On correctly specifying estimation models and the level of covariates: A reply to MARAM/IWS/2019/PENG/P7

Henning Winker[1] & Richard Sherley[2,3]

[1]*Department of Environment, Forestry and Fisheries (DEFF), South Africa*

[2]*Environment and Sustainability Institute, University of Exeter, UK*

[3]*FitzPatrick Institute of African Ornithology, University of Cape Town, South Africa*

Summary

In this paper, we present simulation testing results showing that an adequately formulated hierarchical mixed effects estimation model (EMF) prevents overstating the precision of the estimated model coefficients. Based on these results, we retain our position that, by ignoring a nested random effects structure, the incorrectly specified EMs put forward in MARAM/IWS/2019/PENG/P5 & P7 did not facilitate meaningful comparison or inference about the hierarchical mixed effect model structured as considered in MARAM/IWS/2019/PENG/P4. We agree that, similar to a design-based estimator, a linear regression fitted to means of aggregated data provides an unbiased estimator for the mean and standard error, but only if the observations indeed originate from a randomly stratified sampling design. However, the consequences of the unbalanced and somewhat opportunity based sampling design for the Island Closure Experiment remains unevaluated by simulation testing at this stage. To this end, we suggest that a major advantage of fitting hierarchical mixed-effects models to observations is the flexibility to account for important biological processes that are measured at observational level (e.g. brood mass) or lower hierarchical levels (e.g. monthly variation in chick condition) than is possible with aggregated means for island and year.

Introduction

In MARAM/IWS/2019/PENG/P6, we raised two major concerns regarding the properties of the simulation experiment design as presented in MARAM/IWS/2019/PENG/P5. These were: (1) the simulation experiment failed a critical 'self-test' and (2) the estimation model 'EMA' was misspecified, thus preventing any meaningful inference about the performance of Bayesian hierarchical mixed-effect models applied to real-world data in MARAM/IWS/2019/PENG/P4. We acknowledge that the corrections provided in MARAM/IWS/2019/PENG/P7 resolved the 'self-test failure' issue and that the simulation testing results are now consistent with our expectations given the simulation design. However, we also note that MARAM/IWS/2019/PENG/P7 did neither address our concerns regarding the apparent EMA misspecification, nor considered our proposal of a correctly specified EM with the desirable property that it is also consistent with the model structure used in MARAM/IWS/2019/PENG/P4. Here, we use simulated datasets for Run 10 from OM1 and OM4 (kindly provided by Ross-Gillespie) to demonstrate that our correctly specified hierarchical mixed-effects EM can produce unbiased precision estimates when fitted to observations.

Material and Methods

Due to time constraints it was agreed that additional simulation testing exercises shall be based on a selected simulation Run from MARAM/IWS/2019/PENG/P5. For this purpose, we initially considered Run 10 from the operating models OM1 and OM4 (originally assumed to be OM2). Unfortunately, the documentation of the OM4 Run 10 remains somewhat opaque as this run is not clearly referenced in Table 1 in MARAM/IWS/2019/PENG/P7. To prevent speculation, we therefore focus on Run 10 for OM1 and provide the results for the (undocumented) Run 10 from OM4 for reference in Appendix A. OM1 is formulated as (MARAM/IWS/2019/PENG/P4):

$$F_{i,y,z,j} = a_i + b_y + \delta\left(X_{i,y}\right) + c_{i,z,y} + \varepsilon_{i,y.z,j} \qquad \text{(OM1)}$$

where $F_{i,y,z,j}$ is the response variable on log-scale, a_i is the island effect for *i* = 1, 2 (fixed effect), b_y is the normally distributed year effect $b_y \sim N\left(0, \sigma_b^2\right)$ and δ is the binary closure effect (here common to both islands) for a vector with a sequence of 0's (closed years) and 1's (open years), such that a negative δ implies a positive closure effect (opposite to Sherley et al. 2018). The introduced 'hidden covariate' $c_{i,z,y}$ is realized by the fixed effect term $c_{i,z,y}$ with z factorial levels within each island *i* in year *y* and normally distributed effect sizes $c_{i,z,y} \sim N(0, \sigma_c^2)$, The error term is assumed to be normally distributed for the measurement error of penguin *j*, given covariate *z* on island *i* in year *y*, such that $\varepsilon_{i,y,z,j} \sim N(0, \sigma_\varepsilon^2)$.

