Question Q1: Summary document

This document combines pertinent comments and responses contained in various documents submitted to the Small pelagic Working Group during the course of 2020 that are associated with Question Q1.

QUESTION Q1: It has been asserted (see FISHERIES/2020/SEP/SWG-PEL/96rev) that the estimates of island closure effects provided in FISHERIES/2020/JUL/SWG-PEL/53REV, which are based on individual data-based analyses, are (for reasons given in FISHERIES/2020/AUG/SWG-PEL/82) unreliable and consequently unacceptable for consideration in developing management recommendations regarding possible future island closures. Do the reasons given justify this assertion?

Overview of material included under 4 steps: 1 = assertion, 2 = response, 3 = response to response, 4 = further responses. Note interpretation of document purpose below is that of Janet Coetzee. Author DSB = Doug Butterworth, RS = Richard Sherley, MOB = Mike Bergh.

	Step	Author	Original source document and main reference documents	Pg.
Exchange 1: Butterworth vs Sherley	1	DSB	**FISHERIES/2020/JAN/SWG-PEL/08 (**Debates the merits of using aggregated vs disaggregated data with reference to the 2019 International review panel report (MARAM/IWS/2019/General/5) and MARAM/IWS/2019/PENG/WP3)	2
			FISHERIES/2020/AUG/SWG-PEL/82 (Asserts that FISHERIES/2020/JUL/SWG-PEL/53REV, which updates (MARAM/IWS/DEC19/Peng/P4) to take account of recommendations made by the 2019 review panel, should not be considered reliable)	3
	2	RS	**FISHERIES/2020/AUG/SWG-PEL/83** (Response to FISHERIES/2020/JAN/SWG-PEL/08) **FISHERIES/2020/SEP/SWG-PEL/85** (Response to FISHERIES/2020/AUG/SWG-PEL/82)	15
	3	DSB	**FISHERIES/2020/SEP/SWG-PEL/96rev** (Response to FISHERIES/2020/SEP/SWG-PEL/85 and FISHERIES/2020/SEP/SWG-PEL/86) **FISHERIES/2020/OCT/SWG-PEL/103 (**Notes the absence of a mathematical response to the proof provided in FISHERIES/2020/AUG/SWG-PEL/82) **FISHERIES/2020/OCT/SWG-PEL/110** (Response to FISHERIES/2020/OCT/SWG-PEL/102) **FISHERIES/2020/OCT/SWG-PEL/111** (Response to FISHERIES/2020/OCT/SWG-PEL/105rev, a proposal based on the results in FISHERIES/2020/JUL/SWG-PEL/53REV)	18
	4	RS	No specific document, but extracts from various documents including the International review panel reports of 2016 and 2019 (MARAM/IWS/DEC16/General/7, MARAM/IWS/2019/General/5) and Sherley 2020 documents FISHERIES/2020/SEP/SWG-PEL/85-87.	23
Exchange 2: Bergh vs Sherley	1	MOB	**FISHERIES/2020/AUG/SWG-PEL/84** (Comments on FISHERIES/2020/JUL/SWG-PEL/53REV and asserts that the debate on the use of individual bird data versus standardised aggregated data needs to be resolved).	29
	2	RS	**FISHERIES/2020/SEP/SWG-PEL/87** (Response to FISHERIES/2020/AUG/SWG-PEL/84)	30
	3	MOB	**FISHERIES/2020/OCT/SWG-PEL/107** (Response to FISHERIES/2020/SEP/SWG-PEL/87) **FISHERIES/2020/OCT/SWG-PEL/113** (Response to FISHERIES/2020/OCT/SWG-PEL/105rev, a proposal based on the results in FISHERIES/2020/JUL/SWG-PEL/53REV) **FISHERIES/2020/SEP/SWG-PEL/99 (**Points out that there are still outstanding technical issues regarding the analyses presented in FISHERIES/2020/JUL/SWG-PEL/53REV and notes the existence of a mathematical proof against using those results for informing decisions	32
	4	RS	No specific document, but extracts from various documents including the International review panel reports of 2015, 2016 and 2019 (MARAM/IWS/DEC15/General/8, MARAM/IWS/DEC16/General/7, MARAM/IWS/2019/General/5) and Sherley 2020 documents FISHERIES/2020/SEP/SWG-PEL/85-87.	35

EXCHANGE 1: Butterworth/Sherley

Step 1 – Butterworth assertion:

Document FISHERIES/2020/JAN/SWG-PEL/08

On estimates of the impact of fishing from analyses of the island closure experiment which model individual penguin responses directly

D.S. Butterworth

The Panel for the 2019 International Workshop (Die *et al.*, 2019) made a number of comments about analyses of the island closure experiment results involving the use of individual observations. These included:

- For natural experiments such as the closure experiment, it is a working hypothesis that including random effects chosen using model selection methods will appropriately account for the pseudo-replication.

Implications of empirical comparisons currently available

MARAM/IWS/2019/PENG/WP3 (Sherley and Winker, 2019) provides comparisons for closure effect SE estimates based on the maximum forage distance variable for south coast penguin colonies (St Croix and Bird) and the condition variable for west coast colonies (Dassen and Robben), as follows:

South coast:	aggregated data	0.084	
	individual data	0.098	random effect: year-bird ID
		0.102	random effect: year-island
West coast:	aggregated data	0.038	
	individual data	0.023	random effect: year-month
		0.039	random effect: year-island

Notably, the individual-based estimates of SE are not necessarily robust to which random effects structure has been used. In one of the two examples above, one choice suggests precision almost twice as good as the other. In simple terms then, the "working hypothesis" (that including some random effects will appropriately account for pseudo-replication) is not supported.

However, the Panel qualified their comments about this working hypothesis by referring to the need for an appropriate random effects structure to be used, also mentioning the use of model selection approaches in that regard.

Does use of individual-based approaches remedy the limited degrees of freedom problem of estimators based on annually aggregated data?

Obtaining estimates of high precision of the fishing effect parameters in the island closure experiment, when these are based on annually aggregated data, is hampered by the low number of degrees of

freedom (dof), together with their slow accumulation over time. Effectively, adding results from one further year provides two additional data points, but adds one further estimable parameter, and so increases the dof by no more than one (though this is ameliorated somewhat if the year factor is treated as a random effect in the estimation).

Using individual data appears an attractive approach to address this problem, but does it in fact achieve any better than the aggregated data approach?

First, note that the empirical comparative results shown in the section above hardly suggest so.

But further, consider the following hypothetical limiting case situation of a small-ish number of years (say 10), a large-ish process error, and a large number at individual data for the response variable from each island each year (say 10 000 each), this in circumstances where the observation error is very small. The expected response variable value each year will then be effectively exactly determined, but the closure effect will still be rather poorly estimated because the annual mean response will nevertheless vary substantially from year to year, and the extent of this variation will contribute substantial variance to the estimate of the closure effect. The estimate for the closure effect from the aggregated and from the individual approaches will be effectively identical, and so too the standard errors for the closure effect for each. But though the dof for the former will hardly reach double figures, the dof for the latter will seemingly be close to 20 000. That's plainly in the context of using AIC for model comparison.

Clearly the structure of the problem here, and the different nature of between-year vs within-year information renders model selection involving fixed vs random effects approaches less than straightforward. It seems likely, as far as precision is concerned, that very little if anything is to be gained from pursuing an individual data compared to an annually aggregated data based approach.

References

Die, D.J., Punt, A.E., Tiedemann, R., Waples, R. and Wilberg, M.J. 2019. International Review Panel report for the 2019 International Fisheries Stock Assessment workshop. 20pp.

Sherley, R.B, and Winker, H. 2019. Some observations on comparisons of fitting to the annual means and the observation-level data for the cases in MARAM/IWS/DEC19/Peng/P4 that support a positive effect of the island closures experiment on African penguins. Document MARAM/IWS/2019/PENG/WP3. 5pp.

Document FISHERIES/2020/AUG/SWG-PEL/82

A response to Sherley: FISHERIES/2020/JUL/SWG-PEL/53REV

D. S. Butterworth

Note: For readers' ease, responses have, in the main, been inserted at appropriate points in the original document below in red, and in *italics* in the main text though not in the Annex added.

Primary overview comments

Sherley's document below, as it states, provides a response (in commendable detail) to some suggestions made by the 2019 International Review Panel regarding the selection of random effects structures for models to estimate the closure effect from the island closure experiment which Sherley and colleagues have submitted previously. That goes to the question of how best such models might remove the effects of

non-independence (or pseudo-replication) in the individual measurement data they use to prevent their providing negatively biased estimates of the standard errors of these closure effects.

However, the document fails to address the more basic question of whether, even if perhaps such removal may be achieved, the use of such individual data can provide improved (lower standard error) estimates of such precision compared to those based on annually aggregated values of the corresponding response variables. This is the issue raised, for example, in the last section of FISHERIES/2020/JAN/SWG-PEL/08, where a limiting case example is used to suggest that this may not be so.

The Annex added to this document provides a mathematical-statistical demonstration that ***this is indeed not so****. Thus even if the random effects approach to making use of individual data can fully account of their non-independence, and hence prevent this from negatively biasing estimates of the standard error of the island closure effect, the resultant estimates could not have better precision than those provided by corresponding models based on annually aggregated values of measurements of the response variables.*

The underlying reason for this is the absence of any inter-year linkage of the data sources available to provide the response variables in these analyses. Sherley confirmed at the SWG meeting on 30 July when this document was discussed that in all the instances examined, there was no such connection: for chick condition and for chick survival there is no linkage between parents or nests from one year to the next, and similarly for foraging length (maximum distance travelled) the birds used to obtain the data are not linked inter-annually. This means that the observations from one year to the next are independent, which in turn leads to a diagonal structure in terms of annual sub-matrix blocks in the variance-covariance matrix and its inverse which are used for the closure effect estimation, and that in turn leads to the key result of the Annex.

Explained less formally:

a) *Unlike the case of individual linkage, which for example enables a paired-t test to have more discriminatory power than a comparison based on means only, in this instance annual means contain all the information content of pertinence to the key effect being estimated.*
b) *Estimation of the island closure effect is one relating to inter-year (not intra-year) variation, so that use of individual data in the estimation is unable to improve estimation precision for this parameter.*

At the 30 July meeting, Sherley stated that in his view the estimates from analyses based on annually aggregated data (such as the approach developed in collaboration with the International Panel, and (re-)implemented in FISHERIES/2020/JAN/SWG-PEL/09) should not be considered (or words to similar effect) because the ratio of degrees of freedom to number of parameter values estimated is too small. The analyses in the Annex show that the individual based approach (applied in such a way that there is adequate correction for non-independence/pseudo-replication effects) cannot improve on this precision. From this it therefore follows that Sherley's results below should also not be considered. I would not, however, agree that this is an adequate reason not the consider the results of FISHERIES/2020/JAN/SWG-PEL/09. Admittedly the nature of the island closure experiment is such that the number of degrees of freedom is limited, with effectively each extra year adding two more data points and one additional (year-effect) estimable parameter only, and hence only one further degree of freedom. Nevertheless, the results for standard errors of the closure effect estimates in Table 2 of FISHERIES/2020/JAN/SWG-PEL/09 do show that meaningful results may be obtained from the annually aggregated data that have become available from this experiment.

The individual-data-based approach can thus at best equal the aggregated data approach in terms of precision of the estimates of island closure effects, but only provided that there is complete adjustment for the non-independence effects through the use of random effects models. This requires first that the random effects structure is appropriately chosen, as the document below addresses; but even if that is the case, as the 2019 Panel stated, in natural experiments such as the island closure experiment, it remains only "a working hypothesis that including random effects chosen using model selection methods will appropriately account for the pseudo-replication". Thus, even if best practice is used to select the random effects structure, this provides no guarantee that the closure effect standard error estimates arising will not be negatively biased, and to an unknown extent. Hence, why consider the results from these models, when the aggregated approach already accounts for within-year data non-independence without raising this concern?

