Contract reference n°: CC/13/04

Consultancy Report: March 2013

Author: Brian McArdle

Objective: To improve the estimation of species and size composition of the western and central Pacific purse seine fishery from observer based sampling of the catch

Executive summary and recommendations

1) The multinomial approach should be employed. Personally I have a feeling the modelling approach of Lawson might have some advantages if combined with a multinomial model. Basically the problems presented by the data seem to require exploitation of every opportunity to improve the quality of the predictions (smoothing, covariates, strata). Evidence for their utility would be from simulation.
2) However the model is developed, it is clear that robust and appropriate measures of the quality of the estimates must be put in place, validated by simulation
3) A comprehensive multipurpose simulation program is therefore essential.
4) The available data should be examined to establish what stratification is necessary and simulation studies performed to see if useful results can then be achieved.
5) The same analysis should estimate parameters so that the simulation can reproduce realistic scenarios.
6) At the design stage of the simulation, information essential for the useful implementation of the full simulation should be identified and if not yet available then, where possible, data should be collected. For example: the true magnitude of the layering effect in brails.

I read the reports from Joseph Powers and Patrick Cordue. I was impressed by both of them. They tended to focus on very different aspects of the problem and were thus complementary, but reached substantially similar conclusions. Both agreed that a multinomial approach would be more appropriate than the current model based approach presented in Tim Lawson's report. I will discuss this later.

Powers emphasized the importance of a simulation to evaluate different sampling designs and methods, and to investigate the impact on these of biases intrinsic to these methods (like layering). Cordue also recommended a simulation be prepared.

Cordue concentrated more on deriving a multinomial based procedure for estimating a correction term for the bias due to grab sampling using the paired sample data and applying this to the historical records. He suggested two approaches, a simple one, straightforward to implement and convenient to use, and a more complex one based on an explicit likelihood approach that would be tricky to implement, and might not function significantly better than the simpler approach. I concentrated on the simpler approach.

The multinomial approach.

The importance of the multinomial approach is mainly that the number of fish in a sample must be used to weight the contribution of that sample to the length frequency distribution when samples are pooled or modelled. If you pool two samples of different numbers of fish then the larger sample has a more reliable estimate of the population length frequency distribution than the small and should therefore contribute more. It was one of the drawbacks of the existing system that this did not happen. Of course, the size of the resulting bias might be small but it makes defending it more difficult.

One striking feature of his simplified approach is that it works on heavily pooled data, pooling over brails, sets, even trips. His objective is to produce strata specific "average" correction terms. By doing this he appears to sidestep one of the main problems with these data that there is so much between sample variability that the large sample sizes that are the norm with spill samples are effectively illusory (with effective sample sizes between 4-10% of the observed, see later discussion). So, very large sample sizes are necessary before the resulting bias correction terms are reliable enough to be useable. It is one of my concerns that the much more sophisticated maximum likelihood that is cleverly designed to cope with between brail variability and sources of bias etc might actually be used for small scales where the estimates are so unreliable to be useless. Better to lose detail by using larger strata to get reliable if slightly biased estimates, rather than get unbiased but unreliable estimates. Cordue is clearly aware of this problem with the likelihood model:
"In the simplest case the selectivity of the brail would be assumed constant over all sets. However, more than one brail selectivity could be modelled. Likewise, the observer selection-preference parameters could be assumed constant over all sets or they could be assumed to vary *over a small number of classes* (my italics)". In other words, strata within which the correction terms can be assumed constant must be large (few in number). In which case, one of the advantages of the more sophisticated method is lost. Until work has been done to investigate how many strata are actually suggested by the data and simulations performed to see how many are unnecessary for effective analysis, it is difficult to tell if the more sophisticated method will give any practical advantage over the simpler method. Being simple would make a method easier to defend comprehensibly, I am not in favour of defence by obfuscation (unnecessary complexity).
A further observation. The more complex model offers the prospect of correcting for biases produced by the phenomenon of layering. It might be worthwhile demonstrating that these biases are of a magnitude worth worrying about before committing to the more complex model. This would require data that showed the magnitude of the problem, and simulation to show the magnitude of the consequences once brails have been pooled into sets or trips.

The Lawson model.

