generation.

Values of $C_{y,i,p}$ generated that are either more than two standard deviations above \bar{C} or less than 5% of \bar{C} will be truncated to avoid unduly influential values or values mimicking closure. This truncation is achieved by setting any value generated outside these boundaries to the value at the boundary.

Since reliable estimation of the linear regression parameters is possible only for the full time series available, which corresponds closely to the periods for which the longer penguin response values are available, an adjustment to the above is needed for the shorter period response series which correspond roughly to the 2004-2013 period, as the ‘‘centre’’ of the α_y distribution will no longer correspond to the mean value of B_y for the full time period. Here then $\alpha_y = \kappa(B_y - \bar{B}^*)$, where \bar{B}^* is average value of B_y over 2004-2013, and the computation of κ also takes only B_y into account.

For years where there is a closed area around the island, C_{10} and C_{closure} are taken to be 0. There are two options for C_{20} and for C_{30} :

- i) They are generated as above, i.e. the assumption is made that the catch that would have been taken in the closed area is taken immediately outside that in the C_{20} blocks intersected by the 18 km radius circle around the island.
- ii) They are set equal to $C_{20} - C_{\text{closure}}$ and $C_{30} - C_{\text{closure}}$ respectively, with all values sampled with replacement from the same year of the historical series. This assumes that any catch that would have been made within 18 km of the island is taken instead to far away from the island to have any impact, so that i) and ii) bound the plausible range.

The values of ϕ to be used are 0.2 and 0.4 for anchovy, and 0.4 and 0.6 for sardine; these are intended to span a plausible range, based on regressions of C_{10} and C_{20} against regional biomass from surveys for which results are shown in Table 1 and Fig 1. To allow for possible greater correlation at the island level, cases where each of these values is increased by 0.1 will also be run. The values of \bar{B} and \bar{C} will be as given for the scenario concerned in Table 1 (once values for C_{closure} become available, the corresponding regression results will be added to Table 1).

The length of the time series of data simulated will correspond to that available for the scenario for which series are under consideration.

The residuals $\varepsilon_{y,i,s}$ for equation (3) are generated from:

$$\varepsilon_{y,i,s} = \mu \varepsilon_{y-1,i,s} + \sqrt{1 - \mu^2} \eta_{y,i,s} \quad (8)$$

where the first value of ε and η are from $N(0, \sigma_{\varepsilon}^2(y))$ with $\sigma_{\varepsilon}^2 = \sigma_0^2 + \sigma_1^2/N(y)$ and $N(y)$ is sampled with replacement from the historical series in question while σ_0^2 and σ_1^2 are set to their values respectively estimated and input for the conditioning, and

μ is the autocorrelation, for which values of 0, 0.2 and 0.5 are considered.

Note that under this approach, any interaction term between the α_y and γ_s is subsumed in the variance σ_0^2 under the assumption that those interaction terms are normally distributed.

Regional biomass approach

$$\ln(F_{y,i,s}) = \psi B_y + \gamma_s + \lambda_i \frac{C_{y,i,p}}{C_{i,p}} + \delta_i X_{y,i} + \varepsilon_{y,i,s} \quad (9)$$

where B_y is the biomass within the pertinent region (e.g. SA coast west of Cape Agulhas for Dassen and Robben islands). B_y is taken to be either the estimate from the spawner biomass survey of the preceding November (which measures fish on which the penguins would feed before commencing breeding, and hence may relate to their pre-breeding condition), or the May recruitment survey for that same year (which relates to fish present during the penguin breeding season).

In what follows, only changes from the procedures detailed above for the biomass surrogate approach are listed.

Conditioning

Instead of REML, MLE is used for conditioning, which the resultant estimate of (the process error part of) the variance of the residuals adjusted upwards by the bias correction factor $n/(n-p)$ where n is the number of values available for the response variable and p is the number of estimable parameters, as is the standard result for linear fixed effect models.

Generating pseudo-data

$\hat{\psi}$ is the best estimate of ψ from the conditioning.