Overview summary

The individual-data-based estimates of the closure effect are indicated by the Annex to be unable in principle to provide any improvement on annually aggregated analyses. To the extent that they might appear to do so, no guarantee can be provided that this appearance is not a consequence of a failure of the random effects approach used to account for all sources of non-independence in the data.

Why therefore proceed further with any comments on the document below, since its results cannot be used as a basis for decisions regarding the implications of the island closure experiments? This is only because other useful discussion points arise co-incidentally from this text, which may well assist in further analyses (e.g. standardisation of annual aggregate measurements for co-variates), consideration and interpretation of results for this experiment.

3.1. Chick Condition, Western Cape

Table 1. Model selection results for the candidate models with different random effect structures, tested to assess the impact of the fishing closures on African penguin chick condition at Robben and Dassen Islands. M3 (Year/Month) corresponds to the original model presented in Sherley et al. (2019). Effect sizes marked in bold text are credibly different from zero (≥ 97.5% of the posterior > 0). Models are ranked by PSIS–LOO value (the *smaller* the PSIS–LOO, the better the relative model fit).

Model Number	Random effects structure	WAIC	PSIS–LOO	Stacking weight	Robben Closure effect mean (95% HPDI)	Dassen Closure effect mean (95% HPDI)
M1	Island/Year/Month	10365.9	10366.2	0.946	0.07 (−0.01–0.14)	0.03 (−0.03–0.10)
M3	Year/Month	10680.5	10680.7	0.022	**0.10 (0.05–0.14)**	−0.002 (−0.05–0.04)
M4	Island/Month	11348.0	11348.0	0.002	**0.10 (0.08–0.12)**	0.01 (−0.02–0.03)
M6	Month	11449.9	11449.9	0.019	**0.10 (0.08–0.12)**	−0.002 (−0.03–0.02)
M2	Island/Year	11499.6	11499.6	0.000	0.08 (−0.01–0.16)	0.02 (−0.07–0.10)
M5	Year	11582.6	11582.6	0.012	**0.11 (0.06–0.15)**	0.01 (−0.03–0.06)
Model-averaged results					0.07 (−0.01–0.14)	0.03 (−0.03–0.10)

Notes: / denotes nesting of the random effects, thus Island/Year/BirdID = Month nested in Year, nested in Bird Identity. WAIC = Widely Applicable Information Criterion (Watanabe 2010). PSIS–LOO = Pareto smoothed importance sampling, leave-one-out cross-validation (PSIS–LOO; Vehtari et al. 2019a). HPDI = highest posterior density interval.

It is of importance to note that when Island/Year is included in the random effects structure, the 95% HPDI widens appreciably, and to the extent that for Robben Island it is no longer credibly different from zero. This corroborates the concern expressed in FISHERIES/2020/JAN/SWG-PEL/08 that failure to include this interaction was leading to negatively biased estimates of standard errors (i.e. unduly high precision) for the estimates of island closure effects. It should be noted that some earlier analyses of this nature have also not included this interaction in their random effects structure, e.g. MARAM/IWS/DEC19/Peng/P4 included only a Year/Month interaction term; this renders their conclusions questionable.

The model averaged closure effect at Robben Island represented an improvement during closed years of 23.6% (−4.9–51.9%) with 96% of all the posterior estimates > 0 and 82% > 10% (Figure 1). To put this in perspective, to be considered credibly different from zero (and thus bold in Table 1), 97.5% would need to be > 0. For Dassen Island, the corresponding model averaged estimates represented an increase of 13% (−12–39%) with 83% of all the posterior estimates > 0 and 55% > 10% (Figure 1).

The estimates and standard errors of the island closure effect given in Table 2 for the aggregated-data-based analyses of FISHERIES/2020/JAN/SWG-PEL/09 are in log-space, and so correspond roughly to proportions (multiply by 100 for percentages). Adjusting here, and also in similar comparisons below, for the sign change in the convention used, these are 0.14 (se 0.13) for Robben and 0.03 (se 0.14) for Dassen. These PEL/09 values are thus notably different for both the magnitude of the effect itself (smaller) and the associate precision (less).

3.3. Chick Survival, Western Cape

The best fitting model contained the random 'Island/Year/Month' intercept and yielded an estimated Closure effect size of 0.38 (HPDI: 0.21–0.58). This corresponds to an improvement in survival of 10.3% (5.4–15.2%) at Robben Island and 10.6% (5.2–16.2%) at Dassen Island when the closure was in place. And, as with Max. foraging distance at the Easern Cape islands, this effect was unambiguous, with 100% of the posterior indicating a positive effect of the Closure. Respectively, 53% and 57% of the posterior distribution exceeded the 10% threshold for management action at Robben Island and Dassen Island.

The estimates and standard errors of the island closure effect given in Table 2 for the aggregated-data-based analyses of FISHERIES/2020/JAN/SWG-PEL/09 are 0.04 (se 0.11) for Robben and 0.13 (se 0.10) for Dassen. Similar to results for chick condition then, these PEL/09 values are thus notably different for both the effect itself (smaller) for Dassen and the associated precision (less) for both islands. Part of the improved precision here is arising from assuming that the closure parameter is the same for both islands – see comment immediately preceding Figure 1 above which questions the appropriateness of this assumption.

[The comment referenced immediately above:

*Model selection must also be informed by relevant external information when available, and not only the analysis of the data from the experiment in question **alone**. The totality of the estimates of the island closure effect from various sources in Table 2 of FISHERIES/2020/JAN/SWG-PEL/09 are strongly suggestive of a real difference in the values for each island. Thus, the results above without an Island/Closure interaction, while of themselves indicative of some average value across the two islands, cannot be used reliably to draw inferences about values for either island separately.]*

3.5. Conclusions and next steps

It has been argued that the effects presented in Sherley et al. (2019) were not robust because the allegedly poorly chosen random effect structure resulted in fixed effect estimates that were overly precise

(Butterworth and Ross-Gillespie 2019). However, the results presented here suggest little meaningful change in inference whether or not Island was included as a higher-level random effect.

This comment is completely at variance with the results shown earlier in Table 1 (see also the comments made immediately below that). Quite clearly results for precision are distinctly non-robust to decisions made about which factors to include in the random effects considered in that case. Consequently, results reported in earlier papers have been incorrect because this selection process was not carried out appropriately.

This result has also just been supported by a new simulation study on the robustness of linear mixed-effects models which concluded that "missing random effect predictors had little effect on the fixed effect estimates but had systematic effects on the estimates of random effects. The variance due to unmodelled higher level predictors [like Island in our case] was almost completely absorbed by the nested random effect variance of interest" (Schielzeth et al. 2020).

This study provides results similar to those reported in FISHERIES/2016/NOV/SWG-PEL/65 (MARAM/IWS/DEC16/Peng Clos/P3) which showed that the appropriate adjustment using a random effects approach of estimates of precision for non-independent data were reasonably robust to choice of the unknown time scale at which that effect modelled was operating. Whatever, the Schielzeth et al. study is not pertinent to the key aspect of the analyses in question here, because the models which it considers do not include year-independence in a diagonal sub-matrix block structure as pertinent here to the estimation of the parameter of interest – the closure effect (see Annex and the Primary overview comments above).

Some final comments

Similarly, the results of Table 1 seemingly make clear that the inclusion of Island/Year interactions (elsewhere termed "process errors") are essential. After all, the basis underlying the whole experiment is that closure and year effects are confounded, but that their distinction is made possible by considering two nearby islands (e.g. the Dassen-Robben pair) for which the year effect (related to forage fish densities) is likely to be similar because they are not that far apart. But of course the differential effect between the two islands will vary from year to year, necessitating the presence of this process error term in the model equations. Furthermore, this term is quantitatively important, as shown by the analyses of MARAM/IWS/DEC15/PengD/P2 which indicated that this process error is far larger than the observation errors remaining after averaging over the number of island-specific measurements typically available each year. It may be that improved standardisation for covariates might further reduce the extent of that observation error, but there have been no indications to date that that might be the case, and in any case this seems unlikely given the process error dominance which is already apparent in the results from the experiment.

References

Butterworth DS. 2016. On the use of aggregated vs individual data in assessment models. MARAM/IWS/DEC16/Peng Clos/P3. Pp.1-6

Butterworth DS and Ross-Gillespie A. 2019. Is pseudo-replication biasing results from analyses from the island closure experiment which model individual penguin responses directly? FISHERIES/2019/NOV/SWG-PEL/34. Pp. 1-10.

Penguin Island Closure Task Team (M.O. Bergh, D.S. Butterworth, K.L. Cochrane (chair), T.L. Morris, R.B. Sherley and H.Winker). 2015. Consolidated analyses produced in implementation of the approaches described in document MARAM/IWS/DEC15/PengD/P1. MARAM/IWS/DEC15/PengD/P2. Pp. 1-40.

Ross-Gillespie A and Butterworth DS. 2020. Updated implementation of the Algorithm recommended by the Panel for the 2016 International Stock Assessment Workshop for assessing whether or not to continue with the penguin island closure experiment. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/JAN/SWG-PEL/09. Pp. 1–17.

Schielzeth H, Dingemanse N, Nakagawa S, Westneat DF, Allegue H, Teplitsky C, Réale D, Dochtermann NA, Garamszegi L and Araya-Ajoy Y. 2020. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*. DOI: 10.1111/2041-210X.13434.

Annex

A demonstration why use of individual data constitutes pseudo-replication in the estimation of the island closure effect from penguin response variables

This demonstration is best facilitated by first considering a simpler case – the estimation of a mean of values collected over a number of years.

(I) Estimation of a mean

(a) An exact measurement each year

Basic model: $x_y = \mu + \eta_y$ where $\eta_y \sim N(0, \sigma^2)$ (1)

where: x_y is measured exactly
the annual variance is $\sigma^2 = v$
there are $\mathcal{N}$ years of observations

μ is estimated by minimising the following equation with respect to μ.

$$-lnL = \sum_{y=1}^{\mathcal{N}} \left[ln\sigma + \frac{1}{2\sigma^2}(x_y - \mu)^2 \right] \quad (2)$$

For convenience here, focus on

$$SS = \sum_{y=1}^{\mathcal{N}} \frac{1}{2\sigma^2}(x_y - \mu)^2 = \frac{1}{2v}\sum_{y=1}^{\mathcal{N}} (x_y - \mu)^2 \quad (3)$$

Then the estimate of the variance of μ is provided by:

$$var(\mu) = 1/\frac{d^2SS}{d\mu^2} = 1/\left\{ \frac{1}{2v}\sum_{y=1}^{\mathcal{N}} 2 \right\} = \frac{v}{\mathcal{N}} \quad (4)$$

In vector-matrix notation, this can be written as:

$$SS = \underline{z}^T V^{-1} \underline{z} \quad (5)$$

where $\underline{z}^T = (x_1 - \mu, x_2 - \mu, \ldots, x_{\mathcal{N}} - \mu)$ and the variance-covariance matrix V in this case is the $(\mathcal{N} \times \mathcal{N})$ matrix:

$$V = \begin{bmatrix} v & 0 & \ldots & 0 \\ 0 & v & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & v \end{bmatrix} \quad (6)$$

so that

$$V^{-1} = \begin{bmatrix} 1/v & 0 & \ldots & 0 \\ 0 & 1/v & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/v \end{bmatrix} = \frac{1}{v}\begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \quad (7)$$

(b) A number of measurements each year

Basic model: $x_{y,i} = \mu + \eta_y + \epsilon_{y,i}$ (8)

where: $\eta_y \sim N(0, \sigma_\eta^2)$; $\epsilon_{y,i} \sim N(0, \sigma_\epsilon^2)$

$\sigma_\eta^2 = v$; $\sigma_\epsilon^2 = \omega v$ (i.e. $\omega = \sigma_\epsilon^2/\sigma_\eta^2$)

and $i = 1,2,\dots,n$ i.e. for simplicity the case of the same number n of measurements each year is considered.

