Some comments on FISHERIES/2020/JAN/SWG-PEL/08

Richard B. Sherley[1,2]

1. Centre for Ecology and Conservation, College of Life and Environmental Sciences, University of Exeter, Penryn Campus, Cornwall, TR10 9FE, United Kingdom.
2. FitzPatrick Institute of African Ornithology, DST-NRF Centre of Excellence, University of Cape Town, Rondebosch 7701, South Africa.

Email: r.sherley@exeter.ac.uk

FISHERIES/2020/JAN/SWG-PEL/08 (Butterworth 2020) was submitted as a PDF document. Thus, to facilitate responding, I have extracted sections of text from FISHERIES/2020/JAN/SWG-PEL/08, marked with "**E:**" at the start of each extract, and inserted them here in blue text. My comments are then in black directly below, marked with "**C:**" at the start of each section.

E: 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.

C: The statement above mis-specifies the statement made by the Panel at the 2019 International Workshop (Die et al. 2019). The panel said: *"Random effects models are used to account for such "latent" covariates in designed experiments. However, in the 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. This is a working hypothesis because it can never be guaranteed that including random effects will fully address pseudo-replication when the "true" sampling design is not known (i.e., how individuals are selected)".*

Contrary to what Butterworth's (2020) implies, the panel did not say that including some (the implication being "any") random effect will necessarily account for pseudo-replication. It is well recognised that random effects should be chosen carefully to ensure that they (as far as is practically possible) account for latent covariates common to individuals and the hierarchical nature of the sampling design in both natural and designed experiments (Zuur et al. 2009, Crawley 2013, Harrison et al. 2018, Arnqvist 2020, Silk et al. 2020). The fact that the SEs change depending on which random effects structure has been used is not surprising, as not every possible choice of a random effects structure will account for pseudo-replication equally well. But this was never claimed to be the case.

Nor do these few extracted SE estimates demonstrate that the working hypothesis is not supported (as claimed by Butterworth 2020). First, we need to look again at what the panel said in their report: the working hypothesis was that "*including random effects chosen using model selection methods will appropriately account for the pseudo-replication*" (Die et al. 2019). At the time that they were writing, the panel and the SWG-PEL had not been presented with the results of FISHERIES/2020/JUL/SWG-PEL/53REV (Sherley 2020), which do indeed confirm that the inference on the value of the island closures is essentially unchanged regardless of which random effect structure was used. Moreover, the results in Sherley (2020) – crucially – indicate that the closure effects based on the random effect structure chosen using model selection methods recommended by the panel (Die et al. 2019) were actually slightly stronger than those originally reported in MARAM/IWS/2019/PENG/P4 (Sherley et al. 2019).

Second, in three of the four cases provided as examples by Butterworth (2020), the random effects models applied to the individual-level data returned a larger SE estimate than the model applied to the aggregated data. If, in that case, the concern is one of making a Type I error (because the estimates from the individual-level data are alleged to be overly precise), then one would actually need to conclude on the bases of the examples provided by Butterworth (2020) that the models using the aggregated data are in fact at equal or greater risk of making that error.

It should also be noted that the panel had been presented with the same results quoted by Butterworth (2020) during the workshop in MARAM/IWS/2019/PENG/WP3 (Sherley and Winker 2019), prior to them writing their report. They then concluded in their report that *"results presented to the Workshop suggest that estimates of closure parameters using models fitted to aggregated and individual data had similar standard errors"* (Die et al. 2019). Thus, if we take Butterworth's (2020) argument at face value, we are being asked to make the assumption that the panel came up with the working hypothesis that "including random effects chosen using model selection methods will appropriately account for the pseudo-replication" after they had already seen results that, in Butterworth's opinion, refute that working hypothesis.

E: 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.

This points for the need for the PWG to consider whether these aspects, including robustness to alternative choices for random effect terms, have been adequately addressed - this before taking account of results from such analyses in drawing inferences about the impacts of island closures on penguins.

C: At the time that Butterworth (2020) was writing, the SWG-PEL had not been presented with the results of Sherley (2020). Sherley (2020) does indeed assess the robustness of the inference in Sherley et al. (2019) to alternative choices for random effect terms and it does so in line with the recommendations made by the 2019 panel (Die et al. 2019). Crucially, that analysis indicates that the closure effects based on the random effect structure chosen using model selection methods recommended by the panel (Die et al. 2019) were actually slightly stronger than those originally reported in MARAM/IWS/2019/PENG/P4 (Sherley et al. 2019). Thus, the key inference of Sherley et al. (2019) hold up to checks of "robustness to alternative choices for random effect terms".

E: 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).

C: Yes, I agree that the aggregation approach advocated by Butterworth (e.g. Ross-Gillespie and Butterworth 2020) is hampered by a low number of dof and that this is only "*ameliorated somewhat*" if the year factor is treated as a random effect (see also Sherley and Winker 2019).

Essentially, the concern here is that 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 non-conservative 3 data points per parameter (Crawley 2013), through 10 data points per parameter (Draper and Smith 1998) to the very conservative m^n 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. Moreover, aggregation always leads to loss of information; unknown factors at the finer-scale that could play a role cannot necessarily always be incorporated at the coarse-scale. This aggregation error can result in bias (Fritsch et al. 2020).

