A HYBRID SPATIAL-AND-HAZARD-PROB MODEL FOR SOWER MINKE DATA

MARK BRAVINGTON, MAY 2011

1. Introduction

Neither the SPLINTR nor OK method is theoretically optimal for estimating Antarctic minke whale abundance from SOWER/IDCR data. In particular, SPLINTR uses a spatial model to compensate for imperfections of coverage and correlations between location, weather , but for g_0 it assumes trackline independence (AKA "point independence") which will lead to some degree of overestimation of g_0 and corresponding underestimation of abundance. Conversely, OK does not use a spatial model and its abundance estimates will likely be biassed high to some extent, but it does use a more flexible and potentially less biassed hazard-probability model[1] for g_0. This paper reports progress with developing SPHAZ, a hybrid model that uses a hazard-probability model for sightings within spatial models for school size and density.

Paper AE5 at the 2011 Intersessional Workshop (Bergen, 16-20 Jan 2011— referred to as "Bergen" hereafter) presented an empirical examination of hazard-probability in SOWER, using sightings close to the trackline of reported size 1, for which differences between the methods are likely to be greatest. The analysis was based on the idea of "duplicate trials": e.g. amongst cases where platform A saw the school in front of the boat and B had not seen it by then, how often did B in fact see the school subsequently? Although the empirical analysis is approximate, because it neglects school size error and has limited sample size, it is nonparametric (ie makes no assumptions about the form) and should provide a model-free check[2] on g_0. Unsurprisingly, the SPLINTR TLCI estimates of g_0 were higher than the empirical estimates, by about 25% for reported school size 1; more surprisingly, the OK g_0 estimates were about 25% *lower* than the empirical estimates.

The empirical analysis cannot be used directly in a full-likelihood setting such as SPLINTR, because it uses a nonparametric model of hazard probability, whereas full-likelihood models pretty much have to be parametric. The SPHAZ hybrid model in this paper uses the same parametric functional form for hazard-probability as OK, although the implementation details are somewhat different (Section 2). Section 3 presents preliminary results, including diagnostics. In particular, if the functional form chosen is insufficiently flexible to represent the "truth", then there may be bias in the g_0 estimates, and variants of the empirical-hazard diagnostics developed for Bergen are used to investigate this. The Discussion in Section 4 gives some comments on the implications for g_0 estimates in OK as well as SPHAZ itself.

2. SPHAZ formulation here

The hazard-probability model specifies the function $Q(r,\theta)$, which is the probability that a cue at radial distance r and angle θ will be seen by a given platform, given that the school which makes the cue has not been seen before by that platform. The SPHAZ model follows Appendix A of OK 2010 (SC/62/IA3):

$$Q(r,\theta) = 1/\left(1+\exp\left(\tau_r r^{\gamma_r} + \tau_\theta \theta^{\gamma_\theta} + \omega\right)\right) \tag{2.1}$$

where the parameters τ_r, τ_γ and ω are log-linearly related to Platform, School size, and "Z" (sighting-conditions covariate, in this case Sea.state=Good/Bad). In principle, the two γ parameters could also depend on these covariates, but I have followed OK in estimating a single overall value for each γ. (The γ parameters were introduced to alleviate a suggestion of misfit in diagnostics of observed and predicted radial distance in an earlier incarnation of OK.) The dependencies on covariates test here were:

$\log\tau_r$˜Platform + INC(Beauf,SS)
$\log\tau_\theta$˜Platform * Beauf
$\log\omega$˜Platform + INC(Beauf,SS)
$\log\lambda_Q$˜ INC(SS)

[1] IE a model for the probability of this event: given there is a cue at a particular distance and angle from the boat, and the platform in question has not seen the school before, will the platform see the cue? Platforms are assumed independent *per cue*, and cues are also assumed to occur independently, according to a memoryless Poisson process.

[2] That is: "model-free" subject to the independence-within-and-between-cue assumptions made by all (?) hazard-probability models.

where INC() is an increasing function . These are fairly similar to OK 2010, except there is no Vessel effect on ω, no SS effect on τ_θ, and several interactions (constrained to be increasing) where OK use additive terms.