Note $var(x_{y,i}) = \sigma_\eta^2 + \sigma_\epsilon^2 = v(1+\omega)$ (9)

because $\epsilon_{y,i}$ is independent of η_y.

Consider first the example $\mathcal{N} = 2, n = 3$. The variance-covariance matrix V takes the form:

$$V = \begin{bmatrix} \sigma_\eta^2 + \sigma_\epsilon^2 & \sigma_\eta^2 & \sigma_\eta^2 & 0 & 0 & 0 \\ \sigma_\eta^2 & \sigma_\eta^2 + \sigma_\epsilon^2 & \sigma_\eta^2 & 0 & 0 & 0 \\ \sigma_\eta^2 & \sigma_\eta^2 & \sigma_\eta^2 + \sigma_\epsilon^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_\eta^2 + \sigma_\epsilon^2 & \sigma_\eta^2 & \sigma_\eta^2 \\ 0 & 0 & 0 & \sigma_\eta^2 & \sigma_\eta^2 + \sigma_\epsilon^2 & \sigma_\eta^2 \\ 0 & 0 & 0 & \sigma_\eta^2 & \sigma_\eta^2 & \sigma_\eta^2 + \sigma_\epsilon^2 \end{bmatrix}$$

$$= v \begin{bmatrix} 1+\omega & 1 & 1 & 0 & 0 & 0 \\ 1 & 1+\omega & 1 & 0 & 0 & 0 \\ 1 & 1 & 1+\omega & 0 & 0 & 0 \\ 0 & 0 & 0 & 1+\omega & 1 & 1 \\ 0 & 0 & 0 & 1 & 1+\omega & 1 \\ 0 & 0 & 0 & 1 & 1 & 1+\omega \end{bmatrix} \quad (10)$$

Note that at the "year block" level, V has a diagonal form; the off-diagonal blocks are zero because $x_{y1,i}$ is independent of $x_{y2,i}$ for $y1 \neq y2$.

The inverse is then:

$$V^{-1} = \frac{1}{v} \frac{1}{3\omega + \omega^2} \begin{bmatrix} 2+\omega & -1 & -1 & 0 & 0 & 0 \\ -1 & 2+\omega & -1 & 0 & 0 & 0 \\ -1 & -1 & 2+\omega & 0 & 0 & 0 \\ 0 & 0 & 0 & 2+\omega & -1 & -1 \\ 0 & 0 & 0 & -1 & 2+\omega & -1 \\ 0 & 0 & 0 & -1 & -1 & 2+\omega \end{bmatrix} \quad (11)$$

Again $SS = \underline{z}^T V^{-1} \underline{z}$

where now $\underline{z}^T = (x_{1,1} - \mu, x_{1,2} - \mu, x_{1,3} - \mu; x_{2,1} - \mu, x_{2,2} - \mu, x_{2,3} - \mu)$ (12)

so that

$$SS = \left(\frac{1}{2v}\frac{1}{3\omega+\omega^2}\right)\times$$
$$\left[(2+\omega)(x_{1,1}-\mu)^2+(2+\omega)(x_{1,2}-\mu)^2+(2+\omega)(x_{1,3}-\mu)^2\right.$$
$$-2(x_{1,1}-\mu)(x_{1,2}-\mu)-2(x_{1,1}-\mu)(x_{1,3}-\mu)-2(x_{1,2}-\mu)(x_{1,3}-\mu) \qquad (13)$$
$$+(2+\omega)(x_{2,1}-\mu)^2+(2+\omega)(x_{2,2}-\mu)^2+(2+\omega)(x_{2,3}-\mu)^2$$
$$\left.-2(x_{2,1}-\mu)(x_{2,2}-\mu)-2(x_{2,1}-\mu)(x_{2,3}-\mu)-2(x_{2,2}-\mu)(x_{2,3}-\mu)\right]$$

$$\frac{d^2SS}{d\mu^2} = \frac{1}{2v}\frac{1}{\omega(3+\omega)}\sum_{y=1}^{2}[(2+\omega)(2)(3)-(2)(2)(3)]$$
$$= \frac{1}{2v}\frac{1}{\omega(3+\omega)}\sum_{y=1}^{2}6\omega = \frac{1}{v}\left(\frac{3}{3+\omega}\right)(2) \qquad (14)$$

Hence:
$$var(\mu) = 1/\frac{d^2SS}{d\mu^2} = \frac{v}{2}\left(1+\frac{\omega}{3}\right) \qquad (15)$$

For general $\mathcal{N}$ and n, the V^{-1} matrix takes the blocked form:

$$V^{-1} = \frac{1}{v}\frac{1}{\omega(n+\omega)}\begin{bmatrix} B & 0 & \dots & 0 \\ 0 & B & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & B \end{bmatrix} \qquad \mathcal{N}\times\mathcal{N} \text{ blocks} \qquad (16)$$

where each $n \times n$ block sub-matrix B has the form:

$$B = \begin{bmatrix} n-1+\omega & -1 & \dots & -1 \\ -1 & n-1+\omega & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & n-1+\omega \end{bmatrix} \qquad (17)$$

from which it follows that:

$$\frac{d^2SS}{d\mu^2} = \left(\frac{\mathcal{N}}{v}\right)\left(\frac{n}{n+\omega}\right) \qquad (18)$$

so that
$$var(\mu) = \frac{v}{\mathcal{N}}\left(1+\frac{\omega}{n}\right) \qquad (19)$$

Note then that:

$$var(\mu) = \frac{v}{\mathcal{N}} + \frac{v\omega}{\mathcal{N}n} = \frac{\sigma_\eta^2}{\mathcal{N}} + \frac{\sigma_\epsilon^2}{\mathcal{N}n} \qquad (20)$$

In contrast, had all the $\mathcal{N} \times n$ measurements been considered independent, the variance of μ would have been estimated (incorrectly) as:

$$var^*(\mu) = \frac{\sigma_\eta^2+\sigma_\epsilon^2}{\mathcal{N}n} = \frac{\sigma_\eta^2}{\mathcal{N}n} + \frac{\sigma_\epsilon^2}{\mathcal{N}n} \qquad (21)$$

i.e. $var^*(\mu) < var(\mu)$ suggesting better precision than actually applies.

The error in $var^*(\mu)$ arises because of the $\frac{1}{n}$ factor in the first term on the RHS of equation (21), which has failed to take account of pseudo replication/non-independence of the data.

(II) <u>Extension to estimation of the island closure effect</u>

(a) An exact measurement each year

Consider the simple case of alternate closures of one of two islands each year, the same size of the closure effect δ for each island, and $\mathcal{N}$ an even number. The basic model is then:

$$\begin{aligned} &\text{Island } j=1 \qquad x_{y,1} = a_1 + b_y + \delta X_y + \eta_{y,1} \qquad X_y = \begin{cases} 1 \ for\ y\ odd \\ 0\ for\ y\ even \end{cases} \\ &\text{Island } j=2 \qquad x_{y,2} = a_2 + b_y + \delta X'_y + \eta_{y,2} \qquad X'_y = \begin{cases} 0 \ for\ y\ odd \\ 1\ for\ y\ even \end{cases} \end{aligned} \tag{22}$$

where: a_j is an island effect,

b_y is a year effect common to both islands, and

$\eta_{y,1} \sim N(0, \sigma_\eta^2)$ and $\eta_{y,2} \sim N(0, \sigma_\eta^2)$ are independent of each other.

The SS function to be minimised to estimate δ then takes the form:

$$\begin{aligned} SS = \frac{1}{2\sigma_\eta^2}\Bigg[&\sum\nolimits_{y=1(2)\mathcal{N}-1}\left\{\left(x_{y,1}-a_1-b_y-\delta\right)^2+\left(x_{y,2}-a_2-b_y\right)^2\right\} \\ &+\sum\nolimits_{y=2(2)\mathcal{N}-1}\left\{\left(x_{y,1}-a_1-b_y\right)^2+\left(x_{y,2}-a_2-b_y-\delta\right)^2\right\}\Bigg] \end{aligned} \tag{23}$$

so that

$$\begin{aligned} \frac{\partial^2 SS}{\partial\delta^2} &= \frac{1}{2\sigma_\eta^2}\left[\sum\nolimits_{y=1(2)\mathcal{N}-1}\{2+0\}+\sum\nolimits_{y=2(2)\mathcal{N}-1}\{0+2\}\right] \\ &= \frac{1}{2\sigma_\eta^2}\left[\frac{\mathcal{N}}{2}(2)+\frac{\mathcal{N}}{2}(2)\right] = \frac{1}{2\sigma_\eta^2}(2\mathcal{N}) \end{aligned} \tag{24}$$

Thus:
$$var(\delta) = 1/\frac{\partial^2 SS}{\partial\delta^2} = \frac{\sigma_\eta^2}{\mathcal{N}} \tag{25}$$

(b) A number of measurements each year

As for the example above for the mean, the model now extends to the following, assuming the same number *n* of measurements at each island each year.

$$\begin{aligned} &\text{Island } j=1 \qquad x_{y,1,i} = a_1 + b_y + \delta X_y + \eta_{y,1} + \epsilon_{y,1,i} \\ &\text{Island } j=2 \qquad x_{y,2,i} = a_2 + b_y + \delta X'_y + \eta_{y,2} + \epsilon_{y,2,i} \end{aligned} \tag{26}$$

where $\epsilon_{y,1,i} \sim N(0, \sigma_\epsilon^2)$ and $\epsilon_{y,2,i} \sim N(0, \sigma_\epsilon^2)$ are independent of each other and of $\eta_{y,1}$ and $\eta_{y,2}$. (Note that elsewhere η_y is conventionally termed process error and $\epsilon_{y,i}$ observation error.)

Because of this independence, the function SS to be minimised:

$$SS = \underline{z}_p^T V_p^{-1} \underline{z}_p \qquad (27)$$

will have both its variance-covariance matrix V_p and the inverse thereof V_p^{-1} of exactly the same blocked form with year (and now also island) as in equations (16) and (17) above.

For exactly the same reasons as in section (I) for the example of the mean then:

$var(\delta) = \frac{\sigma_\eta^2}{\mathcal{N}}$ for an exact measurement each year is generalised to

$var(\delta) = \frac{\sigma_\eta^2}{\mathcal{N}} + \frac{\sigma_\epsilon^2}{n\mathcal{N}}$ for *n* measurements at each island each year. (28)

(c) Implications

For equation (26), instead of treating each measurement individually, estimate for the mean of these measurements ($\bar{x}_{y,i}$) each year:

$$\begin{aligned} &\text{Island } j = 1 \qquad \bar{x}_{y,1} = a_1 + b_y + \delta X_y + \eta'_{y,1} \\ &\text{Island } j = 2 \qquad \bar{x}_{y,2} = a_2 + b_y + \delta X'_y + \eta'_{y,2} \end{aligned} \qquad (29)$$

where $\eta'_{y,1} \sim N(0, \sigma_{\eta'}^2)$ and $\eta'_{y,2} \sim N(0, \sigma_{\eta'}^2)$ are independent and

$$\sigma_{\eta'}^2 = \sigma_\eta^2 + \frac{\sigma_\epsilon^2}{n} \qquad (30)$$

as the $\eta's$ and $\epsilon's$ are independent so that the second term on the RHS follows from the equation for the standard error of the mean.

Consequently, from equation (25)

$$var(\delta) = \frac{\sigma_{\eta'}^2}{\mathcal{N}} = \frac{\sigma_\eta^2}{\mathcal{N}} + \frac{\sigma_\epsilon^2}{n\mathcal{N}} \qquad (31)$$

i.e. identical to equation (28).