Indeed, the above point (that the approach using the aggregated data is hampered by a low number of dof and that this is only "*ameliorated somewhat*" if the year factor is treated as a random effect) 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".*

It is true, on the other hand, that there is a risk of making a Type I error when using the individual data approach if (as outlined above and elsewhere) the random effect structures is not chosen carefully (Arnqvist 2020, Silk et al. 2020). 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 Sherley (2020) has been presented to address that recommendation.

Finally, it should be noted that (generalised) LMMs are generally fairly robust to these issues: "*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 was almost completely absorbed by the nested random effect variance of interest*" (Schielzeth et al. 2020). And that providing "*statements of credibility based on Bayesian analyses with correct specification of design hierarchy*" (as done in Sherley et al. 2019, Sherley 2020) rather than relying on p-value based approaches is one of the recommended solutions for possible pseudoreplication when using LMMs (Silk et al. 2020).

E: 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.

C: These two sections seem to be both confused and confusing. The argument being made here that both the closure effect estimate and the standard errors will be effectively identical for the aggregated and individual approaches (in this hypothetical situation) appears to directly contradict the argument made in FISHERIES/2019/NOV/SWG-PEL/34 that using individual data "*will lead to over-optimistic estimates of the precision of estimates of the impact of fishing*" (Butterworth and Ross-Gillespie 2019). Moreover, the last two sentences (that mention AIC and model selection) appear to be suggesting that models applied to the aggregated and individual datasets should be compared using AIC. If so, it should be noted that AIC can only be used to appropriately compare models that are fit to the same response variable (y-values, or dataset). Thus, an AIC-based comparison of models fit to the aggregated and individual datasets would be neither informative nor helpful.

E: 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.

C: The argument here that "*pursuing an individual data approach*" offers "*very little if anything*" compared to an annually aggregated data-based approach "*as far as precision is concerned*", appears to directly contradict the key objection that Butterworth and colleagues have to using an individual data approach. Butterworth and colleagues have "*questioned for some time*" results from approaches that "*have directly modelled data from individual penguin observations*" on the basis that they "*will lead to over-optimistic estimates of the precision of estimates of the impact of fishing*" (Butterworth and Ross-Gillespie 2019). How can it be the case that an approach that directly models data from individual penguin observations can simultaneously (a) provide estimates of the closure effect that are overly precise (relative to analysis of the aggregated data) and (b) offer very little if any gain compared to an annually aggregated data based approach as far as precision is concerned? The two cannot both be true. Does Butterworth no longer feel that the approach used by Sherley et al. (2018, 2019) and Sherley (2020) is fundamentally flawed on the basis that it produces results that are overly precise, given that he now feels that "*very little if anything is to be gained from pursuing an individual data compared to an annually aggregated data based approach*" in terms of precision?

I suggest that we now cease having a technical debate on whether (generalised) LMMs can adequately balance Type I error rates and statistical power when random effects are selected by model selection approaches (since exactly this point has been established by simulation studies elsewhere; Matuschek et al. 2017) and focus on the key conclusion offered by Sherley et al. (2019) that:

"we have now iterated to a place where two independent sets of analyses agree that biologically meaningful effects of fishing around African penguin breeding colonies are apparent and, importantly, that some of those effects are on variables (chick survival, fledging success) that contribute to the demographic process".

References:

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

Butterworth DS. 2020. On estimates of the impact of fishing from analyses of the island closure experiment which model individual penguin responses directly. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/JAN/SWG-PEL/08. pp. 1–2.

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? Department of Environment, Forestry and Fisheries Report: FISHERIES/2019/NOV/SWG-PEL/34. Pp. 1–10.

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.

Draper NR and Smith H. 1998. *Applied Regression Analysis* (Third Edition). New York: Wiley.

Fritsch M, Lischke H and Meyer KM. 2020. Scaling methods in ecological modelling. *Methods in Ecology and Evolution*. DOI: 10.1111/2041-210X.13466.

Good PI and Hardin JW. 2006. *Common Errors in Statistics (and How to Avoid Them)*. Hoboken, NJ: Wiley.

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.

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-PEI/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.

Sherley RB. 2020. Revisiting the key results in MARAM/IWS/2019/PENG/P4 in light of the 2019 Panel recommendations. Department of Environment, Forestry and Fisheries Report: FISHERIES/2020/JUL/SWG-PEL/53REV. Pp. 1–27.

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.

Sherley RB, Barham BJ, Barham PJ, Campbell KJ, Crawford RJM, Grigg J, Horswill C, McInnes A, Morris TL, Pichegru L, Steinfurth A, Weller F, Winker H and Votier SC. 2018. Bayesian inference reveals positive but subtle effects of experimental fishery closures on marine predator demographics. *Proceedings of the Royal Society B: Biological Sciences* 285: 20172443.

Sherley RB, Barham BJ, Barham PJ, Campbell KJ, Crawford RJM, de Blocq A, Grigg J, Le Guen C, Hagen C, Ludynia K, Makhado AB, McInnes A, Meyer A, Morris T, Pichegru L, Steinfurth A, Upfold L, van Onselen M, Visagie J, Weller F and Winker H. 2019. A Bayesian approach to understand the overall effect of purse-seine fishing closures around African penguin colonies. Department of Environment, Forestry and Fisheries Report: MARAM/IWS/2019/PENG/P4. pp. 1–25.

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

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