Question Q3: 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 Q3.

QUESTION Q3: It is acknowledged that the limited degrees of freedom available with the aggregated data approach hampers efforts to obtain precise closure (fishing effect) estimates from the island closure experiment. In 2016, a power analysis approach was finalised to advise on the period needed for the closure experiment to have to continue before being able to provide biologically meaningful results; this was based on an aggregated data approach. In 2019, the Panel recommended that "given the nature of the experiment, use of individual data is to be preferred" (first bullet, page 10, MARAM/IWS/DEC19/General/5). Does it therefore follow that the aggregated approach should not be used to provide results on which management advice for island closures is to be based?

EXCHANGE 1: Sherley/Butterworth

Step 1 – Sherley assertion:

In light of the 2019 panel recommendation, the use of the aggregated approach no longer represents the best-available science on which management advice for island closures should be based: Concerns about the small number of residual degrees of freedom (dof) in the models presented by the MARAM group have been repeatedly presented to previous IWS, dating back to at least 2014 when Bergh et al. [MARAM/IWS/DEC14/PENG/A2] wrote: "*There are two main concerns with the structure of the GLMs performed by Robinson (2013). Firstly, group means are modelled for each year, hence biologically important variation is averaged out… Secondly, the models contain too many predictors for the number of outcomes available, and therefore would have little power to detect significant results*", "*it is not clear whether the problems with small degrees of freedom are addressed by the use of year as a random effect*" and "*as such, it would be concerning if the analyses of Robinson and Butterworth (2014a, 2014b) were used for management-decision making*".

More recently, it has been agreed in SWG-PEL documents that the aggregated approach is hampered by a low number of degrees of freedom [dof] and that this is only "*ameliorated somewhat*" if the year factor is treated as a random effect [see statement by Butterworth on page 2 of FISHERIES/2020/JAN/SWG-PEL/08 and agreement by Sherley at the bottom of page 2 in FISHERIES/2020/AUG/SWG-PEL/83].

The essential point is "if the models have a low number of dof relative to the number of parameters estimated, this lowers their explanatory power" (Crawley 2013). Rules of thumb for the minimum number of data points per parameter range from the nonconservative 3 data points per parameter (Crawley 2013), through 10 data points per parameter (Draper and Smith 1998) to the very conservative mn datapoints (where m = the number of data points required to determine a univariate regression line with sufficient precision and n = the number of parameters to be estimated; Good and Hardin 2006). Many of the models in Ross-Gillespie and Butterworth (2020) fall closer to the less conservative (3 data points per parameter) end of the scale, thus are just about statistically acceptable (now that the experiment has been running for 12 years). In other words,

models applied to the aggregated data potentially lack statistical power and we risk making a Type II error… [This] is why I understood the panel to make the following recommendation in their 2019 report: "*Given the nature of the experiment, use of individual data is to be preferred. However, this is only the case if an appropriate random effects structure is chosen*". [Extract from page 2 and 3 of FISHERIES/2020/AUG/SWG-PEL/83].

Step 2 – Butterworth response:

Direct responses to the assertion made in Step 1 in relation to QUESTION Q3: Step 1

Paragraph 1
This paragraph relates to concerns about the (aggregated data) GLMs in Robinson (2013). These concerns have subsequently been addressed, *inter alia* by robustifying the estimator by treating the year factor as a random rather than a fixed effect. This approach was accepted by the Panel in its 2015 and 2016 meetings, and the approach itself involves simulation testing of the estimator, the results from which are used to adjust for any bias though this is usually small (see e.g. FISHERIES/2020/JAN/SWG-PEL/09rev). (Note that although the ultimate purpose of discussions at those Panel meetings was to develop a power analysis, the results of such a power analysis would apply only to the estimator to be used in practice, so adoption of that specific power analysis approach also implied acceptance of the aggregated data estimator which underlay it.) The results from using that estimator for the standard errors of the closure effect estimates shown in Table 2 of FISHERIES/2020/JAN/SWG-PEL/09rev confirm that meaningful results may be obtained from the annually aggregated data (see also Figure 1 below). Subsequent to the 2016 Panel meeting, there has been no other indication expressed that the concerns referenced in Paragraph 1 had not been adequately addressed to general satisfaction.

Paragraph 2
This paragraph is written in a way that implies that because the aggregated approach is agreed to be "hampered by a low number of degrees of freedom", it is "unacceptable" in some sense. It also implies that this fact has been realised "only recently". Rather, this has been appreciated since 2014 at least, including at the 2015 and 2016 meetings referenced above at which the approach was generally considered acceptable. The "hampering" comment was intended to imply only that this is an unfortunate (but unavoidable) fact of life (the assertion that the use of individual data can avoid this and improve the situation is refuted in comments provided on Q1 in Step 1 and Step 3).

Paragraph 3
Ultimately the acceptability or otherwise of the use of an estimator does not depend on the ratio of degrees of freedom to number of estimable parameters *per se*, but on the standard errors achieved on application and satisfactory performance under simulation testing. As per the paragraph two above, both have proved to be satisfactory in this instance.

Does this question have practical import?

The 2019 Panel commented, *inter alia*, that "*Results presented to the Workshop suggest that estimates of closure parameters using models fitted to aggregated and individual data had similar standard errors*". This suggests that in making their recommendations they may have had the

impression that Q3 is not a question of particular consequence (though they did add the qualifier *"(at least for the effects of closure parameter for the South Coast colonies evaluated)"*).