Hence:

(1) Taking additional measurements at each island each year (increasing *n*), reduces the standard error of the estimate of δ.

(2) However, **the *same standard error for* δ is estimated whether the estimation is carried out for the individual measurements, or for their annual means**.

The result (2) shows that as far as estimates of the precision of the closure effect in the island closure experiment is concerned, there is no advantage to be gained from analysing the individual data each year rather than an aggregate value such as their means. The former involves pseudo replication unless the covariance of the data is taken into account; provided that is done properly, exactly same results for this precision are necessarily to be expected.

Notes

(A) The analyses above have been conducted for special cases (e.g. equal numbers of measurements each year). However, if the key result that an individual-response-variable-based estimate of the island closure effect cannot have better precision than one based on the annual aggregated values of such variables, it must also be valid for more general cases.

(B) In practice (e.g. FISHERIES/2020/JAN/SWG-PEL/09) estimation using the equivalent of equation (22) treats b_y as a random rather than a fixed effect. In essence this does no more than stabilize/ robustify estimates against the possible undue influence of outlier measurements. Strictly variance estimates in these fixed and random effects structures differ in that the latter include also the variance of the random effects, but in practice the difference is minimal for the situation under consideration here (compare for example the results for standard errors of δ for EMB and EMC for operating model OM2 in Figure 1B of FISHERIES/2019/NOV/SWG-PEL/34 – note that EMB and EMC differ only with respect to whether b_y is treated as a random or a fixed effect respectively in the estimation).

Step 2 – Sherley responses: Extracts (in blue) from:

Document FISHERIES/2020/AUG/SWG-PEL/83 "Some comments on FISHERIES/2020/JAN/SWG-PEL/08" by Richard B. Sherley

And

FISHERIES/2020/SEP/SWG-PEL/85 "A response to Butterworth: FISHERIES/2020/AUG/SWG-PEL/82" by Richard B. Sherley

HEADLINE: SWG-PEL/08 and SWG-PEL/82 cover old ground already dealt with by the 2016 and 2019 IWS panels.

Extract from R19, page 20 of FISHERIES/2020/SEP/SWG-PEL/85: "the 2019 IWS panel have already given an opinion… "*Given the nature of the experiment, use of individual data is to be preferred… Die et al. (2019)*".

HEADLINE: Butterworth's two key arguments against the disaggregated approach are contradictory.

Extract from R2, pages 3 and 4 of FISHERIES/2020/SEP/SWG-PEL/85: "On the one hand, he says that the results cannot be used as a basis for decisions because they might produce negatively biased estimates of the standard errors of closure effect compared to the aggregated data approach (i.e. the standard errors are smaller than obtained from the aggregated data approach). On the other hand, he also contends that the results presented in FISHERIES/2020/JUL/SWG-PEL/53REV should not be considered further because that document has NOT shown that they can provide improved estimates of such precision (i.e. smaller standard errors than those obtained from the aggregated data approach".

"Is FISHERIES/2020/JUL/SWG-PEL/53REV to be disregarded because Butterworth… maintains that it DOES yield smaller standard errors than those obtained from the aggregated data approach? Or are we expected to accept the complete opposite contention that FISHERIES/2020/JUL/SWG-PEL/53REV should be disregarded because its approach DOES NOT produce smaller standard errors than those obtained from the aggregated data approach?

HEADLINE: Statistical text books and the peer-reviewed literature are clear that mixed-effects models are reliable and should be used in this situation.

Extract from Comment 3, page 3 of FISHERIES/2020/SEP/SWG-PEL/83: Simulation studies have, however, demonstrated that linear mixed-effects models (LMMs) can be used in these circumstances, when random effects are chosen based on the known sampling structure in the data (Silk et al. 2020) and when model selection methods are used to choose the random effect structure (Matuschek et al. 2017). As Matuschek et al. (2017) put it: "Our simulations have shown that determining a parsimonious model with a standard model selection criterion is a defensible choice to find this middle ground between Type I error rate and power". Again, this is why I understood the panel to make the following recommendation in their 2019 report: "Model selection methods should be applied to select an appropriate random effects structure" and why [FISHERIES/2020/JUL/SWG-PEL/53REV] has been presented to address that recommendation.

Extract from R1, page 2 and 3 of FISHERIES/2020/SEP/SWG-PEL/85: "[The] use of parsimonious mixed models… improves the balance between a Type I error and statistical power (e.g. Matuschek et al. 2017,

Bates et al. 2018, Silk et al. 2020), and so allows exactly this kind of one-step approach in FISHERIES/2020/JUL/SWGPEL/ 53REV... mixed-effects models have even been advocated for and used in fisheries management for more than a decade (Venables & Dichmont 2004, Punt et al. 2006, Thorson & Minto 2015, Thorson et al. 2016), including by Butterworth himself (Brandão et al. 2004)".

HEADLINE: Even the maximal mixed-effects models in SWG-PEL/53REV, which include 'Island' in both the fixed and random components of the model, yield closure effects with >95% probability.

Extract from R6, page 8 of FISHERIES/2020/SEP/SWG-PEL/85: "there is still the question of whether Island should simultaneously be included in both the fixed and random components of these models and whether Island, which only has two levels, should be included in the random effect structure at all…

M1 in the table above is the maximal model (the most complex possible random effect structure); maximal models are "generally wasteful and costly in terms of statistical power for testing hypotheses" (Stroup 2012, pg. 185) and maximal models – even when they converge – can result in overparameterization that leads to uninterpretable models (Bates et al. 2018).

Furthermore, the maximal model may actually trade-off power for some conservatism beyond the nominal Type I error rate, even in cases where the maximal model matches the generating process exactly (Matuschek et al. 2017). Nevertheless, it presents a > 96% probability (given the data and model structure) of a closure effect at Robben Island. Ignoring this, particularly given that an independent analysis by Ross-Gillespie and Butterworth (FISHERIES/2020/JAN/SWG-PEL/09) concluded that there was "a biologically meaningful fishing effect" on chick condition at Robben Island, using the 2004 to 2018 aggregated data would certainly risk making a Type II error about the impact of the closure".

References

Arnqvist G. 2020. Mixed models offer no freedom from degrees of freedom. Trends in Ecology and Evolution 35: 329–335.

Bates D, Kliegl R, Vasishth S and Baayen 2018. Parsimonious Mixed Models. arXiv:1506.04967v2.

Brandão A, Butterworth DS, Johnston SJ, Glazer JP. 2004. Using a GLMM to estimate the somatic growth rate trend for male South African west coast rock lobster. Fisheries Research 70: 399–349.

Crawley M. 2013. The R Book (Second Edition). Chichester: Wiley.

Die DJ, Punt AE, Tiedemann R, Waples R and Wilberg MJ. 2019. International Review Panel Report for the 2019 International Fisheries Stock Assessment Workshop, 2–5 December 2019, UCT. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/2019/General/5. Pp. 1–18.

Harrison XA, Donaldson L, Correa-Cano ME, Evan J, Fisher DN, Goodwin CED, Robinson BS, Hodgson DJ and Inger R. 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ 6: e4794.

Matuschek H, Kliegl R, Vasishth S, Baayen H and Bates D. 2017. Balancing Type I error and power in linear mixed models. Journal of Memory and Language 94: 305–315.

Punt AE, Hobday DK and Flint R. 2006. Bayesian hierarchical modelling of maturity-at-length for rock lobsters, Jasus edwardsii, off Victoria, Australia. Marine and Freshwater Research 57: 503–511.

Silk MJ, Harrison XA and Hodgson DJ. 2020. Perils and pitfalls of mixed-effects regression models in biology. PeerJ 8: e9522.

Stroup WW. 2012. Generalized linear mixed models: Modern concepts, methods and applications. Boca Raton: CRC Press.

Thorson JT and Minto C. 2015. Mixed effects: a unifying framework for statistical modelling in fisheries biology. ICES Journal of Marine Science 72: 1245–1256.

Thorson JT, Rindorf A, Gao J, Hanselman DH and Winker H. 2016. Density-dependent changes in effective area occupied for sea-bottom-associated marine fishes. Proceedings of the Royal Society B: Biological Sciences 283: 20161853.

Venables WN and Dichmont CM 2004. GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. Fisheries Research 70: 319–337.

Zuur AF, Ieno EN, Walker NJ, Saveliev AA, and Smith GS. 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, New York.

Step 3 – Butterworth response to responses (extracts from previous documents)

Note: On occasions, simply to provide the necessary context, extracts need to include more material than pertinent to Q1 in isolation. In such instance, the text concerned is shown in yellow highlight.

Document FISHERIES/2020/SEP/SWG-PEL/96rev

Summary comments on analyses of the island closure experiment

D.S. Butterworth

Analyses by Sherley (and colleagues) based on the use of individual data

Sherley and colleagues have motivated this approach as providing greater precision and reliability (through achieving a greater number of data to estimable parameters ratio) for estimates of δ. However, such approaches may provide estimates of precision (e.g. standard errors – se's) for δ that are negatively biased because of the effects of pseudo-replication. This is of concern, because it could lead to an estimate of δ being considered to be reliably established as meaningful when this is not the case. Sherley and colleagues have attempted to address this concern by the use of estimation approaches incorporating random effects terms.

Two concerns have been raised concerning their approach. The first relates to the selection of the random effects structure used. The 2019 IWS Panel recommended a procedure for choosing the best such structure. This has been reasonably implemented for more recent results, simultaneously confirming the previous associated concern that estimation of δ is not robust to alternative selections. Earlier results reported using this approach, which failed to apply this selection procedure, are therefore confirmed to have been invalid.

Nevertheless, even when such a selection approach is incorporated, such random effects approaches cannot be guaranteed to fully account for pseudo replication effects (so may still yield negatively biased estimates of se's); but in this specific case there is a second and much more important concern (which has also been raised regularly, but has never received an adequate response, in the past). This concern is related to the structure of the data available from this experiment as this impacts the optimal precision which is achievable for estimates of δ. What is critical here is that the estimates of δ are informed by inter-annual changes in the data. For all response variables considered in the experiment, there is no linkage between elements of the individual data from one year to the next (e.g. there is no information collected that provides the ability to link a penguin or nest sampled one year to a sample taken the next year), so that the individual data are statistically independent from one year to the next. A little thought makes clear then that these individual data cannot add any further information content to the estimation of δ than is already contained in their annually aggregated value, and therefore cannot improve the estimation precision for δ. This contention has now been confirmed by what amounts to a mathematical-statistical proof (the Annex of FISHERIES/2020/AUG/SWG-PEL/82). Continued acceptance of results from this individual-based approach would therefore necessarily require that proof to be shown to be invalid.

Two recent contributions by Sherley serve to strengthen concerns about results from the individual-based approach. Comparisons in Table 1 of FISHERIES/2020/SEP/SWG-PEL/86 indicate in some cases much smaller standard errors for δ for individual- compared to aggregated approaches. Given the above, this makes clear that in those cases even though the random error structure selection procedure has been applied, it has been unable to account completely for pseudo-replication, hence providing false impressions of the precision of the result. In FISHERIES/2020/SEP/SWG-PEL/85 Sherley quotes Maunder (2001)[1] as a fisheries-related

[1] CJFAS 58 (2001) 795-803

example of the equivalent of the two-step aggregated estimation approach leading to worse precision than a single step process (as in the individual-based approach). But Maunder (2001) failed to make any adjustment for pseudo-replication (the non-independence of his "individual-equivalent" data). In principle, that case could see potential utility for the Sherley individual-based approach, as those individual data involve vessels which are identifiable from one year to the next, and hence provide more inter-annual information content than annually aggregated values, unlike in this island closure case. However, this one-step process is generally not attempted in fisheries assessments for reasons which include that although random effects models may be used in the "standardization" process concerned, they are generally unable to account for all contributors to pseudo-replication effects. This necessitates a two-step process to estimate the size of "process error" (additional variance), as in the case of the Panel Algorithm of the previous section. Essentially, for this closure experiment as is typical in fisheries assessments, process error dominates observation error (MARAM/IWS/DEC15/PengD/P2).