The $Q()$ function must be used to compute the probability of the observed sighting history of each school, given that the school was seen at all. The sighting history is discretized into a number of forward-distance bins[3]. Each bin amounts to a mini-trial. When the school first enters the bin, some combination of platforms A, B, and C is "live" (i.e. not having seen the school earlier, and in the case of C, of neither A nor B having seen the school earlier either). By the time the school leaves the bin, some of those live platforms may have seen the school; the sighting history for that bin is {who was live at the start, who was live at the end}. Computing the probability of each trial outcome given the starting possibilities (AUBUC live; AUB live; A live; B live) is a straightforward albeit tedious exercise in Poisson-probability calculations. It is possible for both A and B platforms to see a school within the same distance bin regardless of whether they see the same cue or different cues; no attempt is made to label simultaneous duplicates in the analysis, but the possibility of a simultaneous duplicate is allowed for in the probability calculations.

Platform C requires special attention. C will only be live if no-one has seen the school yet. If either A or B sees the school within a bin, then C's sighting (if any) is disregarded, and the "live" platforms at the start of the trial are reduced to AUB. The reason is that C's near-simultaneous sightings cannot be assumed to be reliably recorded, even though there are specific codes in the datasheets; if a school makes several cues in quick succession— which would be treated as a single "megacue" for purposes of assessing simultaneous duplicates between A and B— then A or B may report an early cue in the set before C has a full opportunity to see the "megacue".

Since the function $Q()$ can change quite rapidly within a single bin (at least in formulation [2.1]), the probability calculations are broken down into several mini-bins within each bin (currently 0.05nmi, i.e. ~100m). This means that the overall probability calculation is computed as a finely-discretized sum, and is effectively the same as the integration in OK and other hazard probability models. To simplify calculations, all probabilities for all possible trials are first computed over a grid of perpendicular distances at 0.1nmi spacing; for the real data, the grid probabilities are interpolated to the actual perpendicular distance of each sighting using 3-point (quadratic) interpolation assuming symmetry about the trackline.

Other than the treatment of data from C (possibly), the main difference from OK's hazard probability model is the use of binned-forward-distance rather than time-in-front-of-abeam as the driving variable. The decision is partly to allow for measurement errors in distance and time, and partly to cope more cleanly with whale movements parallel to the trackline, which are substantial for at least *some* groups in SOWER. For example, in a time-based model, it is not obvious what to do with a school where the first sighting is 1nmi in front of the boat, and the second sighting occurs after a 10-minute gap (something that does happen in SOWER); the school should be comfortably past abeam long before the second sighting happens. In a forward-distance-based model, this causes no problem: whales that move towards/away from the boat parallel to the trackline simply have lower/higher cueing rates. The rate variability is actually ignored during estimation, since the Poisson-process variability in cueing which is assumed in hazard-probability models probably dominates any variability in the underlying rate, but at least there is no philosophical problem about impossible events.

The Appendix describes some remaining minor details of data-processing for a forward-distance-bin-based approach to hazard-probability, to do with possible simultaneous duplicates and with occasional schools that move *really* fast away from the boat.

Overall, though, the hazard-probability treatment in SPHAZ is fairly similar to that in OK.