For the pseudo B_y values, a measurement error term is added to the value drawn from the distribution used to reflect the range of the pertinent historical values (see Fig. 2), drawn from a normal distribution with standard deviation $\sigma_B(y)$ equal to that arising from survey sampling for the historical observation. The distributions in Fig. 1 were selected in the main by fitting the levels and change-points for three constant values, but reducing this to two in cases where there was clearly no justification in including the third given the low number of data. The two periods chosen are broadly reflective of the years covered by the longer and shorter series of response variable values in PENG/DATA1.

Since in reality this measurement error is a contributor to the process error component of the residual variance σ_0^2 , when these are generated according to equation (5), the value estimated for σ_0^2 will be adjusted:

$$\sigma_0^2 \rightarrow \sigma_0^2 - [\hat{\psi} \cdot \sigma_B(y)]^2 \quad (10)$$

Note that this adjusted σ_0^2 also incorporates the variance around the relationship between the regional biomass and that in the vicinity of the island concerned.

$$C_y \text{ values are generated from } C_{y,i,p} = \bar{C} + m(B_y - \bar{B}) + \eta_i \quad \eta_i \sim N(0, \sigma_\eta^2) \quad (11)$$

where $\sigma_\eta = \sqrt{1 - \phi^2} \sigma_C$ as in equation (7) and the B_y are generated as detailed above.

Other considerations

- The following scenarios were selected for a first set of conditioning attempts (subject to the sampling standard errors associated with the data concerned becoming available very shortly):
 - For the biomass surrogate approach – equation (1)
 - a) Chick growth – anchovy
 - b) Chick growth – sardine
 - c) Forage trip duration – anchovy
 - For the regional biomass approach – equation (9)
 - a) Forage trip duration – sardine
- Any prospective estimator must have bias estimated across the full suite of operating models.
- The following list of output statistics were agreed upon by the task team:
 - a) histogram of observed and generated biomass values across all years
 - b) histogram of observed and generated catch values across all years
 - c) histogram of catch to biomass correlation values
 - d) mean and variance of catch to biomass correlations
 - e) histogram of observed and generated response variable across all years
 - f) mean and variance of observed and generated response variable across all years
- For any of the operating models specified above, a range of estimators can be tested. This would be intended to include estimators corresponding to the form of the model itself, of other operating models in the set, of simpler forms of these operating models (e.g. *in extremis* no more than a closure/non-closure differentiation as amongst the suggestions by the Panel), and perhaps also other estimation models (e.g. ones based on state-space models).

Table 1: Results for linear regressions of catch on biomass for three different biomass series: (a) Anchovy May-June recruitment estimates (Table 12 of MARAM/IWS/DEC15/PengD/BG1), (b) Sardine May-June recruitment estimates (Table 12 of MARAM/IWS/DEC15/PengD/BG1) and (c) Sardine Oct-Dec biomass estimates (Table 11 of MARAM/IWS/DEC15/PengD/BG1). In each case, regressions are performed for the whole biomass series available up to 2013. Catches taken within 10nm of the islands (Table 8 of MARAM/IWS/DEC15/PengD/BG1) and 20nm (Table 9 of MARAM/IWS/DEC15/PengD/BG1) are considered. Note that for the recruitment biomass series (a) and (b), catch in year y is regressed against biomass from year $y-1$. Results are given for both Dassen and Robben island in terms of \bar{B} and $se(B)$, the mean and standard error of the biomass series under consideration, \bar{C} , the mean of the catch series under consideration, \hat{m} and $se(\hat{m})$, the estimate and standard error of the slope of the regression, the correlation coefficient r and coefficient of determination r^2 , and lastly the regression residual standard deviation.