Tim Lawson's model approaches the problem from the smaller scale (sets) and then pools the data (gets predicted mean values) using covariates and strata variables in a linear model. While all the available evidence suggests it works adequately and by modelling the correction term (smoothing the size class distribution of it) it may even have an advantage over the strictly multinomial methods, as currently implemented the main statistical objection is that it does not explicitly incorporate the numbers of fish in each set, leading (almost certainly) to bias. Having said this, the bias may not be large enough to worry about, but starting from a multinomial framework avoids this problem. Modelling the grab bias explicitly using the multinomial (as Cordue does: $\hat{r}_k = \frac{n_{jk} / n_j}{N_{jk} / N_j}$)[1] also makes method description (and therefore defence) easier. Using the availability (the probability of a fish being taken in a grab sample from all the fish in the set) as Lawson does, ignores the problem of combining such numbers from different sized sets, and makes the bias correction less intuitive ($\frac{1}{\hat{r}_k}$ is pretty intuitive).

Cordue on the Lawson model.

I feel Cordue wrote his critique of the Lawson model before he produced his own (simpler one) as some of the criticisms he aims at one apply equally to the other.

I will treat them in order (more or less).

1) That availability depends only on the number of fish in the set from that size class, though clearly this probability depends on the total number of fish all classes.

   I feel that here Lawson's own notation lets him down. $\hat{A}_k = \frac{n_{jk}}{\hat{N}^T_{jk}}$ (using Cordue's subscripts as above – with N^T as the true number of fish) implies that A is the same over all sets differing only between size classes. Cordue clearly interpreted it that way. But, as Cordue explains, it actually depends on the number of fish in the set (in the other size classes). It is therefore unique to each set. Insofar as the covariates in the final model explain differences in the set size and composition (i.e. the strata in Cordue's approach) this could be less of a problem than might appear at first sight. This variation might even be left in the error term (with unknown effects on inference due to non-independence – intra-class correlations) and still get usable estimates. However, having the crucial correction factor as a set dependent random variable makes it harder to explain to people (even without noting the fact in the formula which needs to be rewritten with a set subscript for *A:* $\hat{A}_{jk} = \frac{n_{jk}}{\hat{N}^T_{jk}}$)

2) "Between-brail structure is also ignored by the model" As it is in his simple model. At present we do not have enough information to know if this matters or not when data are being pooled over large strata.

[1] *n* is grab sample, *N* spill sample, *j* is set number, *k* is size class number.

3) "Equation (2) introduces a mathematical error. The total set weight divided by the mean fish weight (in the set) is the total number of fish in the set. This needs to be multiplied by the proportion of fish in the *j*th length class, by *number*, to obtain the number of fish in the *j*th length class. Instead, equation (2) wrongly uses the proportion of fish in the *j*th length class by *weight*."(Note: he is using Lawson's original subscripts not his own convention that I use throughout)
   I believe here he failed to see the model clearly. It seems to me clear enough that $T_{jk}W_j$ is the estimate of the weight of fish in class *k*, dividing this by $\overline{w}_k$ the average weight of fish in this (*k*th) class turns it into an estimate of the number of fish in this class. I believe Cordue failed to note the *k* subscript on the $\overline{w}$. Were it absent this would lead to the interpretation that he uses.

4) "The substitution of the spill-sample data for the set's length composition and mean weight is yet another flaw in the approach as the substitution of random variables for unknown parameters introduces additional error structure which is not captured by the single additive error."
   This criticism can again be made of his own simple model. By pooling over significantly different brails and sets, his length frequency distribution is no longer multinomial due to the extra-multinomial variation. He is therefore ignoring important additional error structure himself. Of course he is well aware of it, stressing the importance of identifying strata whose use would remove some of this structure (in the same way that covariates would do the same in Lawson's model), and advocating the use of effective sample size to give some measurement of the effects of this additional error structure.

Conclusions

1) Tim Lawson's model is not as bad (given the limitations of the data) as Cordue suggests. However a multinomial based method, though not necessarily producing better estimates, would undoubtedly be easier to defend and develop further.
2) The simple multinomial approach that both suggest provides a sensible basis for further development.
3) It is unclear that the much more complex likelihood based multinomial model is justified by the structure and quality of the data.
4) The simulation recommended by both is indispensable for the development of an appropriate correction method and any modification of the sampling programme.