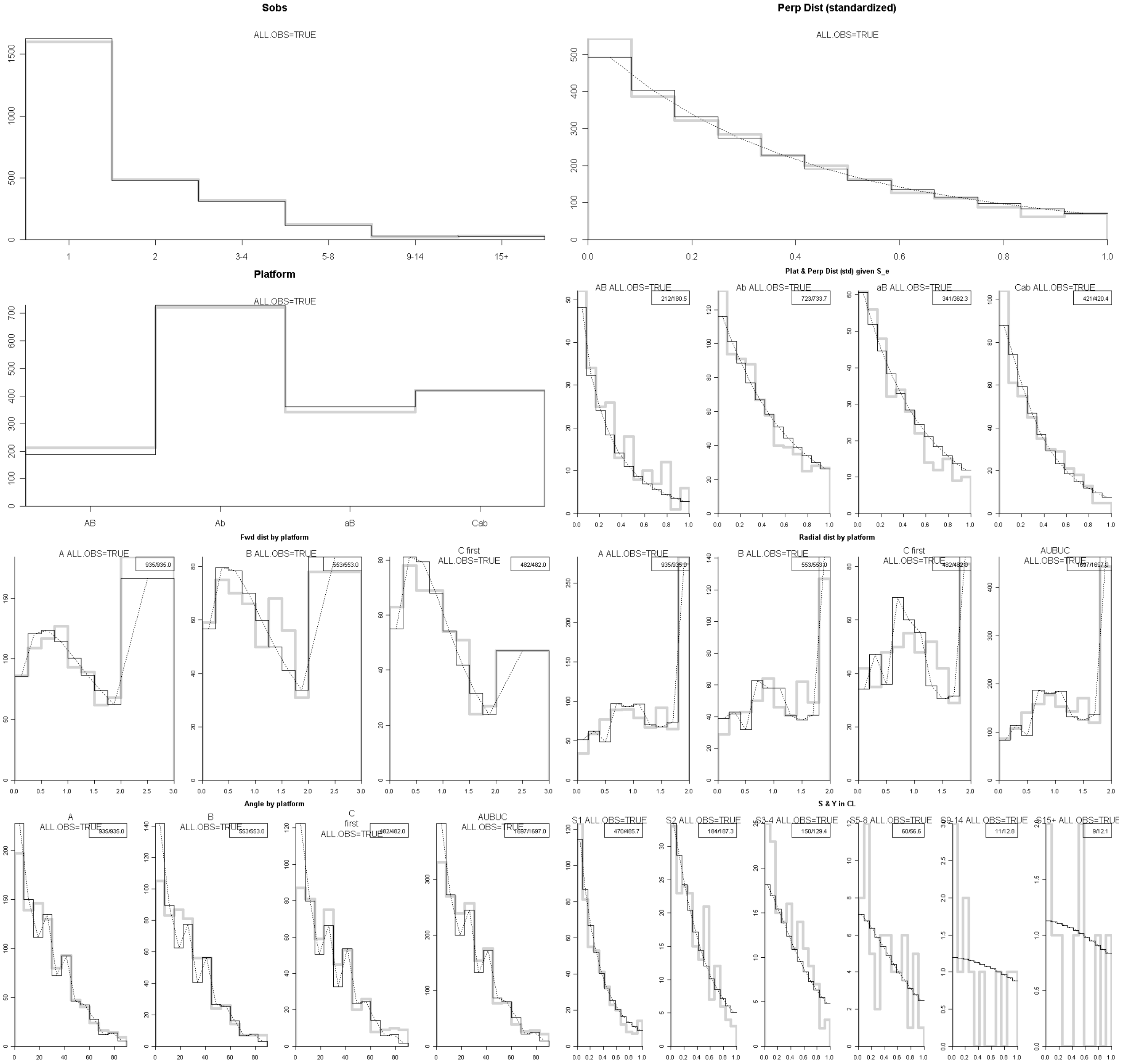
3. Results

The g_0 estimates from SPHAZ so far are slightly lower than those in OK (Table); note that the Beaufort categories used here (0-2=Good, 3-5=Bad) which is different from OK and from the results in the SPLINTR summary paper. I have computed but not extracted the abundance estimates yet for SPHAZ (I stoppped, because the diagnostics shown below look dubious), but presumably the estimates will be much closer to OK's because the g0s are similar or lower.

CP	Beauf	g0A	g0B	g0C	g0AUBUC
2	0-2	0.17	0.11	0.22	0.34
2	3-5	0.14	0.08	0.17	0.28
3	0-2	0.21	0.14	0.25	0.42
3	3-5	0.17	0.11	0.24	0.38

[3] For the results here, the forward bins were set at 0.25nmi out to 2nmi, then one bin between 2nmi and 3nmi, with any sightings claimed at >3nmi forward being forcibly moved backwards into that bin.

Figure 3.1. Standard diagnostics for SPHAZ. Only shown for CP3, to save space; CP2 looks very similar.



The standard diagnostics that have been produced for SPLINTR and OK in recent years look reasonable (plus a forward-distance diagnostic); the perpendicular-distance curves are more sharply peaked than in SPLINTR, which has (deliberately) flat shoulders, and are more similar to those in OK. However, the number of duplicates is underestimated by the model, by 8% (CP2) and 15% (CP3); since duplicate proportion is closely linked to g_0, this suggests there may be negative bias in the g_0 estimates and positive bias in the abundance estimates.

Also, a new set of diagnostics motivated by the empirical-hazard work in AE5 do not look so good. These diagnostics are “forward-truncated BT trials”; given that we know there was an group at forward distance x (because that is where it was first seen by one platform), we can compute the probability that it would be seen afterwards by a different platform, and then compare this with the “observed value” of whether it actually was seen. The idea is to remove the effect of simultaneous duplicates, and then check whether the proportion of delayed-duplicates is well-predicted. This is an independent check on g0; too few successful trials implies that g0 has been estimated too low. For this paper, the AE5 diagnostic has been adapted to include all perpendicular distances and to allow for school size error when calculating expected values.