No counter to the proof in FISHERIES/2020/AUG/SWG-PEL/82 has been offered by Sherley or his earlier co-authors. Estimates based on a methodology which an unchallenged proof has shown to be flawed are necessarily unreliable. The results from Sherley (and colleagues) based on their individual data-based analyses are consequently quite **un**acceptable for consideration in developing management recommendations regarding possible future island closures.

Possible further steps needed in moving towards management recommendations

6) <u>The process needed to compare results from different models in a comparative exercise</u>

The 2015 Panel provided the default model to be used for providing estimates of δ. Certainly use of alternative approaches for estimation is desirable to check estimation robustness, but standardly in fisheries assessments this is required to be done through a "building-a-bridge" approach whereby factors that differ from the default are changed one at a time to enable an understanding, if there is a difference in results, of what aspect it is that is driving that difference. Compared to the agreed default approach, Sherley's results relate to applications that change many if not all of the following aspects: the data used, the period considered, inclusion of covariates, and working in normal rather than log space which leads to difficulties in relating his models' estimates of the closure effect parameter to those based on the default approach. There may be a case for some of these changes, but a comparative exercise is not assisted when the associated requirement to build-a-bridge is not followed. This becomes particularly relevant when Sherley claims that two independent sets of analyses have iterated to the stage where they are in effective agreement about impacts of fishing on penguins **(**FISHERIES/2020/JUL/SWG-PEL/53REV). The comparisons shown in Figure 1 show that this is hardly the case, with some important differences in the values and especially the variances for δ estimates readily evident.

References (some are also quoted below)

de Moor, C. L. 2020a. A simple summary of the penguin island closure analysis. Document FISHERIES/2020/SEP/SWG-PEL/95. Pp. 1-7.

Penguin Island Closure Task Team (M.O. Bergh, D.S. Butterworth, K.L. Cochrane (chair), T.L. Morris, R.B. Sherley and H.Winker). 2015. Consolidated analyses produced in implementation of the approaches described in document MARAM/IWS/DEC15/PengD/P1. MARAM/IWS/DEC15/PengD/P2. Pp. 1-40.

Ross-Gillespie A and Butterworth, D.S. 2019(a). 2019 updated GLMM results for the South Coast penguin colony foraging data. Document MARAM/IWS/2019/PENG/P2. Pp. 1-12.

Ross-Gillespie A and Butterworth, D.S. 2019(b). Results for GLMM analyses of the South Coast penguin colony chick condition data. Document FISHERIES/2019/NOV/SWG-PEL/33. Pp. 1-5.

Ross-Gillespie A and Butterworth DS. 2020. Updated implementation of the Algorithm recommended by the Panel for the 2016 International Stock Assessment Workshop for assessing whether or not to continue with the penguin island closure experiment. Document FISHERIES/2020/JAN/SWG-PEL/09. Pp. 1–17.

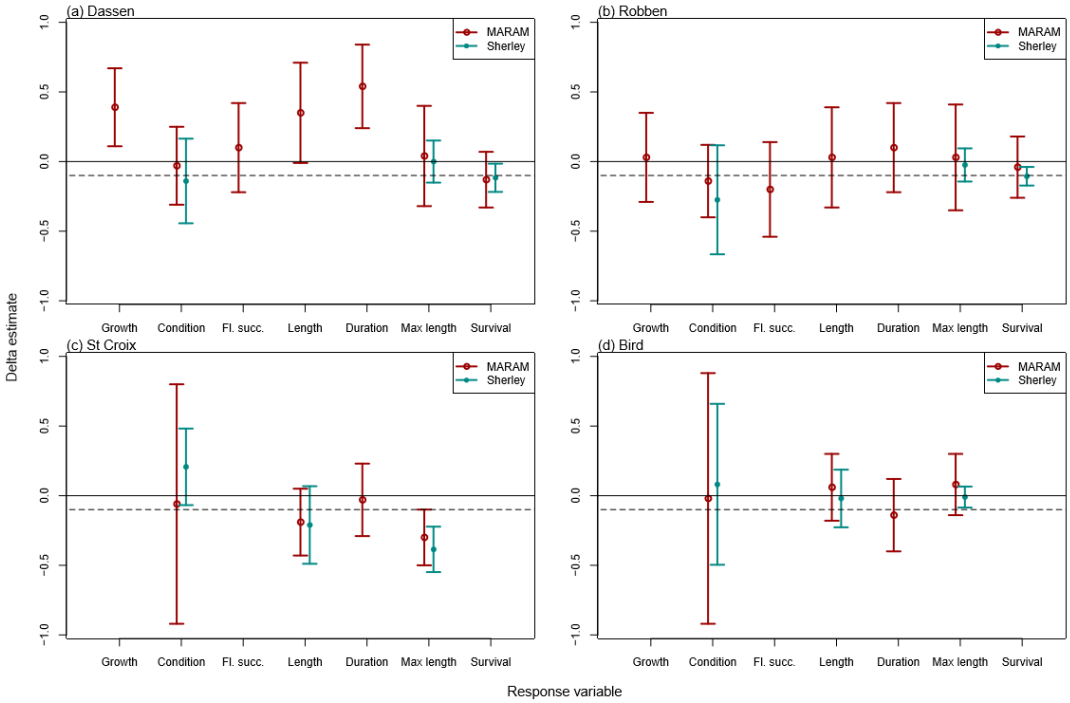

Figure 1: Zeh plots of the δ estimates and rough 95% confidence intervals are shown for the MARAM (aggregated data-based) and Sherley (individual data-based) models. The results for the MARAM models are taken from FISHERIES/2020/JAN/SWG-PEL/09rev for Robben and Dassen islands, from MARAM/IWS/2019/PENG/P2 for the foraging data for St Croix and Bird islands, and from FISHERIES/2019/NOV/SWG-PEL/33 for the chick condition data for St Croix and Bird islands. The values for the Sherley models have been derived from the last table of FISHERIES/2020/SEP/SWG-PEL/95 by use the following formula: $\delta = \ln(1 - p/100)$ where the p values are those reported in that last table as a simple approach to transform from normal to log-space to achieve improved comparability. The confidence intervals have been converted in a similar manner, and a rough standard error may be calculated as (max(CI)-min(CI))/4.. The Figure has been kindly provided by A. Ross-Gillespie.

Document FISHERIES/2020/OCT/SWG-PEL/103

A proposal for a basis to consider future island closures, taking account especially of the current results from the island closure experiment

D. S. Butterworth

On the biological basis for a proposal

The rationale that follows is based primarily on the results from the most recent application of the estimation model component of the algorithm recommended by the Panel for the 2016 International Stock Assessment Workshop, developed in collaboration with and endorsed by subsequent IWS Panels, to data obtained from the island closure experiment. For the island closure effect parameter δ, these

results are reported in FISHERIES/2020/JAN/SWG-PEL/09rev for Dassen and Robben islands and, based on the same default model, for St Croix and Bird islands in MARAM/IWS/2019/PENG/P2 and FISHERIES/2019/NOV/SWG-PEL/33 (see Figure 1). These applications utilise the most recent data made available to DEFF at the time of those analyses, under pre-agreed procedures. In terms of this algorithm, annually aggregated data are input to these analyses. This document does not consider other estimates of δ based on models using individual-penguin-data-based estimates. This follows a mathematical-statistical demonstration (see the Annex of FISHERIES/2020/AUG/SWG-PEL/82) that such estimates are unreliable, together with the current absence of any mathematical response to falsify that demonstration, as would be a scientifically required pre-requisite for their further consideration.

Document FISHERIES/2020/OCT/SWG-PEL/110

A Response to FISHERIES/2020/OCT/SWG-PEL/102

D. S. Butterworth

Note: For readers' ease, responses have, in the main, been inserted at appropriate points in the original document below in red and in *italics.*

Dassen Island

- ***Both sets of analyses currently show that the closure of Dassen Island to fishing will benefit penguin chick survival.***

While in principle chick survival data have relatively high potential information content for the purposes of the closure experiment, present results need to be considered in the context of a number of caveats:

- *A mathematical-statistical demonstration (see the Annex of PEL/82) indicates that individual data based estimates of closure effects (including those for chick survival) are unreliable; this demonstration has not been falsified, rendering consideration of such estimates scientifically unjustifiable at this time.*

Document FISHERIES/2020/OCT/SWG-PEL/111

Response to FISHERIES/2020/OCT/SWG-PEL/105

D. S. Butterworth

Note: For readers' ease, responses have, in the main, been inserted at appropriate points in the original document below in red and in *italics.*

Island closure results

Moreover, the updated Overall Closure Effect found 2 - 3 times more evidence for a positive closure effect than for no effect. Given that these methods have been scrutinised thoroughly in an international scientific journal (Sherley et al. 2018) *[Proc Roy Soc B: 285, 20172443]*

That scrutiny failed to detect the problems with the methodology used, which had been brought to the attention of the authors earlier. Attention to one aspect of those flaws was drawn by the 2019 Panel, and has been addressed with commendable thoroughness in PEL/53rev. The results reported in Table 1 of PEL/53rev indeed confirmed that the

approach used earlier had been flawed, and produced negatively biased estimates of standard error, confirming earlier criticisms relating to the reliability of that methodology.

Remaining flaws in the method raised earlier, and more specifically in PEL/08 and especially PEL/82, have yet to receive any cogent mathematical response.

The results of Sherley (2020) were subsequently presented at a Small Pelagic Scientific Working Group (SWG-PEL) convened by the Fisheries branch of DEFF in July 2020. Despite having addressed the recommendations of the IWS 2019, which favoured the individual based approach over the aggregated approach (provided certain further analyses were adopted - see above), the merits and flaws of both sets of analyses (aggregated versus individual-based approaches) were revisited. Comments by each party that represented these different methods were subsequently submitted and debated once again at a SWG-PEL meeting in September 2020. This has been a protracted process that has essentially revolved around a dichotomy in methods that have been revisited on numerous occasions. It is unlikely that any further progress will be made until there is a transparent and objective review process that assesses the most appropriate methodology. It is worth mentioning that regardless of the different approaches used, there was common ground in positive estimates of chick survival and/or fledging success for penguins from both Robben and Dassen islands when closed to fishing (de Moor 2020a) *[PEL/95]*. These were the only islands where these parameters, which reflect the demographic process most effectively when compared to all parameters used, were estimated during the experiment.

See comments above. The fledging success result for Dassen island is in the reverse direction. Furthermore, given a mathematical-statistical demonstration (see the Annex of PEL/82) that estimates based on the approach using individual data are unreliable, together with the current absence of any mathematical response to falsify that demonstration, as would be a scientifically required pre-requisite for their further consideration, those results cannot be considered further in a scientific forum such as the PWG. A distinction must be made here between an inexact science (such as fisheries science) where different interpretations of data can defensibly be argued, and an exact science (such as mathematics) where a proof must be respected, unless or until it might be falsified (potentially, for example, by some further review process).

Recommendation

In light of the above, we would like to re-iterate the following:

The results from Sherley (2020) *[PEL/53REV]* have adopted sound due diligence in terms of meeting peer-reviewed assessments and addressing important recommendations.

These results have addressed only one of the problems identified with the approach. Furthermore, they failed to include what are generally standard requirements for a proposed alternative estimator of building a bridge to allow proper comparison of results with those from an existing approach, and demonstration by simulation that the estimator provides satisfactory statistical performance.

Step 4 – Sherley further responses: Extracts from previous documents in blue:

HEADLINE: This issue covers old ground dealt with by the 2016 and 2019 IWS panels.