Run 10 was provided in the form of 1000 simulated datasets with each *N* = 1800 observations, comprising 30 annual observations per island in each of 30 sampling years. The closure effect δ was set to 0.1, the island effect sizes were $a_1 = 0$ and $a_2 = 0.2$ and the year effect was generated form $b_y \sim N(0, 0.1^2)$. The 'hidden covariate' $c_{i,z,y}$ was implemented with 10 factorial levels *z* that were redrawn for each island *i* and year *y* with $c_{i,z,y} \sim N(0, 0.35^2)$ and the observation error was

$\varepsilon_{i,y,z,j} \sim N(0, 0.2^2)$. It is important to note that the simulation assumes a perfectly balanced, randomly stratified sampling design.

Here, we consider three estimation models (EMs). The first two EMs correspond to EMA and EMB in MARAM/IWS/2019/PENG/P5 and are used as control. The third EM is our proposed correctly specified 'EMF' with an additional nested random effect of island within year. The three EMs are specified as:

$$F_{i,y,j} = a_i + b_y + \delta(X_{i,y}) + \varepsilon_{i,y,j} \qquad \text{(EMA)}$$

$$\bar{F}_{i,y} = a_i + b_y + \delta(X_{i,y}) + \varepsilon_{i,y} \qquad \text{(EMB)}$$

$$F_{i,y,j} = a_i + b_y + d_{i,y} + \delta(X_{i,y}) + \varepsilon_{i,y,j} \qquad \text{(EMF)}$$

where b_y denotes a random effects term for the year effect, $\bar{F}_{i,y}$ denotes the mean value of the measured penguin response variable for island *i* and year *y* in EMB and $d_{i,y} \sim N(0, \sigma_d^2)$ is the additional random effect for island *i*, nested within year *y*. Note that the fixed effect models EMC and EMD are not considered further here because they produce (unsurprisingly) qualitatively the same results as EMB and EMA, respectively (c.f. MARAM/IWS/2019/PENG/P5 & P7).

For ease of comparison, we computed the same performance metrics as described in MARAM/IWS/2019/PENG/P5 & P7. Accordingly, the bias in precision of the closure effect δ is defined as the difference between the $mean(SE_\delta)$, computed from the mean of the estimated $SE_{\delta,k}$ across the simulation replicates *k,* and the 'true' $SE_\delta(true)$, computed from the standard deviation of the estimated δ_k across 1000 simulations replicates k . We reiterate that this bias calculation seems to have the undesirable property to rely on unbiased estimates of δ_k in the first instance, which may not necessarily hold in a less idealized future OM setups that consider, for example, an unbalanced sampling design. As a more robust alternative, we recommend the conventional metric of confidence interval coverage (e.g. Winker et al. 2020), which computes the probability of the 'true' value falling within the 95% CIs against the nominal 95% probability expectation.

Results

The simulation results for Run 10 from OM1 are shown in Figure 1. Our simulation results for EMA and EMB confirm the results by MARAM/IWS/2019/PENG/P5, in that both EMs produced unbiased (though variable) estimates closure effect δ, but EMA resulted in a notably underestimated $mean(SE_\delta)$ relative to $SE_\delta(true)$. Our results also show that the correctly specified EMF produced approximately unbiased estimates of both the effect δ and its precision in the form of the $mean(SE_\delta)$ (Fig. 1). This also holds for Run 10 of OM4 (Fig. A1).