(a) Anchovy May-June (Table 12)			\bar{B}	$se(B)$	\bar{C}	\hat{m}	$se(\hat{m})$	r	r^2	resid sd
1987-2013	10nm	Dassen	0.803	0.637	6.03	2.96	1.39	0.405	0.164	4.34
		Robben	0.850	0.656	7.18	1.58	1.69	0.195	0.038	5.34
		Average	0.827	0.647	6.61	2.27	1.54	0.300	0.101	4.84
	20nm	Dassen	0.803	0.637	30.76	9.30	5.30	0.343	0.118	16.56
		Robben	0.850	0.656	22.77	4.33	4.47	0.202	0.041	14.07
		Average	0.827	0.647	26.77	6.82	4.89	0.273	0.080	15.32
(b) Sardine May-June (Table 12)			\bar{B}	$sd(B)$	\bar{C}	\hat{m}	$se(\hat{m})$	r	r^2	resid sd
1987-2013	10nm	Dassen	0.166	0.184	1.55	3.04	1.51	0.387	0.150	1.36
		Robben	0.165	0.190	0.99	3.20	1.47	0.422	0.178	1.33
		Average	0.166	0.187	1.27	3.12	1.49	0.405	0.164	1.35
	20nm	Dassen	0.166	0.184	10.97	17.03	9.77	0.342	0.117	8.80
		Robben	0.165	0.190	5.62	20.43	5.36	0.631	0.398	4.87
		Average	0.166	0.187	8.30	18.73	7.57	0.487	0.258	6.84
(c) Sardine Oct-Dec (Table 11)			\bar{B}	$sd(B)$	\bar{C}	\hat{m}	$se(\hat{m})$	r	r^2	resid sd
1987-2013	10nm	Dassen	0.453	0.360	1.55	2.02	0.72	0.503	0.253	1.27
		Robben	0.455	0.370	0.99	1.53	0.76	0.393	0.155	1.35
		Average	0.454	0.365	1.27	1.78	0.74	0.448	0.204	1.31
	20nm	Dassen	0.453	0.360	10.97	12.42	4.65	0.487	0.237	8.18
		Robben	0.455	0.370	5.62	8.55	3.03	0.515	0.266	5.38
		Average	0.454	0.365	8.30	10.49	3.84	0.501	0.252	6.78

Table 2: Chick growth data used for results presented to date. Data are from Table 2 of MARAM/IWS/DEC15/PengD/BG1.

Year	Island	Growth rate (median growth coefficient)	Sample size (no. of chicks)
1989	Dassen	0.0317	86
1995	Dassen	0.0374	111
1996	Dassen	0.04	145
1997	Dassen	0.0446	136
1998	Dassen	0.0395	46
2008	Dassen	0.0268	14
2009	Dassen	0.0315	70
2010	Dassen	0.0328	45
2011	Dassen	0.0288	19
2004	Robben	0.0461	142
2008	Robben	0.0369	50
2009	Robben	0.0312	88
2011	Robben	0.0329	54
2012	Robben	0.0347	80

Table 3: Forage trip duration data used for results presented to date. Data are from Table 5 of MARAM/IWS/DEC15/PengD/BG1.

Year	Island	Forage trip duration	Sample size
2003	Dassen	9.6	3
2004	Dassen	12.5	29
2008	Dassen	22.6	10
2009	Dassen	17	3
2010	Dassen	13.8	14
2011	Dassen	17.6	22
2012	Dassen	12.7	37
2013	Dassen	10.7	21
2003	Robben	11	9
2008	Robben	25.76	10
2010	Robben	18	10
2011	Robben	16.8	25
2012	Robben	11	35
2013	Robben	19.3	12