Sherley (2016) [MARAM/IWS/DEC16/Peng Clos/P4] "Linear mixed effect models were used because of "the flexibility they offer in modeling the within-group correlation often present in grouped data" (Pinhero & Bates 2000) because they can "account for dependencies within hierarchical groups through the introduction of random-effects" (Zuur et al. 2009). Their use is now commonplace in ecological analyses and they have been advocated for and used in fisheries management for some time (Venables & Dichmont 2004;… Punt et al. 2006; Thorson & Minto 2015; Thorson et al. 2016), including by members of MARAM (e.g. Brandão et al. 2004). To quote Venables & Dichmont (2004), writing over a decade ago: "One of the most important benefits of using mixed models is their capacity to 'borrow strength' from one part of the data to another, thus often providing a more realistic analysis of large fragmentary data sets, which are the norm in fisheries research". Or as Punt et al. (2006) put it: "there is value in using a mixed-effects approach to allow the years for which the dataset is large to 'provide support' for the years for which the data are sparse".

Dunn, Haddon, Parma and Punt (2016) [MARAM/IWS/DEC16/General/7]: "*MARAM/IWS/DEC16/Peng_Clos/P4 … suggests that analysing disaggregated data can lead to different estimates of the impact of closures on chick condition as well as more precise estimates. Table 2 explores some of the consequences of the impact of lower standard errors for these estimates and found them to be small*".

Die, Punt, Tiedemann, Waples and Wilberg (2019) [MARAM/IWS/2019/General/5]: "*Given the nature of the experiment, use of individual data is to be preferred*"...

"*results presented to the Workshop suggest that estimates of closure parameters using models fitted to aggregated and individual data had similar standard errors*".

HEADLINE: Comments to the effect that I haven't built a bridge between the two approaches are deliberately disingenuous.

Sherley (2020) [FISHERIES/2020/SEP/SWG-PEL/87]: "And it overlooks the point that I have already submitted documents attempting to provide this common ground on more than one occasion – see Sherley (2016) and Sherley and Winker (2019). In the interests of doing so again, please see [FISHERIES/2020/SEP/SWG-PEL/86] and Table 1 overleaf".

Table 1: Common ground between FISHERIES/2020/JUL/SWG-PEL/53REV (and associated analyses), and FISHERIES/2020/JAN/SWG-PEL/09 and FISHERIES/2019/NOV/SWG-PEL/27rev). Positive (+) and negative (−) signs are the direction of the effect: + here means positive for the penguins in the sense of Sherley (2020c; FISHERIES/2020/SEP/SWG-PEL/89, so indicates a negative δ value in FISHERIES/2020/JAN/SWG-PEL/09 and FISHERIES/2019/NOV/SWG-PEL/27rev. – means the reverse. The colour scheme denotes how meaningful each effect is and is based on the colour scheme used in Table 2 of Ross-Gillespie and Butterworth (2020; FISHERIES/2020/JAN/SWG-PEL/09. For entries in the R-G&B column: Green = "There is no evidence in the current data to support a biologically meaningful fishing effect, as the lower bound of the normal distribution, $\delta_{data}^{EM*} - 2(se)$, lies above the threshold". Blue = "The experiment needs to continue for more than 10 years before a biologically meaningful fishing effect is likely to be detected, if it is present". Orange = "The experiment needs to continue for 2 to 5 years before a biologically meaningful fishing effect is likely to be detected, if it is present". Red = There is evidence in the current data of a biologically meaningful fishing effect because $X > P_{min}$. For entries in the Sherley et al. column: Blue = There is no evidence at present for a closure effect either way in this dataset; < 95% of the posterior distribution has the same sign as the mean. Orange = There is a 95% probability of a closure effect in the direction indicated by the sign in this dataset. Red = There is a greater than 97.5% probability of a closure effect in the direction indicated by the sign in this dataset. Green is not used in the Sherley et al. column as there are no cases where there is evidence for a negative impact of the closures that is credibly different from zero at the 97.5% level.

Island	Data type	R-G&B	Sherley et al.	Sources
Dassen	Chick growth	−	+	[1,2]
	Chick Condition	+	+	[1,3]
	Fledging success	−	NA	[1]
	Path Length	−	NA*	[1]
	Trip Duration	−	NA*	[1]
	Max. Distance	−	+	[1,3]
	Chick survival	+	+	[1,3]
Robben	Chick growth	−	+	[1,2]
	Chick Condition	+	+	[1,2]
	Fledging success	+	NA	[1]
	Path Length	−	NA*	[1]
	Trip Duration	−	NA*	[1]
	Max. Distance	−	+	[1]
	Chick survival	+	+	[1,3]
Bird	Chick Condition	NA	−	[3]
	Path Length	−	+	[3,4]
	Trip Duration	−	NA*	[4]
	Max. Distance	−	+	[3,4]
St. Croix	Chick Condition	NA	−	[3]
	Path Length	+	+	[3,4]
	Trip Duration	+	NA*	[4]
	Max. Distance	+	+	[3,4]

Notes: NA = Not analysed by that group. R-G&B = Ross-Gillespie and Butterworth. *Not analysed due to issues of heterogeneity of variance (see R8 below). Sources: [1] Table 2b of Ross-Gillespie and Butterworth (2020; FISHERIES/2020/JAN/SWG-PEL/09). [2] The last time growth was analysed using a disaggregated approach was in Hagen et al. (2014; MARAM/IWS/DEC14/Peng/A3). This result is used here as the dataset has not been updated since. [3] Figure 1 of Sherley (2020c; FISHERIES/2020/SEP/SWG-PEL/89). [4] Based on option 2 (Closure model, 2018 foraging data) in Figure 1 of Ross-Gillespie and Butterworth (2019; FISHERIES/2019/NOV/SWG-PEL/27rev).

Sherley (2016) [MARAM/IWS/DEC16/Peng Clos/P4] "Finally, Doug's suggestion of calculating the standard deviation of the differences of the logged posterior distribution for open and closed years yielded precision estimates that were no longer an order of magnitude smaller than those of Ross-Gillespie and Butterworth (2016), but about 50% smaller (compare A and 2 in the table below). This was not greatly influenced by which data time period was used (compare 2 and 4, and both with A in the Table below)".

Source/Data range	Model type	Data type	Island	Effect size	SE/SD
(A) Ross-Gillespie and Butterworth 2016 from Table 6[1]	Log LMM(?)	Agg.	Dassen	−0.08	0.22
			Robben	−0.13	0.20
(B) Sherley 2016 (results from fit with biomass omitted)[2]	LMM (JAGS)	Disagg.	Dassen	0.02	0.02
			Robben	−0.11	0.03
(C) As Sherley 2016, but fit to 2004, 2008–2013 data[2]	LMM (JAGS)	Disagg.	Dassen	0.003	0.03
			Robben	−0.12	0.03
(1) Adding fortnight to the hierarchical random effect (Year/Month/Fortnight)[1 and 2 shown]	LMM (JAGS)	Disagg.	Dassen	−0.05	0.03[2] (0.10[1])
			Robben	−0.06	0.03[2] (0.09[1])
(2) As Sherley 2016, but fit to 2004, 2008–2013 data, SD of difference of logged joint posterior[1]	LMM (JAGS)	Disagg.	Dassen	0.003	0.10
			Robben	−0.12	0.09
(3) 2004, 2008–2013 aggregated data, no biomass, year random effect[1]	Log LMM (nlme)	Agg.	Dassen	−0.08	0.23
			Robben	−0.12	0.20
(4) As (2) but fit to 2008–2015 disaggregated data[1]	LMM (JAGS)	Disagg.	Dassen	0.02	0.09
			Robben	−0.11	0.08

Notes: 1. Results are in log space; 2. Results are in normal space; Agg. = aggregated data, meaning that the annual means are used; Disagg. = disaggregated data, meaning each of the original observations made in the field is used; nlme = model fit using the nlme library in R; JAGS = model fit using Bayesian inference and Just Another Gibbs Sampler (JAGS); LMM = linear mixed model; Log LMM linear mixed model on log transformed data.

Sherley and Winker (2019) [MARAM/IWS/2019/PENG/WP3]: "Butterworth and colleagues have argued repeatedly that it is preferable to fit to annual means rather than fit to disaggregated data at the level at which the observations were collected (e.g. from individual birds or nests) and use mixed models with random effect structures that account for hierarchical sources of variation implicit to the sampling design (e.g. Butterworth & Ross-Gillespie 2019). [This] is not consistent with modern approaches in either fisheries or ecological science (e.g. Hilborn and Liermann, 1998; Gelman and Hill, 2007; Pinheiro and Bates, 2009; Zuur et al., 2009; Thorson and Minto, 201[5]). Nevertheless, here I consider whether results from using the annual means remain consistent with the findings in Sherley et al. (2019, MARAM/IWS/DEC19/Peng/P4) for two cases that support a positive effect of the island closures experiment"

"Precision estimates range from 2.63% larger to 45% smaller with the disaggregated data than with the annual means (comparing to 2). All return significant and important closure effects at Robben Island, as demonstrated in Peng/P4, except for the grossly over-parameterized model in 1. Inference is otherwise unchanged by the model used"

"it is clear that, even with the substantial loss of statistical power that comes with the approach advocated by Butterworth and colleagues (e.g. Ross-Gillespie and Butterworth 2019), positive effects on penguins of the island closures are apparent".

Sherley (2020) [FISHERIES/2020/SEP/SWG-PEL/86]: "Although similar results have been offered to the SWG-PEL on at least two occasions in the past (Sherley 2016, Sherley and Winker 2019), it seems that it may be useful to offer some additional observations on comparisons of fitting comparable models to the annual means and the observation-level data for some of the island closures datasets. Here, given time constraints, I have focussed on the datasets analysed in [FISHERIES/2020/JUL/SWG-PEL/53REV]".

"Table 1 below offers empirical results based on fitting comparable models to some of the annually aggregated datasets and individual datasets for the Island Closures experiment… comparable models fit to the aggregated and individual data do not result in radically different precision estimates (as the panel put it, they have "*similar standard errors*" [Die et al. 2019]. And, contrary to the suggestion by Butterworth (2020), the maximal models (each M1 case in Table 1) from (Sherley 2020a) are indeed providing more precise estimates than the fits to the aggregated data in each case. They are not always larger, but they are (as would be expected from a model that has more statistical power) always more precise. Crucially, however, Table 1 also shows that **there is a > 95% probability that the closure effect is genuinely positive for the penguins for chick condition at Robben Island, chick survival at both Robben and Dassen Island and for Maximum foraging distance at St. Croix whether the aggregated data or individual data are used**".

Table 1: Comparisons of (Generalised) Linear-mixed effects models implemented in JAGS and fit to either the annually aggregated datasets or individual datasets. Where annual aggregated data are used, the model structure uses Year as a random effect [as per the structures in model 2 in Sherley and Winker (2019) and the structure used in Ross-Gillespie and Butterworth (2020)]. Where individual data results are reported, they are based on the best fitting model and/or the model with the maximal random effects (M1) in Sherley (2020). If only M1 is given, this was the best-fitting model in Sherley (2020a).