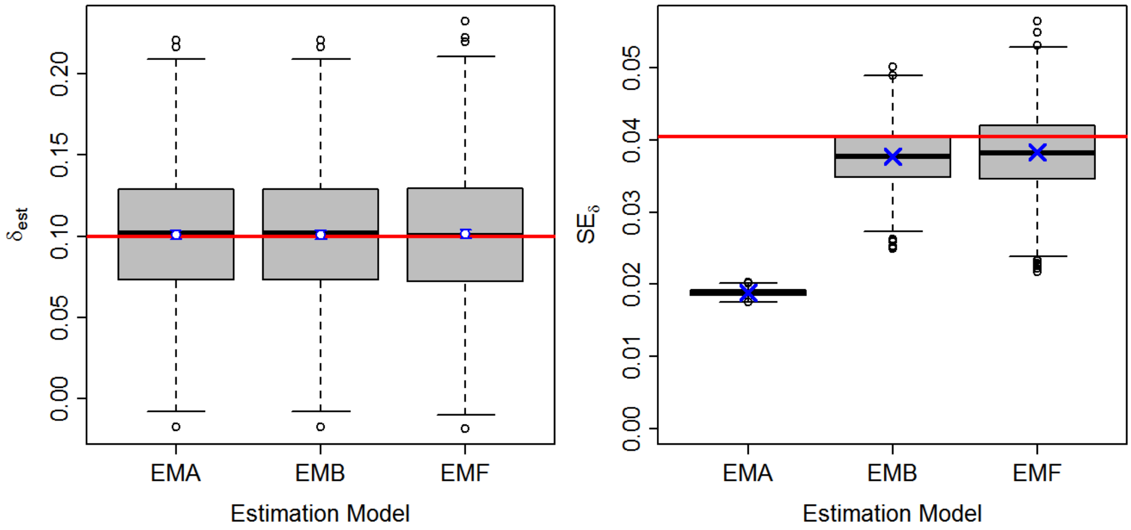

Fig. 1. Boxplots showing the estimates of the closure effect d (left) and the estimated standard errors SE_δ (right) in comparison to the 'true' values (red lines) for 1000 simulation datasets from Run 10 of OM1. The blue circles and error bars (left) represent the mean and 95% CIs computed for the 1000 replicates and the blue X's (right) show the $mean(SE_\delta)$.

Discussion

Here we have used 1000 simulated datasets for Run 10 from OM1 and OM4 (MARAM/IWS/2019/PENG/P7) to demonstrate that a correctly specified EMF produces unbiased precision estimates when fitted to observations directly. All it needs is to introduce a random effect for island nested within year, which is also intuitively compatible with the implicit sampling stratification assumptions of aggregating the observations to annual means for each island and year. We therefore suggest that EMF provides the appropriate random structure for allowing a 'fair' simulation evaluation against OM1 and OM4. We concur with MARAM/IWS/2019/PENG/P5 & P7 in that, similar to a design-based estimator, a linear regression fitted to means of aggregated data also provides an unbiased estimator for the mean and standard error, but only if the observations indeed originate from a randomly stratified sampling design. However, the consequences of an unbalanced and somewhat opportunity based sampling design of the Island Closure Experiment remains unevaluated by simulation testing at this stage.