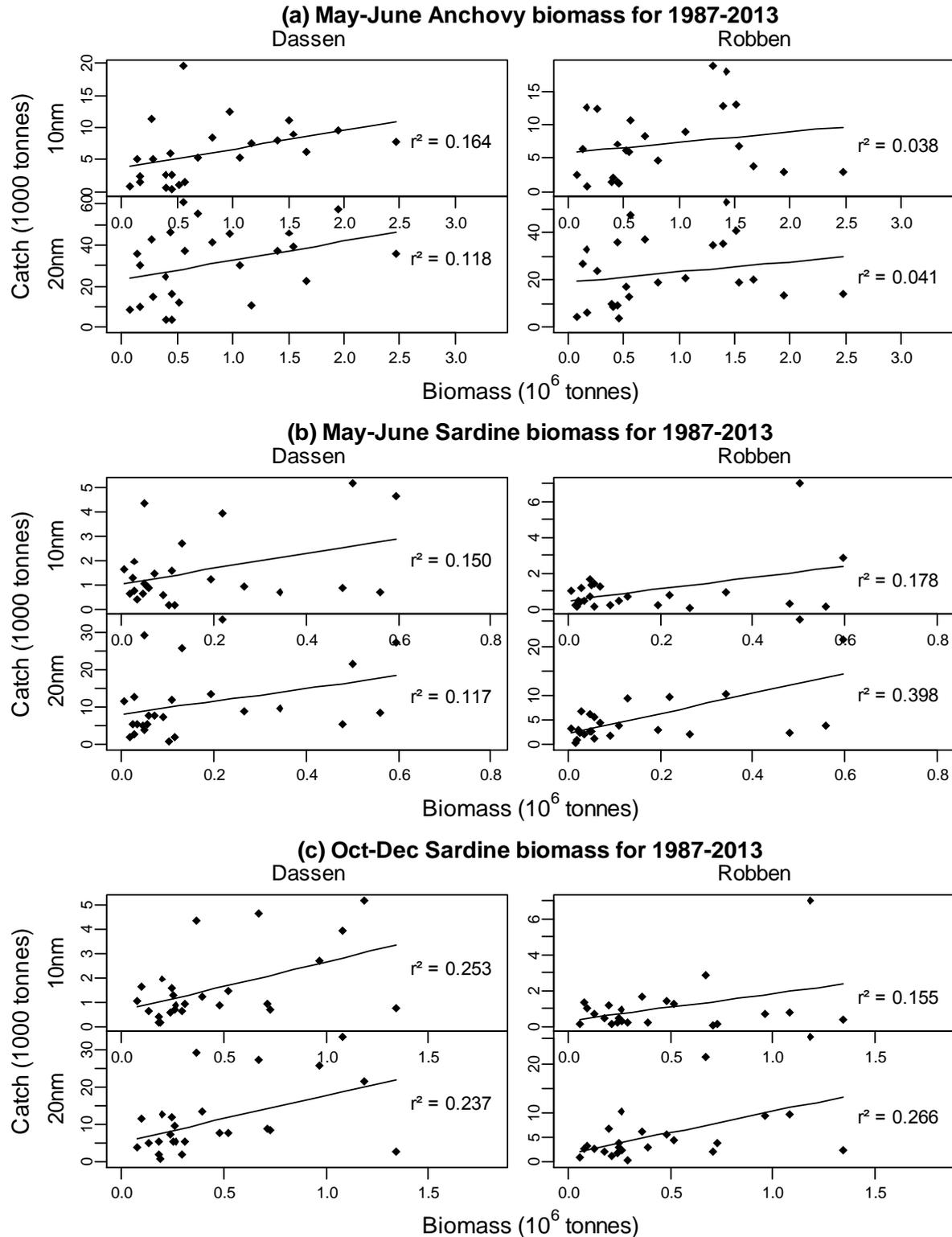
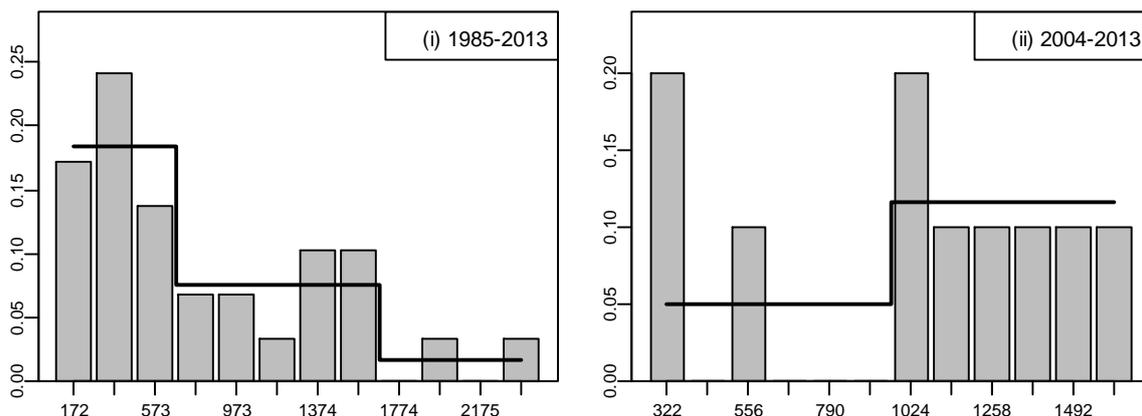
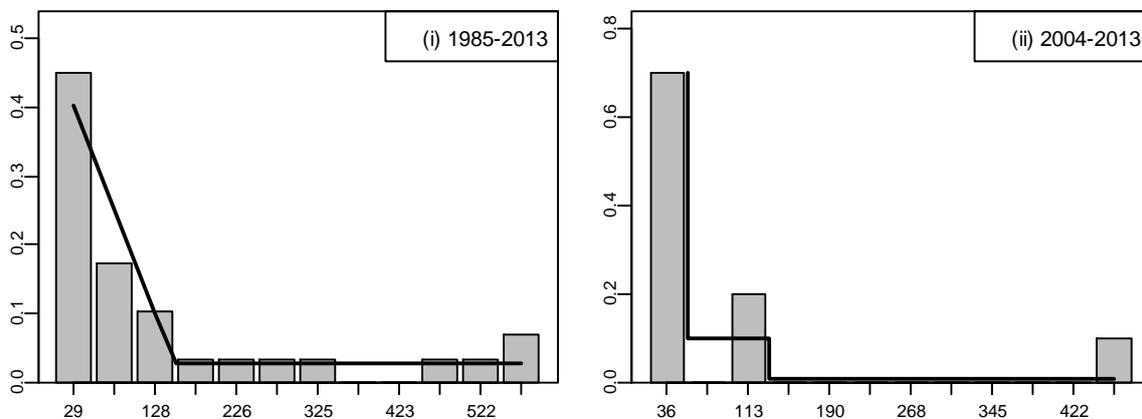