Dataset	Region	Data type	Island	Closure effect mean	Closure effect SD	95% Credible Interval	Probability of a positive effect for penguins
Chick Condition	Western Cape	Aggregated	DASSEN	0.018	0.046	−0.073–0.109	66.4%
		Individual (M1)[1]	DASSEN	0.033	0.033	−0.031–0.098	84.6%
		Aggregated	ROBBEN	0.081	0.046	−0.010–0.172	**96.0%**
		Individual (M1)	ROBBEN	0.066	0.038	−0.009–0.142	**95.6%**
Chick Survival	Western Cape	Aggregated	No interaction	0.270	0.092	0.094–0.461	**99.7%**
		Individual (M1)[2]	No interaction	0.380	0.087	0.211–0.550	**100%**
Max. distance	Eastern Cape	Aggregated*	ST CROIX	−0.345	0.099	−0.539–−0.150	**99.9%**
		Individual (M6)	ST CROIX	−0.322	0.057	−0.434–−0.212	**100%**
		Individual (M1)	ST CROIX	−0.322	0.056	−0.433–−0.213	**100%**
		Aggregated[3]	BIRD	0.057	0.103	−0.145–0.263	27.7%
		Individual (M6)[4]	BIRD	−0.008	0.039	−0.084–0.071	58.2%
		Individual (M1)[5]	BIRD	−0.009	0.039	−0.086–0.069	58.8%

Notes: 1. Random effect structure was: Island/Year/Month; form for model equation = eqn. 1 in Sherley et al. (2019). 2. Random effect structure was: Island/Year/NestID; form for model equation = eqn. A2 in Sherley (2020). 3. Brood mass is omitted from this model and was used in M6 and M1 in Sherley (2020). 4. Random effect structure was: BirdID; form for model equation = eqn. 2 in Sherley et al. (2019). 5. Random effect structure was: Island/Year/BirdID; form for model equation = eqn. 2 in Sherley et al. (2019).

HEADLINE: The comments that chick survival at Dassen is unreliable because the fledging success results trends in the opposite direction are built on a spurious comparison.

Sherley (2020) [FISHERIES/2020/SEP/SWG-PEL/85]: The chick survival dataset… spans 2008–2015 for Dassen Island. The fledging success dataset… spans 1995–1999 and then 2008–2015 (with a gap from 2000 to 2007) at Dassen Island. First, it is difficult to be confident in directly comparing data from the 1990s with data collected from 2008 onwards in this context because there is strong evidence that the ecosystem, the availability of forage fish resources to fisherman and predators, and penguin population dynamics have changed markedly over this timeframe (e.g. van der Lingen et al. 2005, Roy et al. 2007, Robinson et al. 2013, Crawford et al. 2019). We cannot be sure that the trend in the opposite direction is not a consequence of these differences in the state of the ecosystem.

Second, FISHERIES/2020/JAN/SWG-PEL/09 indicates that the experiment would need to continue for more than 10 years before a biologically meaningful fishing effect is likely to be detected for fledging success at Dassen Island. In other words, the fledging success effect at Dassen Island isn't meaningfully different from zero. On the other hand, the chick survival dataset for Dassen Island already provides evidence of a biologically meaningful fishing effect. Thus, the two do not offer equally strong opposing evidence…

Fourth, and most importantly, if we actually do a like for like… comparison between chick survival and fledging success, we find they are positively correlated with one another…

Robben Island 2001 to 2015: Pearson's product-moment correlation, $r = 0.981$, $t_{13} = 18.37$, $p < 0.001$.

Dassen Island 2008 to 2015: Pearson's product-moment correlation, r = 0.818, t_6 = 3.48, p < 0.013.

Robben Island 2008 to 2015: Pearson's product-moment correlation, r = 0.93, t_6 = 6.65, p < 0.001.

Dassen Island 2008 to 2015: Pearson's product-moment correlation, r = 0.817, t_6 = 3.47, p < 0.013.

Thus, Butterworth's concerns about negative correlation effects are unfounded.

References:

Brandão A., Butterworth D.S., Johnston S.J., Glazer J.P. 2004. Using a GLMM to estimate the somatic growth rate trend for male South African west coast rock lobster. Fisheries Research 70: 399–349.

Butterworth, D. S. & Ross-Gillespie, A. (2019) Is pseudo-replication biasing results from analyses from the island closure experiment which model individual penguin responses directly? Department of Environment, Forestry and Fisheries Report: MARAM/IWS/2019/PENG/P5. Pp. 1–10.

Crawford RJM, Sydeman WJ, Thompson SA, Sherley RB and Makhado AB. 2019. Food habits of an endangered seabird indicate recent poor forage fish availability. ICES Journal of Marine Science 76: 1344–1352.

Gelman, A. & Hill, J. (2007) *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge, UK.

Hilborn, R. & Liermann, M. (1998) Standing on the shoulders of giants: Learning from experience in fisheries. *Reviews in Fish Biology and Fisheries* 8: 273–283.

Pinhero J.C. & Bates D.M. 2000. *Mixed-effects models in S and S-Plus*. Springer Verlag, New York.

Pinheiro, J. & Bates, D. (2009) *Mixed-Effects Models in S and S-PLUS*. Springer, New York, NY.

Punt A.E., Hobday D.K. & Flint R. 2006. Bayesian hierarchical modelling of maturity-at-length for rock lobsters, Jasus edwardsii, off Victoria, Australia. Marine and Freshwater Research 57: 503–511.

Robinson WML, Butterworth DS and Plag.nyi .E. 2015. Quantifying the projected impact of the South African sardine fishery on the Robben Island penguin colony. ICES Journal of Marine Science 72: 1822–1833.

Ross-Gillespie A. & Butterworth D. S. 2016. Penguin power analyses using the approach recommended by the international panel: methods and results. FISHERIES/2016/NOV/SWGPEL/ Peng/01. Pp. 1–31.

Ross-Gillespie A and Butterworth DS. 2019. 2019 Updated GLMM results for the South Coast penguin colony foraging data. Department of Environment, Forestry and Fisheries Report: FISHERIES/2019/NOV/SWG-PEL/27rev. Pp. 1–12.

Roy C, van der Lingen CD, Coetzee JC and Lutjeharms JRE. 2007. Abrupt environmental shift associated with changes in the distribution of Cape anchovy *Engraulis encrasicolus* spawners in the southern Benguela. African Journal of Marine Science 29: 309–319.

Sherley RB. 2016. Additional analysis suggested in response to differences in variance estimates between Sherley (2016) and Ross-Gillespie & Butterworth (2016). Department of Environment, Forestry and Fisheries Report: MARAM/IWS/DEC16/Peng Clos/P4. Pp.

1–4.

Sherley RB and Winker H. 2019. Some observations on comparisons of fitting to the annual means and the observation-level data for the cases in MARAM/IWS/DEC19/Peng/P4 that support a positive effect of the island closures experiment on African penguins. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/2019/PENG/WP3. Pp. 1–5.

Thorson J.T. & Minto C. 2015. Mixed effects: a unifying framework for statistical modelling in fisheries biology. ICES Journal of Marine Science 72: 1245–1256.

Thorson J.T., Rindorf A., Gao J., Hanselman D.H. & Winker H. 2016. Density-dependent changes in effective area occupied for sea-bottom-associated marine fishes. Proceedings of the Royal Society B: Biological Sciences 283: 20161853.

van der Lingen CD, Coetzee JC, Demarcq H, Drapeau L, Fairweather TP and Hutchings L. 2005. An eastward shift in the distribution of southern Benguela sardine. GLOBEC
International Newsletter 11: 17–22.

Venables W.N. & Dichmont C.M. 2004. GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. Fisheries Research 70: 319–337.

Zuur A.F., Ieno E.N., Walker N.J., Saveliev A.A. & Smith G.M. 2009. Mixed effects models and extensions in Ecology with R. Springer, New York.

EXCHANGE 2: BERGH/SHERLEY

Step 1 – Bergh assertion: Extracts (in red) from:

Document FISHERIES/2020/AUG/SWG-PEL/84 "Comments on Revisiting the key results in MARAM/IWS/2019/PENG/P4 in light of the 2019 Panel recommendations (FISHERIES/2020/JUL/SWG-PEL/53REV) by Richard B. Sherley". By Mike Bergh

The mathematical proof that is the subject of Q1 first appeared in Document FISHERIES/2020/AUG/SWG-PEL/82 and so was not referenced in any way in the Step 1 comments of Document FISHERIES/2020/AUG/SWG-PEL/84. The following comment has some general relevance to the broader context of the mathematical proof in FISHERIES/2020/AUG/SWG-PEL/82:

"A debate that has been circulating for a number of years is the merit of the use of individual bird data versus standardised aggregated data. This debate has not been resolved in FISHERIES/2020/JUL/SWG-PEL/***53REV***. This dichotomy is one of the important differences between the results reported in FISHERIES/2020/JUL/SWG-PEL/***53REV***, and those reported in FISHERIES/2020/JAN/SWG-***PEL/09*** (and FISHERIES/2019/NOV/SWG-***PEL/27rev***). Further deliberations on the management actions required for island closures need to see this debate resolved".

Step 2 – Sherley responses: Extracts (in blue) from:

Document FISHERIES/2020/SEP/SWG-PEL/87 "A reply to Bergh: FISHERIES/2020/AUG/SWG-PEL/84" by Richard B. Sherley

Document FISHERIES/2020/AUG/SWG-PEL/83 "Some comments on FISHERIES/2020/JAN/SWG-PEL/08" by Richard B. Sherley

And

FISHERIES/2020/SEP/SWG-PEL/85 "A response to Butterworth: FISHERIES/2020/AUG/SWG-PEL/82" by Richard B. Sherley

HEADLINE: This issue covers old ground dealt with by the 2016 and 2019 IWS panels.

Extract from R19, page 20 of FISHERIES/2020/SEP/SWG-PEL/85: "the 2019 IWS panel have already given an opinion… "*Given the nature of the experiment, use of individual data is to be preferred… Die et al. (2019)*".

HEADLINE: The peer-reviewed literature is clear that mixed-effects models can and should be used in this situation.

Extract from R1, pages 1 and 2 of FISHERIES/2020/SEP/SWG-PEL/87: This comment rather overlooks the peer-reviewed scientific literature on this issue (see comments in Sherley 2020a, FISHERIES/2020/AUG/SWGPEL/83).

Extract from Comment 3, page 3 of FISHERIES/2020/SEP/SWG-PEL/83: Simulation studies have, however, demonstrated that linear mixed-effects models (LMMs) can be used in these circumstances, when random effects are chosen based on the known sampling structure in the data (Silk et al. 2020) and when model selection methods are used to choose the random effect structure (Matuschek et al. 2017). As Matuschek et al. (2017) put it: "Our simulations have shown that determining a parsimonious model with a standard model selection criterion is a defensible choice to find this middle ground between Type I error rate and power". Again, this is why I understood the panel to make the following recommendation in their 2019 report: "Model selection methods should be applied to select an appropriate random effects structure" and why [FISHERIES/2020/JUL/SWG-PEL/53REV] has been presented to address that recommendation.

Extract from R1, page 2 and 3 of FISHERIES/2020/SEP/SWG-PEL/85: "[The] use of parsimonious mixed models… improves the balance between a Type I error and statistical power (e.g. Matuschek et al. 2017, Bates et al. 2018, Silk et al. 2020), and so allows exactly this kind of one-step approach in FISHERIES/2020/JUL/SWGPEL/ 53REV... mixed-effects models have even been advocated for and used in fisheries management for more than a decade (Venables & Dichmont 2004, Punt et al. 2006, Thorson & Minto 2015, Thorson et al. 2016), including by Butterworth himself (Brandão et al. 2004)".

References

Bates D, Kliegl R, Vasishth S and Baayen 2018. Parsimonious Mixed Models. arXiv:1506.04967v2.

Brandão A, Butterworth DS, Johnston SJ, Glazer JP. 2004. Using a GLMM to estimate the somatic growth rate trend for male South African west coast rock lobster. Fisheries Research 70: 399–349.

Die DJ, Punt AE, Tiedemann R, Waples R and Wilberg MJ. 2019. International Review Panel Report for the 2019 International Fisheries Stock Assessment Workshop, 2–5 December 2019, UCT. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/2019/General/5. Pp. 1–18.

Matuschek H, Kliegl R, Vasishth S, Baayen H and Bates D. 2017. Balancing Type I error and power in linear mixed models. Journal of Memory and Language 94: 305–315.