In contrast to EMB (or EMC), the hierarchical mixed-effects EMF frees substantial degrees of freedom, which also enables the inclusion of additional important predictor variables (e.g. brood mass for tracking data) and biological processes that operate at finer scales (e.g. bird ID, nest ID or month effects; see detail in MARAM/IWS/2019/PENG/P4 and references therein). Documented examples for the real-world Island Closure Experiment data include that (i) beta (second hatching) chicks do not have the same probability of surviving as their alpha (first hatching) counter-parts in all years (non-independence), (ii) chick condition exhibits substantial monthly changes that vary in timing from year to year, and (iii) not accounting for brood mass when modelling indices of foraging effort ignores the well-established principle that seabirds forage further and for longer as their chicks grow and need more food. In response to MARAM/IWS/2019/PENG/P7 and in particular their statement that "[Sherley et al.] use of individual data appears equivalent to pseudo-reflection" based on their concerns about "[not accounting for] between-year information that relates to the process error effects on precision", we can only reiterate that models in MARAM/IWS/2019/PENG/P4 do in fact account for hierarchical sources of variation in-between years, which we explicitly considered to minimize the risk of pseudo-replication. This is implicit in that all models include a biological plausible random effect that is nested within year.

To this end, we provide simulation based evidence showing that adequate hierarchical random effects structures can indeed be effective in preventing the overstating of the precision of the estimated model coefficients. We therefore retain our position that, by ignoring nested random effects at a lower hierarchical level than year, the incorrectly specified EMs put forward in MARAM/IWS/2019/PENG/P5 & P7 do not facilitate any inference about the hierarchical mixed effect models in Sherley et al. (2018) and MARAM/IWS/2019/PENG/P4. We therefore conclude that the interpretations of simulation results presented in MARAM/IWS/2019/PENG/P5 & P7 are severely overstated.

References

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, 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.

Winker H, Carvalho F, Thorson JT, Kell LT, Parker D, Kapur M, Sharma R, Booth AJ, Kerwath SE. 2020. JABBA-Select: Incorporating life history and fisheries' selectivity into surplus production models. Fisheries Research 222: 105305.

Appendix A

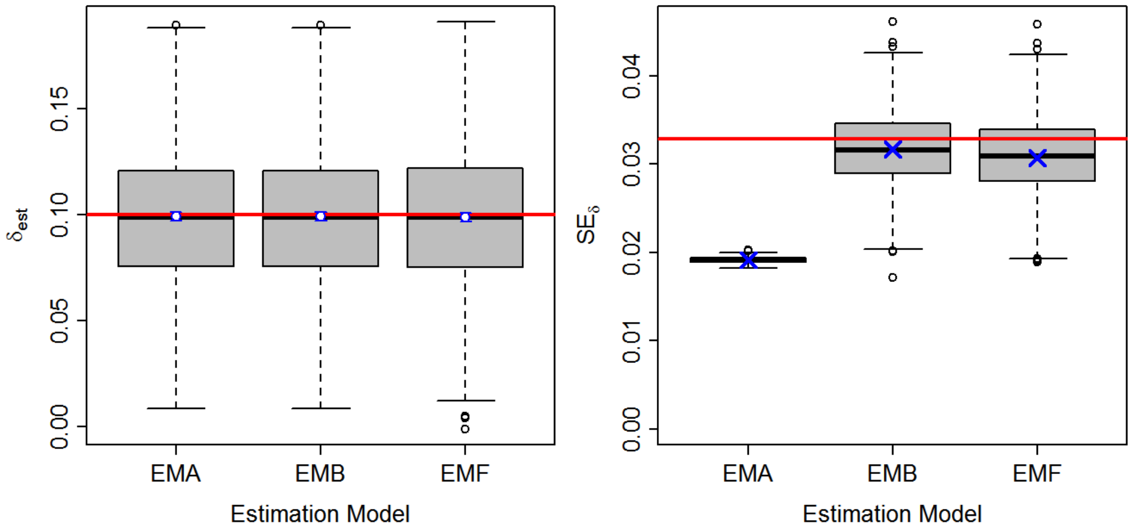

Fig. A1. Boxplots showing the estimates of the closure effect d (left) and the estimated standard errors SE_{δ} (right) in comparison to the 'true' values (red lines) for 1000 simulation datasets from Run 10 of OM4. The blue circles and error bars (left) represent the mean and 95% CIs computed for the 1000 replicates and the blue X's (right) show the $mean(SE_{\delta})$.