Figure 1: Catches taken within 10nm and 20nm of both Robben and Dassen island are plotted against (a) the anchovy May-June recruitment biomass, (b) the sardine May-June recruitment biomass and (c) the sardine October-December biomass (from the previous year). Regression slopes and the regression r^2 are shown.

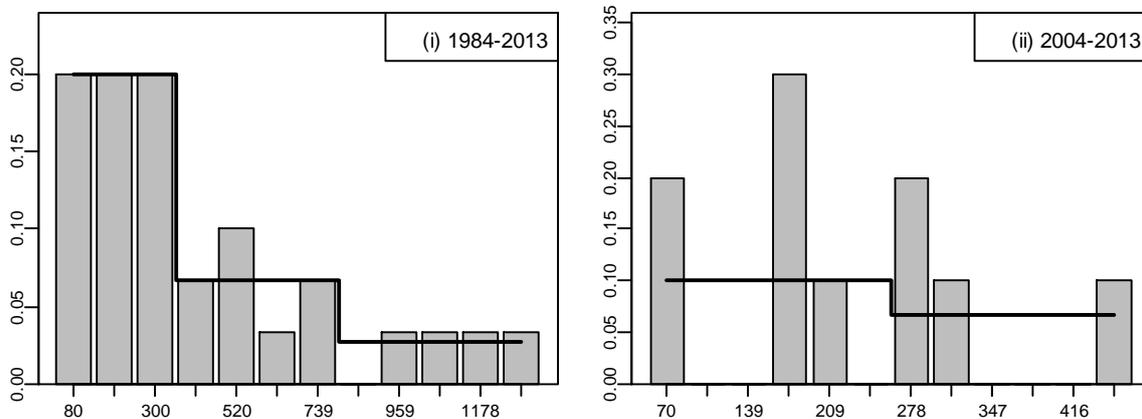
(a) May-June Anchovy biomass



(b) May-June Sardine biomass



(c) Oct-Dec Sardine biomass



Biomass ('000 tonnes)

Figure 2 Illustration of the distributions of the anchovy and sardine biomass estimates (grey bars) along with suggested curves to fit the distributions. Figures 2(a) and (b) show the anchovy and sardine plots for recruitment estimates (May/June) for the standard survey area up to Cape Infanta (Table 12 of Coetzee 2015) for two different time series. Figure (c) gives the distributions for biomass of sardine measured during the October-December acoustic survey to the west and east of Cape Agulhas (Table 11 of Coetzee 2015) for two time periods. Note that Coetzee (2015) incorporates some corrections to data used earlier, so that these plots are changed slightly from those circulated previously.

Reference:

Coetzee, J. 2015. The current set of available data for evaluation of the Island Closure Feasibility study. DAFF document MARAM/IWS/DEC15/PengD/BG1