Punt AE, Hobday DK and Flint R. 2006. Bayesian hierarchical modelling of maturity-at-length for rock lobsters, Jasus edwardsii, off Victoria, Australia. Marine and Freshwater Research 57: 503–511.

Sherley RB. 2020a. Some comments on FISHERIES/2020/JAN/SWG-PEL/08. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/AUG/SWG-PEL/83. Pp. 1–5.

Silk MJ, Harrison XA and Hodgson DJ. 2020. Perils and pitfalls of mixed-effects regression models in biology. PeerJ 8: e9522.

Thorson JT and Minto C. 2015. Mixed effects: a unifying framework for statistical modelling in fisheries biology. ICES Journal of Marine Science 72: 1245–1256.

Thorson JT, Rindorf A, Gao J, Hanselman DH and Winker H. 2016. Density-dependent changes in effective area occupied for sea-bottom-associated marine fishes. Proceedings of the Royal Society B: Biological Sciences 283: 20161853.

Venables WN and Dichmont CM 2004. GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. Fisheries Research 70: 319–337.

Step 3 – Bergh responses to responses (Extracts from previous documents)

('Mathematical proof question')

Document FISHERIES/2020/OCT/SWG-PEL/107

Comments on FISHERIES/2020/SEP/SWG-PEL/87

by

Mike Bergh
20 September 2020

Document FISHERIES/2020/OCT/SWG-PEL/113

Comments on FISHERIES/2020/SEP/SWG-PEL/105REV, "Recommendations for island closures around African Penguin colonies".

by

Mike Bergh
19 October 2020

"For reasons described here, the results of analyses of the island closure experiment as reported in **FISHERIES/2020/JUL/SWG-PEL/53REV** and **MARAM/IWS/2019/PENG/P4** should not form part of the deliberations about island closures.**"**

"There are two sets of results on the island closure experiment that are under discussion, those reported in **FISHERIES/2020/JUL/SWG-PEL/53REV** and those reported in **FISHERIES/2020/SEP/SWG-PEL/97REV**. **FISHERIES/2020/AUG/SWG-PEL/84** pointed out that the analytical methods underpinning **FISHERIES/2020/JUL/SWG-PEL/53REV** have not following the full set of analytical recommendations made by the IWS panel dating back to 2014, which included recommendations to carry out simulations to address biases that arise from the application of GLM techniques for the specific situation pertaining to the island closure experiment. Those in **FISHERIES/2020/SEP/SWG-PEL/97REV** have. In addition, the methods in **FISHERIES/2020/JUL/SWG-PEL/53REV** are based on the use of individual bird data which results in standard errors (se's) that are at times considerably smaller than those reported in **FISHERIES/2020/SEP/SWG-PEL/97REV**, the latter being based on the use of year and island aggregated data. A mathematical proof presented in an annex to **FISHERIES/2020/AUG/SWG-PEL/82** shows that unbiased se's of island closure effects cannot be smaller than those produced from analyses based on aggregated data. The se's of island closure effects reported in **FISHERIES/2020/JUL/SWG-PEL/53REV** are in some cases smaller than those that are based on the use of aggregated data, indicating that some biases have arisen in calculating these standard error estimates. The soundness of the mathematical proof in an annex to **FISHERIES/2020/AUG/SWG-PEL/82** has not been questioned or shown to be questionable. **Until such time** that the proof might be refuted, the results based on the work contained in **FISHERIES/2020/JUL/SWG-PEL/53REV** must

be disregarded, and conclusions/recommendations can only be based on the results reported in **FISHERIES/2020/SEP/SWG-PEL/97REV"**

2.6 Section headed "Recommendation"

1. This section cites results in Sherley (2020) which it is argued elsewhere in this document should be disregarded until and if the mathematical proof in an annex to **FISHERIES/2020/AUG/SWG-PEL/82** might be refuted.

Document FISHERIES/2020/OCT/SWG-PEL/107

Comments on FISHERIES/2020/SEP/SWG-PEL/87
by
Mike Bergh
20 September 2020

R6. Impact of the Use of Different Random Effects. Comment 6 in FISHERIES/2020/AUG/SWG-PEL/84 notes that the use of different random effects has a large impact on the standard error of certain of the closure effects reported in FISHERIES/2020/JUL/SWG-PEL/53REV, and argues that the reason for these differences needs an explanation. FISHERIES/2020/SEP/SWG-PEL/87 responds that
"As outlined in FISHERIES/2020/JUL/SWG-PEL/53REV, one possibility is that this is an issue with having Island in both the fixed and random components of the model. M1 in all cases in the maximal model (the most complex possible random effect structure); maximal models are "generally wasteful and costly in terms of statistical power for testing hypotheses" (Stroup 2012, pg. 185) and maximal models – even when they converge – can result in overparameterization that leads to uninterpretable models (Bates et al. 2018). The maximal model may actually trade-off power for some conservatism beyond the nominal Type I error rate, even in cases where the maximal model matches the generating process exactly (Matuschek et al. 2017)"

This is speculative and should be backed up by numerical results which would probably need to be derived from data analyses as well as simulation studies. It would also need to address the annex of FISHERIES/2020/AUG/SWG-PEL/82 which demonstrates mathematically that "**as far as estimates of the precision of the closure effect in the island closure experiment is concerned, there is no advantage to be gained from analysing the individual data each year rather than an aggregate value such as their means**."

Document FISHERIES/2020/SEP/SWG-PEL/99

Summary comments on the Penguin Island Closure Experiment
By
Mike Bergh
20 September 2020

7. There are a number of outstanding technical issues with the methods and results in FISHERIES/2020/JUL/SWG-PEL/***53REV*** that have not been answered.

These unresolved matters weigh heavily on the scientific deliberations which are now ongoing, and force participants to take a position on one or the other set of results, since both cannot be reliable.

In addition, there is now a mathematical proof (see the annex of FISHERIES/2020/AUG/SWG-PEL/82) that the standard error of the island closure effect achieved using aggregated bird data cannot be improved upon by using data from individual bird data. In the absence of any submission that contradicts this proof, there is no reason to question the correctness of this proof. It follows that any results that provide estimates with standard errors that are smaller than the s.e. achieved using aggregated bird data must either be in error, or be negatively biased (presumably because the random effect used to adjust for pseudo-replication in the case of analyses using individual bird data is failing to account fully for this pseudo-replication). These results are therefore producing a misleading impression of the precision of estimates of the island closure effect.

Another consideration is that since decisions on the Penguin Island Closure Experiment must be made this year, it is likely that, given the complexities associated with the statistical analyses and the time it will take to resolve these, decisions will have to be made on the basis of results that have been tabled thus far. Given the problems that are pointed out above regarding the results reported in FISHERIES/2020/JUL/SWG- PEL/***53REV***, it is ill-advised to allow these to inform decisions that must be made this year.

Step 4 – Sherley further responses: Extracts from previous documents in blue:

HEADLINE: This issue of the SE or precision covers old ground dealt with by the 2016 and 2019 IWS panels.

Dunn, Haddon, Parma and Punt (2016) [MARAM/IWS/DEC16/General/7]: "*MARAM/IWS/DEC16/Peng_Clos/P4 … suggests that analysing disaggregated data can lead to different estimates of the impact of closures on chick condition as well as more precise estimates. Table 2 explores some of the consequences of the impact of lower standard errors for these estimates and found them to be small*".

Die, Punt, Tiedemann, Waples and Wilberg (2019) [MARAM/IWS/2019/General/5]: "*Given the nature of the experiment, use of individual data is to be preferred*"...

"*results presented to the Workshop suggest that estimates of closure parameters using models fitted to aggregated and individual data had similar standard errors*".

HEADLINE: The above suggests that the approach in FISHERIES/2020/JUL/SWG-PEL/53REV has not followed "the full set of analytical recommendations made by the IWS panel dating back to 2014", but all the recommendations mentioned in FISHERIES/2020/AUG/SWG-PEL/84 are those the panel made for the procedure to conduct a power analysis. FISHERIES/2020/JUL/SWG-PEL/53REV is not conducting a power analysis.

Sherley (2020) [FISHERIES/2020/SEP/SWG-PEL/87]: "The top of page 3 of Dunn et al. (2015) states: "In relation to next steps for a power analysis to evaluate closure effects on penguins:" and points 3 and 4 then immediately follow and pertain to conducting a power analysis, as does point 15 on page 5. Table 1 still pertains to a power analysis ...

"Appendix A of Dunn et al. (2015), "OUTLINE OF THE PROCESS OF CONDUCTING A POWER ANALYSIS FOR AFRICAN PENGUINS", again pertains to conducting a power analysis. FISHERIES/2020/JUL/SWG-PEL/53REV is not conducting a power analysis".

"Again, the Table 2 above is labelled "Summary of the power analysis procedure" and so pertains to the process for conducting a power analysis. It isn't at all clear why it ought to be necessary to go through a process designed for a power analysis when simply using a (generalised) linear mixed model (GLMM) to determine whether a difference between two means is credibly (or statistically) different from zero. Indeed, as FISHERIES/2020/JUL/SWG-PEL/53REV is not, and has never been, conducting a power analysis, suggesting that it is necessary to follow a procedure designed for when conducting a power analysis (in a specific context) before the results can be accepted is tantamount to suggesting that anyone, anywhere must follow the procedure in Table 2 before using a GLM(M) in any analysis. It further suggests that all papers using GLM(M)s need to have their analysis reconsidered until such time that they apply the procedure in Table 2. Surely, that is not what is being suggested?"

HEADLINE: Describing evidence from peer-reviewed papers in international journals as "speculative" is concerning and ignores a key unresolved question from the 2019 IWS, which is whether 'Island' should be included in both the fixed and random components of the model when the random effects used are nested, not crossed:

Extract from R6, page 8 of FISHERIES/2020/SEP/SWG-PEL/85: "there is still the question of whether Island should simultaneously be included in both the fixed and random components of these models and whether Island, which only has two levels, should be included in the random effect structure at all…

M1 in the table above is the maximal model (the most complex possible random effect structure); maximal models are "generally wasteful and costly in terms of statistical power for testing hypotheses" (Stroup 2012, pg. 185) and maximal models – even when they converge – can result in overparameterization that leads to uninterpretable models (Bates et al. 2018).

Furthermore, the maximal model may actually trade-off power for some conservatism beyond the nominal Type I error rate, even in cases where the maximal model matches the generating process exactly (Matuschek et al. 2017). Nevertheless, it presents a > 96% probability (given the data and model structure) of a closure effect at Robben Island. Ignoring this, particularly given that an independent analysis by Ross-Gillespie and Butterworth (FISHERIES/2020/JAN/SWG-PEL/09) concluded that there was "a biologically meaningful fishing effect" on chick condition at Robben Island, using the 2004 to 2018 aggregated data would certainly risk making a Type II error about the impact of the closure".

References:

Bates D, Kliegl R, Vasishth S and Baayen 2018. Parsimonious Mixed Models. arXiv:1506.04967v2.

Dunn A, Haddon M, Parma AM and Punt AE. 2015. INTERNATIONAL REVIEW PANEL REPORT FOR THE 2015 INTERNATIONAL FISHERIES STOCK ASSESSMENT WORKSHOP 30 November–4 December 2015, UCT. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/DEC15/General/8.

Dunn A, Haddon M, Parma AM and Punt AE. 2016. INTERNATIONAL REVIEW PANEL REPORT FOR THE 2016 INTERNATIONAL FISHERIES STOCK ASSESSMENT WORKSHOP 28 November–2 December 2016, UCT. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/DEC16/General/7.

Matuschek H, Kliegl R, Vasishth S, Baayen H and Bates D. 2017. Balancing Type I error and power in linear mixed models. Journal of Memory and Language 94: 305–315.

Stroup WW. 2012. Generalized linear mixed models: Modern concepts, methods and applications. Boca Raton: