ON THE USE OF AGGREGATED *vs* INDIVIDUAL DATA IN ASSESSMENT MODELS

Doug Butterworth
MARAM
Department of Mathematics and Applied Mathematics
University of Cape Town
Rondebosch 7701

Summary

The conventional two-step process in fisheries assessments, whereby data are first aggregated to provide typically annual values before those are input to the assessment model, is compared to a single-step process where the individual data are input directly to the assessment model. The key point at issue is whether or not the latter process would provide estimates of key parameters that are (and are reliably estimated to be) more precise in circumstances where there is non-independence in the individual data. Arguments are offered that this non-independence does not introduce bias into estimates of precision for the aggregated case when observation error variance in the data is much less than process error variance in the assessment model. The utility of the random effects approach for addressing non-independence through working with individual data in a single-step process is queried; this is because of uncertainty about the bias in estimates of precision that may arise because of a lack of knowledge in most situations whether the structure assumed for the random effects will adequately account for the actual (and usually unknown) sources of non-independence in the data. Some aspects of the issue are illustrated by quantitative examples.

Introduction

A debate has developed in the Penguin Task Team (which is co-ordinating the responses being developed to the 2015 International Panel recommendations on analyses related to the impact on penguins of closures to pelagic fishing of areas around Dassen and Robben islands) on an issue that is of wider relevance to stock assessment practice. This concerns whether there are advantages in utilising individual data rather than (say) annual means in fitting (assessment) models. Thus, for example, conventional practice in fitting an assessment model (a Schaefer model, say) to abundance information (say from CPUE data) is to first aggregate those CPUE data into annual means (perhaps standardised in some GLM process to adjust for imbalances in the data in relation to co-variates such as month), and then to fit the population model to those annual aggregates. But what about the alternative of a one- rather than a two-stage process, whereby the assessment model would be fit directly to the individual CPUE data points? Would this use of (apparently) more information in fitting the assessment model yield better (e.g. more precise) results?

Sufficient statistics

Consider the contribution to the negative log-likelihood minimised in the model fitting procedure to individual CPUE data for one year. In the interests of simplicity assume that the individual CPUE values have been converted to biomasses x_i using the swept area approach, with catchability q known to be 1. (The result that follows is not dependent on these

assumptions – they simply reduce the complexity of the explanatory algebra needed.) Assuming independence and distribution normality that contribution is (ignoring constants):

$$-\ln L = n \ln \sigma + \Sigma_i (x_i - X)^2/(2\sigma^2) \quad (1)$$

where X is the underlying true biomass and n is the number of individual CPUE data points.

With a little algebra:

$$-\ln L = n \ln \sigma + [(x_{bar} - X)^2 + \sigma_x^2]* n/(2\sigma^2) \quad (2)$$

where x_{bar} is the mean and σ_x^2 the variance of the n individual data points x_i.

Thus the mean, variance and number of data points are sufficient statistics to completely define the likelihood required for fitting the model to these data, and that likelihood contains all the pertinent information. **There is therefore nothing to be gained in terms of improved estimation performance by fitting to the individual data for each year rather than to their means.** (This argument generalises straightforwardly if working in log-space and assuming that log-normal distributions apply.)

Non-independence of data

The exact form of equation (2) follows from the assumption of independence. In practice in fisheries, this seldom if ever holds, even for a designed experiment (e.g. a research survey in contrast to CPUE data to provide an abundance index), because of spatio-temporal relationships amongst "nearby" data. Thus, for example, the standard error of the mean for the typically very large number of individual CPUE data points each year would seriously overestimate the true precision of that mean as an index of abundance. The net effect of this is that the "effective" number n_{eff} of data points each year is (perhaps considerably) less than the actual number of observations n.

Many different methods have been used in fisheries to estimate n_{eff} so that information from a certain source of data is not over-weighted in fitting assessment models. Thus for example random effects models, or models that used lumped sets of data to attempt to integrate over correlations, may be used for standardising CPUE series, while approaches such as suggested by McAllister and Ianelli or by Francis may be applied to down-weight the likelihood contributions of catch-at-age data which have been calculated under an assumption of independence.

Process *vs* observation error

Notwithstanding these attempts to account for non-independence, the general experience in fisheries with annual indices, particularly of abundance, is that process error variance dominates observation error variance. Put another way, this means that the variance of the assessment model residuals (the difference between the index and the corresponding assessment model estimate each year) is substantially greater than can be accounted for by the variance (even if adjusted for non-independence) of the mean of the individual data points for that year (the "observation error" variance). In these circumstances, additional

observations each year will add little to overall estimation precision, to the extent that frequently observation error is ignored in fitting models, with the extent of process error variance (sometimes called additional variance when observation errors are incorporated explicitly) estimated when fitting the assessment model; hence this process error estimate subsumes the smaller observation error which is effectively taken to be constant.

For the penguin closure model analyses, this point was examined in section 2 of document MARAM/IWS/DEC15/PengD/P2. Those analyses indicated that observation errors (SEM's from the individual data) were generally very much less than total errors, i.e. process error variance completely dominated observation error variance (and would still have done so even were the SEMs to have been inflated reasonably to allow for non-independence effects). It was these results that presumably led to the Panel's recommendation A.2.7 that:

> "(**1**, *Allowance for sample size in estimator*) There is no need to account for sample size when generating data in any simulations given the low observation error relative to process error (MARAM/IWS/DEC15/PengD/P2). However, it is also reasonable to exclude data points based on very small sample sizes (perhaps < 5 points) when conditioning the operating model or to estimate the sample size component of the observation error."