The comparison of results from the two approaches shown in Figure 1 below does, however, indicate that the impression gained by the Panel was in serious error. In particular, the precision of the closure effect estimates from chick survival data at Robben and Dassen islands, and for chick condition factor at St Croix, differ substantially, and in ways that would clearly lead to different implications as to the benefits or otherwise of closures to penguins.

The antithesis of Q1

Q3 is essentially the antithesis of Q1, so that arguments offered in support of the assertion of Q1 in Steps 1 and 3 generally also apply in negating the contention of Q3. There is therefore no need to duplicate them all here, but it is perhaps useful to summarise briefly a few of the key points made there.

- A mathematical-statistical demonstration indicates that the independence of individual data from one year to the next (e.g. because there is no information linking a penguin sampled in one year to a penguin sampled the next) means that use of such data cannot improve the precision of estimates of closure effects compared to those from the corresponding annually aggregated data. No falsification of this demonstration has been offered.

- To the extent that individual-data -based analyses appear to provide more precise estimates of closure effects than those based on corresponding annually aggregated data, this simply suggests that the random effects models used has failed to account adequately for the effects of pseudo-replication; such approaches cannot be guaranteed to account fully for pseudo-replication effects. Note the 2019 Panel statement: "*it is a working hypothesis that including random effects chosen using model selection methods will appropriately account for the pseudo-replication. This is a working hypothesis because it can never be guaranteed that including random effects will fully address pseudo-replication*".

- The individual-data-based estimators applied in this situation have not been simulated tested to confirm the claims made that they can achieve improved precision compared to their aggregated data counterparts. Such simulation testing is normally a routine requirement in circumstances where the statistical properties of the estimator as in contention.

- Results from generic simulation studies are indicative only, not conclusive. First, there needs to be a demonstration that the general structure of such studies covers the structure of the estimation problem in question; and even given that, there would still probably be a requirement to condition the simulation on the actual data prior to considering performance claims to be confirmed.

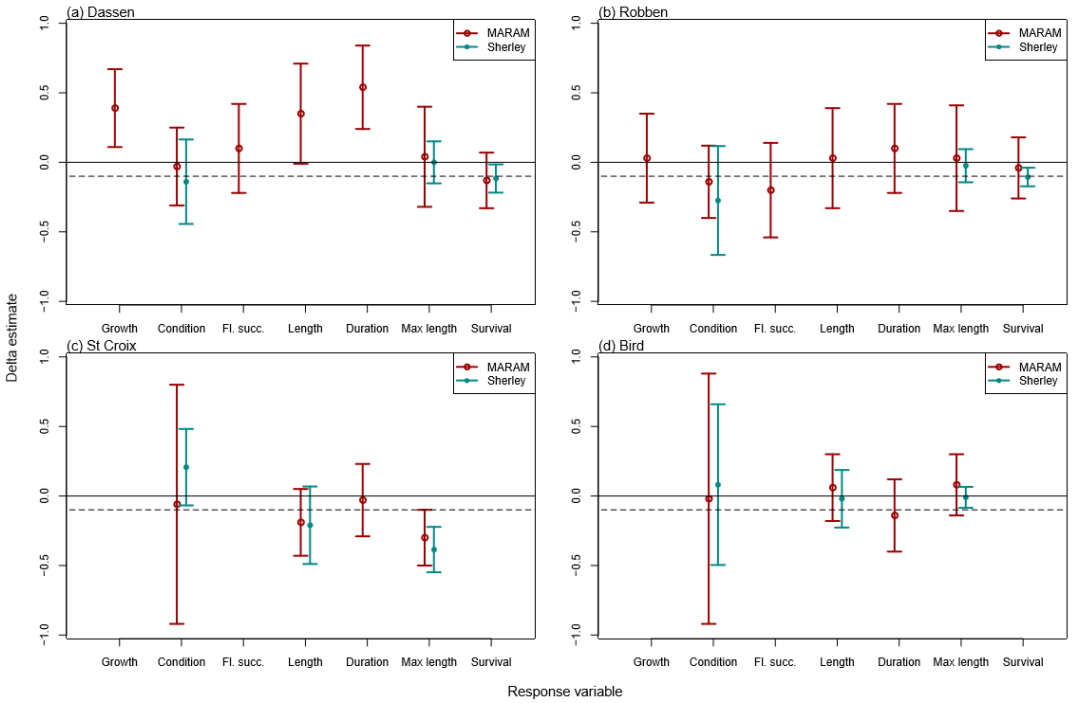

Figure 1: Zeh plots of the δ estimates and rough 95% confidence intervals are shown for the MARAM (aggregated data-based) and Sherley (individual data-based) models. The results for the MARAM models are taken from FISHERIES/2020/JAN/SWG-PEL/09rev for Robben and Dassen islands, from MARAM/IWS/2019/PENG/P2 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.
[Figure taken from FISHERIES/2020/SEP/SWG-PEL/96rev]

EXCHANGE 2: Sherley/Bergh

Step 1 – Sherley assertion: Same as above (Page 1)

Step 2 – Bergh comment: Extracts from previous documents

This question is closely related to Q1. Whereas Q1 asserts that the analytical results based on use of individual data should not be considered as providing reliable estimates of the impact of island closures, Q3 asserts that the analytical results based on use of aggregated data does not represent the best available science, and should therefore nor be considered. As a result all comments that were submitted in relation to Q1 Step 1 by Bergh are relevant here.