Figure 3.2 shows the results, for different variants of the trials depending on who (ie which platform) saw it first and who was subsequently “on trial”. The key columns are the 3rd-from-left in the left-hand graph, and both columns in the right-hand graphs. There are substantial underestimates of the number of subsequent sightings in many of the columns, particularly in the columns AUB|C. This suggests to me that the g0 estimates in SPHAZ are generally too low. Of course, it is not just g0 here, but the whole sighting model across all perpendicular distances. It is notable that the overall proportion of duplicates is nevertheless predicted well by SPHAZ. It seems necessary to look at platform C as well as A/B duplicates in this respect, since the "ignore-C" test (left-hand two columns in the graphs with four columns) look better than the tests when C is included.

The biasses are quite variable across trials, but there is a persistent underestimation overall, which implies that g0 is being underestimated by a corresponding amount. The overall abundance estimate will presumably therefore be overestimated, although it is hard to say by exactly how much because there are so many interlinked aspects in models for SOWER data. The levels of bias in the key trials in the figures below are as follows (showing predicted vs observed; %increase of observed over predicted):

CP2:
AUB|C: 49.8 vs 64 (+29%)
A|BUC: 81.3 vs 90 (+11%)
B|AUC: 95.1 vs 106 (+11%)
CP3:
AUB|C: 49.2 vs 56 (+14%)
A|BUC: 67.7 vs 66 (-3%)
B|AUC: 77.5 vs 103 (+33%)

4. Discussion

TLCI is certainly not perfect for SOWER minke whales. Hazard-probability models for whale sightings are not perfect either; to make a hazard-probability model computationally tractable, it seems to be essential to make assumptions about independence (about the memoryless nature of cue generation, and about the probability that different platforms will see a cue given there is one, which implies that all cues for a given school are intrinsically similar). In SOWER, there are also potential complications arising from large measurement errors in distance and/or time. So, when deciding "which g0 to use", it is important to check as far as possible whether hazard probability is really working well.

The results in this paper suggest that it is possible to get a good fit in terms of the **standard** SOWER diagnostics, but still to fit poorly in terms of the new diagnostics, which should directly reflect on g0. Perhaps the biggest difference between this paper and AE5, which included nonparametric "empirical" hazard-probability estimates of g0s that were considerably higher than in this paper, is that a *parametric* functional form for hazard probability (using the same model formula as OK) has been used here. It is possible that the particular implementation of hazard-probability in SPHAZ is causing a misfit, but I suspect that the problem is more in the strongly-peaked functional form for $Q()$ imposed by equation (2.1). For binocular-based surveys that try to look far away from the boat, the hazard probability should not peak strongly near the boat.

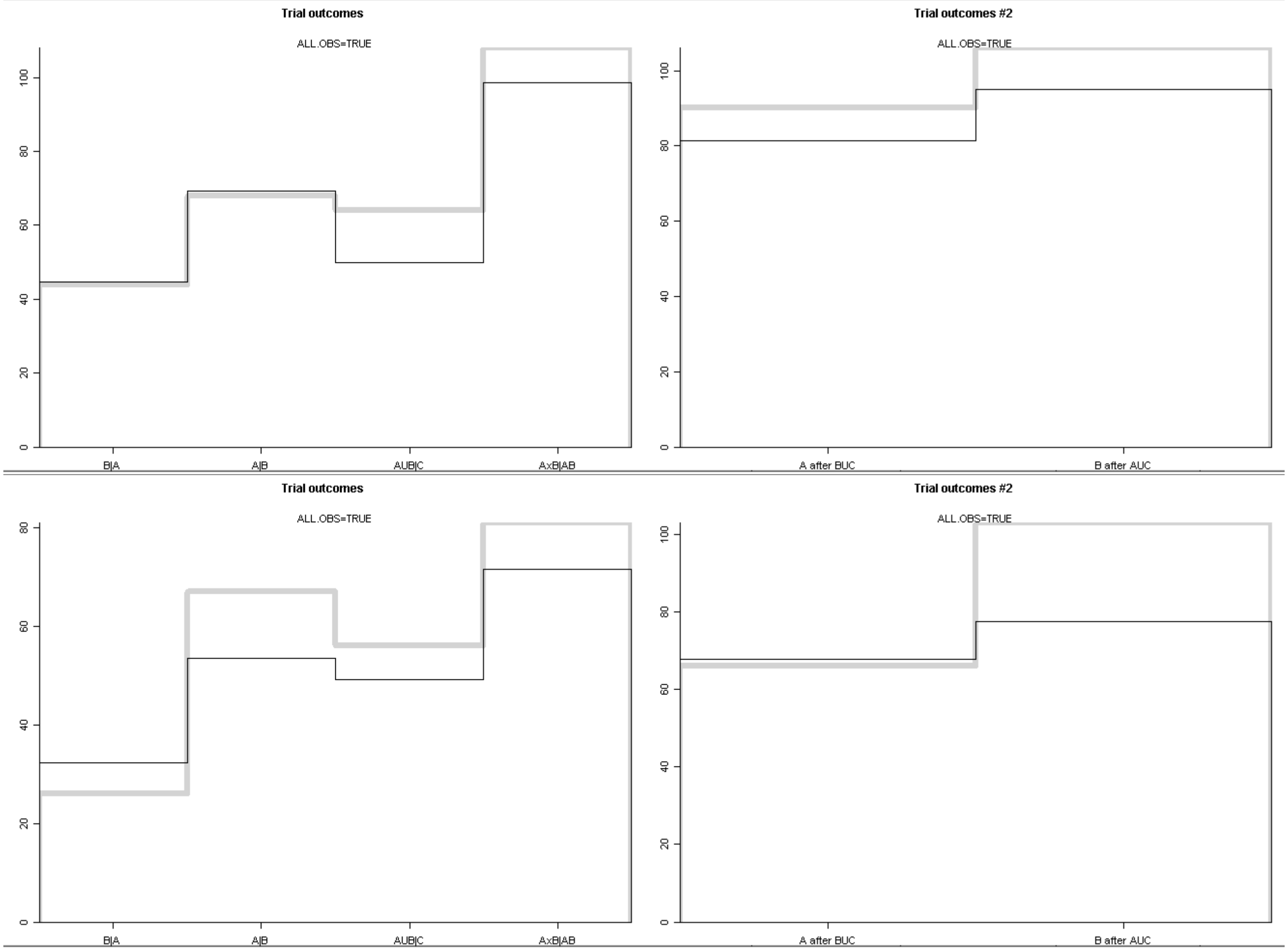
When it comes to "choosing a g0 method" for SOWER minke whales, there are several lines of evidence to be considered, apart from the standard diagnostics.

(1) BT experiment results: $g_{0A} \approx 0.25$ overall (0.31 in good conditions and 0.2 in poor conditions). (though BT assumptions not fully met)
(2) Empirical hazard results from AE5 for school size 1. These may be noisy but should be unbiassed because they do not require parameteric assumptions about functional forms, although they do require sightings within a . TLCI (SPLINTR) estimates are about 25% below Empirical Hazard, OK about 25% above.
(3) Lack of fit of SPHAZ's version of hazard-probability model to new diagnostics
(4) Simulation performance (but do the simulations capture enough real-data nastiness in this respect?). Note that the cueing interval in the simulated datasets (90s) is much shorter than we

Appendix: Pre-processing of duplicate sighting histories

(1) For sightings within 60s, the location estimates for both platforms are replaced by their average (forward and perpendicular). The 60s=0.2nmi of boat travel, so the two sightings "belong" in the same forward-distance-bin, whether truly simultaneous or not.
(2) Each platform's sighting is allocated to a forward-distance bin.
(3) If the first sighting in one forward bin and a subsequent sighting is in a more distant bin (i.e. the whale has apparently outrun the boat), then the bins of the two sightings are swapped, but the time order is preserved (otherwise there is confusion about censoring of platform C). Although a forward-distance-bin model is quite accommodating with respect to whale movement, whales that temporarily or permanently outrun the boat are a philosophical and practical headache. The solution adopted here is ugly but practical;

FIGURE 3.2. Empirical-hazard diagnostics. First two graphs are CP2, second two are CP3.. In the 4-column graphs, the 3 leftmost columns show observed (grey) and predicted (black) numbers of subsequent resights. For example, the leftmost— "B|A"— ignores C, and considers cases where A saw it first and B did not see it in A's bin or any further bin; the question is whether B subsequently saw it in a closer bin. [Ignore the rightmost column, which is something different: how many AB-duplicates fall into the same forward bin?] The 2-column graphs are similar, but the left column now includes all cases where either A-or-C sees the school before B; the trial is whether B subsequently saw the school.



it preserves the duplicate information, but introduces some additional "distance error". Luckily, these cases are quite rare (after binning).

(4) The same rules are applied to all years, whether from the early part of CP2 with coarse timing, or from later with more accurate timing. Note that the "accurate" timing is still subject to potential delays in reporting; with SOWER, measurement errors occur in time as well as distance.