Initial comments on the remaining limitations of MARAM/IWS/2019/Peng/WP4: “Introducing an imbalance in the sampling from the unknown covariate from OM3”

Henning Winker1

1. Department of Environment, Forestry and Fisheries (DEFF), PO Box 52126, Cape Town 8000, South Africa.

Firstly, I note that both the "two-stage" estimations EMB (random year effect) or EMC (fixed year effect) approaches (MARAM/IWS/2019/Peng/P7), fitted to aggregated means by island and year, and the "one-stage" EMF approach (MARAM/IWS/2019/Peng/P8), fitted to observation with an equivalent nested random effects structure, both provide unbiased estimates in terms of the effect size and the standard error, given that the following conditions (C) are met:

(C1) The sampling design randomly stratified

(C2) The sampling strata ("hidden" or known) is balanced (equal sampling of each grouping)

(C3) The available covariates are measured on the same level as the aggregated data

(C4) The number of sampling years is sufficiently large to prevent overfitting of EMB or EBC

MARAM/IWS/2019/Peng/WP4 simulation tested a modification of OM3 where the factorial levels *z* of the “hidden covariate” are sampled randomly instead of imposing an equally balanced sample sizes per group, i.e. violating (2). In the email correspondence with the chair and the authors on the 2nd of December, I had hypothesized that although this may increase the "true" SE, it will not result in any bias of SE if conditions of (1), (3) and (4) are met given that those are the underlying assumptions of a design based estimator that are still be met. Further, I predicted that the nested random effect model in EMF cannot have any meaningful advantage over the 2-stage approach if there is no additional information at a lower hierarchical than the aggregated data, e.g. month (lower level than year) or brood mass (same level as observation). Unfortunately, MARAM/IWS/2019/Peng/WP4 did not even consider EM F. We reiterate that EMA is mispeficied and with regards to the OM and that the apparent bias shown in results thus far is therefore somewhat predictable, has been pointed out in Peng/P6 and Peng/ P8. On this ground I do not effort further considerations on EM A here.

In the following I formulate point out some limitations of the comparative evaluations presented thus far in the form of five questions (Q). The first Q1 addresses the implementation of the ‘hidden covariate’ (random vs nested) and questions (Q2) – (Q4) are then relate to the robustness of model structures to conditions (C2) - (C4) and Q5 is intended to link back to the “real-world” and initial comparisons presented in Peng/WP3.

These are:

(**Q1**) What if the year to year by island variation in the ‘hidden covariate’ c[z,i,y] does not vary at random, but instead preserves some hierarchical  structure, such that the 3 levels z, i, y of effects become nested and additive, where factorial levels z is maintained the same over time series, then i is maintained for each island and added to z and additional noise is added each year for each island i and factorial effect z?

(**Q2**) What if in addition to **Q1**, one factor z of the hidden covariate is systematically over-sampled relative to the others, i.e. violating (C2) given a nested covariate struture?

**Q3**) What if in addition to **Q1** and **Q2**, an additional covariate is observed at a lower hierarchical level, which could therefore only be accounted for EMF but not EMB/EMC, i.e. violating (C2) - (C3)?

(**Q4**) What if the time series is insufficiently long (e.g. 10 years) and the standard error of e.g. EMC (fixed year effect) inflates relative to EMB (random year effect) or e.g. EMB relative to EMF - can this be seen as evidence for overfitting, i.e. violating (**C4**)?  Associated with this - should a singular fit warning be of concern?

(**Q5**) Would a comparison of standard errors of the closure effect among models fitted to the real data be useful?  This is at least partially in the examples in PENG/WP3

Again, I argue that, as long as all conditions (1), (3) and (4) are met, a two-stage approach is indeed an unbiased estimator, and there is no need on spending resources on simulation testing to agree on this (c.f. Peng/P8). As such the results presented in PENG/WP4 were predictable based on first principle (see 2n Dec Communication). The over-arching question remains if this indeed holds in real-world applications?