Implementing Cordue's simple multinomial model.

The estimation of the grab bias $\hat{r}_k$ is trivial, it is the ratio of two multinomial proportions. Code to do this for separate or pooled species was left with Tim Lawson. I regret that at the time I did not incorporate the production of bootstrap estimates of error, since the sample sizes differ widely between size classes and species. This should be done as a priority as part of the investigation for useful stratification of the grab bias.

Further investigation of the appropriate strata for grab bias needs to be done. Preliminary work with set type (associated or unassociated) does not suggest a difference in grab bias. Of course certain species size classes are effectively absent in a given set type but that is not the same thing. With other covariates however, differences may appear. It is possible that between species (independent of size) differences in grab bias may exist (as Cordue assumes in his description of the model), however that is not clearly established. Since larger strata are always an advantage with these data investigating that rigorously would be a priority (probably using multinomial mixed models).

Code was written to apply the bias correction to grab samples (Appendix 1). This code stratifies by set type and performs the bootstrap resampling first on trips (to assemble a random sample of trips) and then separately on the two set types within trips.

The code is organised to follow Cordue's 5 steps:

" To estimate a stratum-wide species-length frequency from the grab-sample data I suggest the following approach.

1) Uncorrected grab-sample data are scaled-up within set (using the estimated *number* of fish in the set) and then added across sets to provide an aggregated species-length frequency.
2) This is then corrected using stratum-appropriate estimates of observer selection preferences (see the earlier section).
3) Scaling up to stratum numbers is then done by multiplying the proportions within each species-length class by the total number of fish caught in the stratum (estimated using the species-length frequency to get the mean fish weight for the stratum and dividing total stratum catch by the mean fish weight).
4) Effective sample sizes for the species length frequencies can be estimated by bootstrapping from the raw data.
5) The stratum-wide catch by species is then estimated from the stratum-wide species-length frequency. The numbers in each species-length class are multiplied by stratum-appropriate mean fish-weight for each species-length class. The uncertainty in the estimates can be estimated by bootstrapping from the raw data."

Steps 1 to 3 are straightforward, however step 4 required investigation of appropriate ways of calculating effective sample size. As explained earlier, since the sets pooled into a stratum estimate of the length frequency distribution are heterogeneous the resulting distribution is not multinomial and, because of this additional variance, the observed sample size conveys an unjustified confidence in the estimates. In fact the effective sample size is a great deal smaller than the observed one.

A definition of effective sample size: " The effective sample, m_{eff}, is defined as the number of fish that would need to be sampled at random so that the sample mean would have the same precision as an estimate based on a sample of *n* clusters." (Pennington et al 2002)

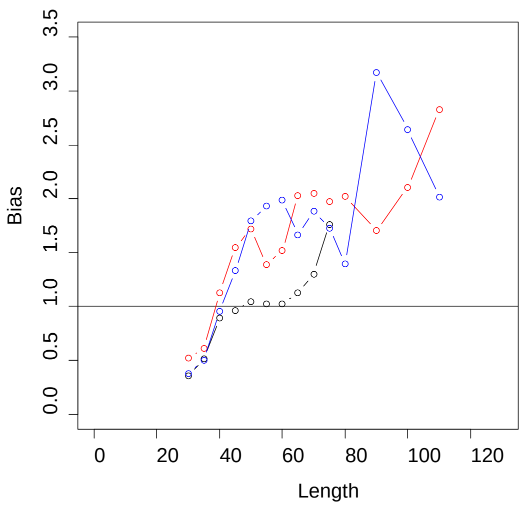
This refers to a mean and the clusters are the sets (the reason the fish are not sampled completely at random). The principle is however the same when the estimates are multinomial proportions rather than a mean.
After preliminary investigation of alternative methods that of Chih 2009 was chosen, implemented, and tested by simulation. It seems to work reasonably well.

Testing Cordue's simple multinomial method.

This was done using the grab samples for which match spill samples existed. This would, we hoped, provide a slightly optimistic test bed for the method. Tim Lawson's method was used for comparison and the Observer eye estimates are also presented.

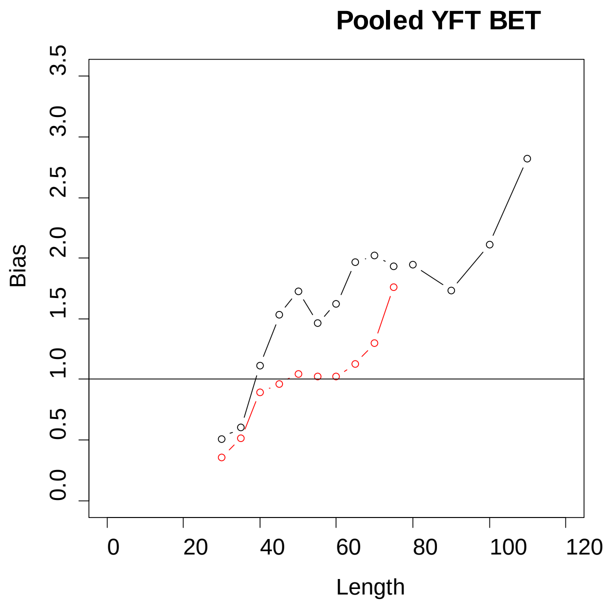
Estimates of grab bias.

Stratifying by species



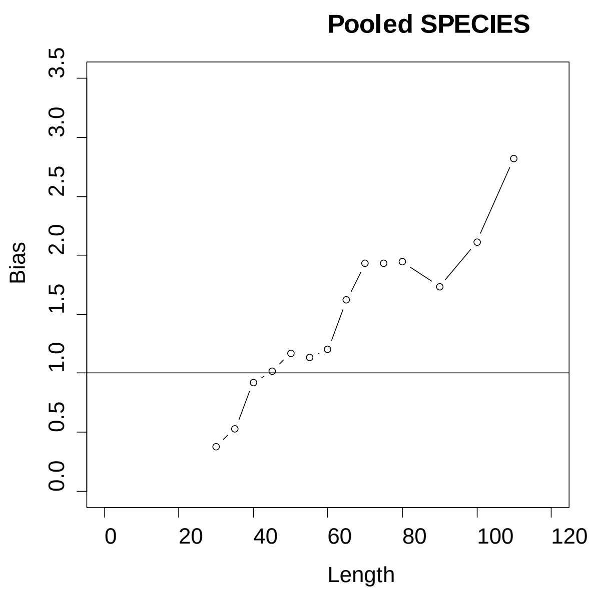
Key: SKJ black, YFT red, BET blue.
Clearly the expected bias against small fish is present. However, less expected is the apparent preference for YFT and BET over SKJ even within the same size class. The values for the largest fish are unreliable as they are based on relatively small numbers of sets and therefore fish (and, as explained earlier, the variation between sets is large).

Since there did not seem to be great deal of difference in length specific grab bias between YFT and BET they were pooled and grab bias calculated again.

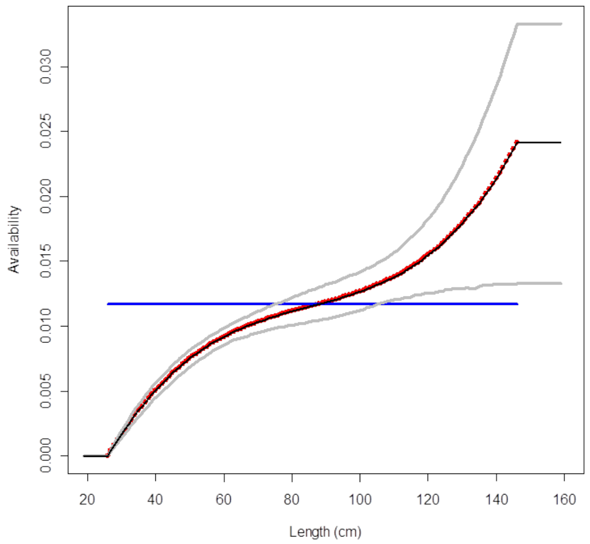


Key: SKJ red, pooled YFT and BET black
The difference between SKJ and the other two is even clearer suggesting this might be an important stratifying difference. Investigating this further would be an important part of the search for useful strata.

For comparison with Tim Lawson's method all the species were pooled in to one stratum and the grab biases calculated.



Without confidence intervals it is difficult to assess the reliability of the values for large fish and therefore whether there genuinely is a flick upward at the end. It might seem somewhat unlikely.
It is interesting to note that the same pattern emerges from Tim Lawson's model.



His availability, though not the same as Cordue's grab bias, shows (after modelling) substantially the same pattern.

Comparison of catch by proportion of weight for Associated and Unassociated sets.

Uncorrected catch

	skj	yft	bet
A	0.618	0.337	0.045
U	0.842	0.152	0.006

Spill sample estimates

	skj	yft	bet
A	0.679	0.286	0.035
U	0.830	0.164	0.007

Observer eye estimates

	skj	yft	bet
A	0.649	0.307	0.044
U	0.863	0.132	0.005

Cordue's correction (pooled species)

	skj	yft	bet
A	0.675	0.288	0.038
U	0.916	0.080	0.005

Bootstrap standard deviations.

	skj	yft	bet
A	0.032	0.030	0.007
U	0.048	0.045	0.004

These standard deviations are quite large. Estimates +/- 2.standard deviations (crude confidence intervals) cover perhaps too wide a range of values to be particularly useful. It would be one of the objectives of a simulation to discover just how reliable could these corrected values be: reliable enough to be useful? There remain questions about how the bootstraps should be taken, given that the trips and sets rather than being random samples could be taken as givens. This would reduce the confidence intervals considerably. Once a simulation is in place it would be relatively simple to design an appropriate way of assessing the reliability of the corrected values.

Tim Lawson's correction

	skj	yft	bet
A	0.650	0.309	0.040
U	0.841	0.152	0.007

This method clearly produces plausible results. Though without an appropriate measure of reliability (and/or simulation results) it is difficult to be sure how good it is

It would be interesting to examine the methods and the data to establish why the methods differ from the spill estimates. It is likely to have something to do with the relative numbers of associated (292) and unassociated (102) sets since the unassociated set estimates seem to be the less successful. It may have something to do with the apparent divergence Skipjack showed from the other species in the grab bias plots.

The problem with sample size (numbers of fish) can be illustrated by looking at the effective grab sample sizes for the strata.

A	Original size=	18275	Effective size=	1480.8	Shrinkage=	0.081

U	Original size=	6641	Effective size=	361.8	Shrinkage=	0.054

The effective sample size for unassociated sets is 362 to estimate an uncorrected frequency distribution for 3 species each with 31 effective size classes between them. Corrected length frequency (or weight) calculations based on these figures are unlikely to be very reliable. The importance of an appropriate simulation program is underlined by this result, it needs checking. If it is true, it has implications for future sampling.

Conclusions

1) Both the Cordue method (with grab biased calculated pooling over species) and Tim Lawson's method both produce plausible corrected values
2) Without a simulation it is impossible to see which is the more reliable.
3) The effective sample sizes calculated for the strata strongly suggest that no method is going to work very well at anything other than the largest strata and largest scales. Simulations may clarify this.

Final conclusions and recommendations

7) The multinomial approach should be employed. Personally I have a feeling the modelling approach of Lawson might have some advantages if combined with a multinomial model. Basically the problems presented by the data seem to require exploitation of every opportunity to improve the quality of the predictions (smoothing, covariates, strata). Evidence for their utility would be from simulation.
8) However the model is developed, it is clear that robust and appropriate measures of the quality of the estimates must be put in place, validated by simulation
9) A comprehensive multipurpose simulation program is therefore essential.
10) The available data should be examined to establish what stratification is necessary and simulation studies performed to see if useful results can then be achieved.
11) The same analysis should estimate parameters so that the simulation can reproduce realistic scenarios.
12) At the design stage of the simulation, information essential for the useful implementation of the full simulation should be identified and if not yet available then, where possible, data should be collected. For example: the true magnitude of the layering effect in brails.

References:

Chih, Ching-Ping (2009): Evaluation of the Sampling Efficiency of Three OtolithSampling Methods for Commercial King Mackerel Fisheries, Transactions of the American Fisheries Society, 138:5, 990-999

Pennington, M., L. M. Burmeister, and V. Hjellvik. 2002. Assessing the precision of frequency distributions estimated from trawl survey samples. U.S. National Marine Fisheries Service Fishery Bulletin 100:74–81.