On Random effects models applied to individual data to account for non-independence

Suggestions have been made that the application of random effects models to individual data in the penguin closure model analyses could lead to results with enhanced estimates of precision. Actual truly comparable results have yet to be calculated to see if this is the case, and further it seems unlikely that differences (if there are such) will greatly impact the key results from these analyses. Nevertheless this suggestion is of importance, because if it holds in this case, it would seemingly offer considerable benefits also for other fisheries assessment exercises in moving from a two- to a one-step calculation process as described above.

A potential concern about the aggregated approach used in the penguin analyses is that combining data at an annual scale may over-correct for non-independence, with information content in the data being lost. This was examined for the case under consideration (chick condition) by computing standard error (SE) estimates for the effect-of-fishing parameter by using a jack-knife approach (with years being treated as the sampling units), and comparing these with the Hessian based estimates for analyses using annual aggregates. If the latter were heavily over-correcting for non-independence, one would expect the jack-knife SE estimates to be notably lower. The results are shown in Table 1. For the catch only estimator, the jack-knife estimates are in fact higher for both islands. For the closure only estimates, for one case the jack-knife estimate is much lower, and the other much higher – a surprising difference which might be pointing to lesser robustness of the closure only estimator. Viewed overall though, these results are not suggestive of substantial under-estimation of precision by the aggregated approach in this instance.

A concern for the random effects approach is robustness. In some cases (with an obvious nesting structure), the best way to implement the random effects approach may be reasonably clear. However in other cases this is not so, and different assumed structures may yield quite different results (as has been found, for example, in some ICCAT CPUE standardisation exercises). Without the source and structure of non-independence (or at least their statistical

properties) being clear, how can one be sure that simply applying some random effects approach will adequately account for non-independence (because if it does not, the results output would indicate spuriously good precision).
To examine this further, a very simple simulation was developed. Data (x) are generated using the following equation reflecting a period of a "year" conveniently considered to be made up of 12 months each of four weeks totalling 360 days:

$$x = a + b + c \quad (3)$$

where a=100, b is a daily random effect ~ N(0; 30^2), and c is random noise ~ N(0; 10^2). Observations are made 100 times each day throughout the year and used to estimate the mean value of x in a variety of ways. These are first a standard mean, and then estimating as $a' + b'$ where b' is a random effect over a period whose length is varied here (because in practice the level at which the non-independence is arising would not be known). The process is repeated 1000 times to allow for measures of estimator performance to be made.

The point of interest is the fraction of the true standard error (known from equation (3)) to which the standard error estimate of a' corresponds. This is reported in the Table 2 below, which also includes results for the case where the 100 observations are taken only once (during a single day) every week rather than every day.

There is fairly strong non-independence in these simulated data, such as that if that was ignored in estimating standard errors, those would be an order of magnitude too small. The random effects approach results are reasonably robust to choice of the unknown time scale at which the effect operates in their provision of a correction for this underestimation, but nevertheless the approach does manifest a bias whose size increases as the difference between the actual and the assumed value for this time scale widens.

The point of this example is not to pretend that it intends a close reflection of some actual (penguin, say) analysis situation, but rather simply to show that the random effects approach for correcting variance estimates for non-independence in data can perform reasonably well and be reasonably robust (as in this example), but nevertheless is not perfect and may fail to secure full correction. Note that this was a deliberately very simple example with the form of the estimator being identical to that of the model used to generate the data; larger differences might be expected for more complex situations where also this structural identity is unlikely to hold.

Thus it seems one cannot assume that a random effects estimator will fully correct for non-independence of data; rather it seems likely to yield estimates of standard errors for parameters which are negatively biased to some extent. This implies a need for checking for each particular case considered, including an investigation of the robustness of the approach to errors in assumptions about the actual error structure underlying the data. [Penguin Task Team members kindly provided some references to the random effects approach and its applications, but though clear expositions, these did not appear to address the key issue here of robustness to mis-specified structure.]

In summary

The aggregated data approach might over-correct for non-independence of data, conceivably thereby sacrificing some estimation power. However, this would not seem to be an issue in instances where process error variance is much greater than observation error variance, and furthermore in a particular (penguin-closure-related) example of interest here, there seemed to be no obvious indication that this approach was resulting in underestimation of precision.

Applying random effects models directly to individual data is a promising approach, but a major problem would seem to be structural robustness – how does one know in any situation that the structure assumed for the random effects will adequately account for the actual sources of non-independence in the data? A related problem, if such models are to be used in a power analysis context, is that such an analysis requires generation of future pseudo-data on an individual observation basis. One cannot simply resample at random with replacement from past individual data, as that will not incorporate any non-independence; generating future data with those statistical properties would seem to require first identification of the statistical properties of the mechanism(s) underlying that correlation structure.

Acknowledgements

Andrea Ross-Gillespie is thanked for carrying out the computations reported in the Tables in this document.

Table 1: A comparison of jack-knife based estimates of the standard error (SE) of the effect-of-fishing parameter for catch only and closure only estimators with those from the Hessian for the standard GLM approach for year-aggregated data for chick condition.

Jack-knife SE		
	Catch only	Closure only
Dassen	0.21	0.11
Robben	0.20	0.38
GLM SE		
	Catch only	Closure only
Dassen	0.18	0.22
Robben	0.15	0.20

Table 2: Simulation results for tests (see text for details) of the ability of a mis-specified random effects approach to adjust for underlying non-independence of data from an effect operating at a daily time scale. The table entries reflect the proportion of the true standard error estimated by the random effects estimator.

Estimation approach	**Daily observations**	**Weekly observations**
Data mean	0.104	0.100
Random effects: time scale		
6 months	0.795	0.752
3 months	0.920	0.876
1 month	0.965	0.945
1 week	0.984	0.959

1 day	0.990	0.959