Abundance estimation of Antarctic minke whales via spatial modelling

Mark Bravington[1] and Sharon Hedley[2]

Contact email: markb@summerinsouth.net

ABSTRACT

This paper describes our work on Antarctic minke whale abundance estimation using line-transect survey data collected between 1986 and 2010 as part of the IWC's IDCR/SOWER programme. While we focus on spatial modelling (producing smooth 'maps' of animal density), much of the detail is concerned with allowing for imperfect detection of whales on the trackline ($g_0 < 1$). We present an account of operational survey details, along with full descriptions of the bespoke methodological minutiae that have to be considered when trying to account properly for survey protocols and Antarctic minke whale behaviour. This paper describes the reality of advanced distance sampling in a complicated setting. A complete mathematical background is given for the spatial models of both school density and school size; for the SOWER-specific adaptation of Trackline Conditional Independence which we developed for dealing with g_0; and for how all these parts of the overall process can be linked. We present some of our own estimates, alongside the IWC's agreed consensus estimates, in part based on our results. The paper ends with comments on interpreting the estimates, the remaining unsolved problems when analysing these particular data, and lessons for future whale abundance surveys.

KEYWORDS: SOWER; ABUNDANCE ESTIMATE; SCHOOL SIZE; SURVEY – VESSEL; SOUTHERN OCEAN; DISTRIBUTION

1 INTRODUCTION

This paper describes a spatial-modelling approach developed to estimate the abundance of Antarctic minke whales (*Balaenoptera bonaerensis*; here AMW). The data come from annual line-transect surveys conducted under the auspices of the International Whaling Commission (IWC). The series of cruises between 1979–1996 were known as the International Decade of Cetacean Research (IDCR) surveys, and subsequently as the Southern Ocean Whale and Ecosystem Research (SOWER) surveys. We use the term SOWER for both. Each year, during about eight weeks over the austral summer, the survey covered a longitudinal region of up to about 60° of the Southern Ocean and generally south of 60°S, resulting in three circumpolar sets of surveys, hereafter referred to as CP1, CP2 and CP3. The design and operation of the CP1 surveys differed substantially from CP2 and CP3 and is not amenable to the same analysis. Hence, we only provide estimates from CP2 and CP3, i.e., from 1986 onwards.[3] A recap of operational details is given in Section 1.1. Full details may be found in Matsuoka *et al*. (2003).

The first AMW abundance estimates from SOWER data used the 'standard method' of distance-sampling (Branch, 2006). The 'standard method' was generally expected to underestimate AMW abundance because it assumes that $g_0 = 1$ – i.e., that detection is certain for schools on the trackline – which is clearly not the case for

[1] Estimark Research, Hobart, Australia
[2] St Andrews, Fife, Scotland, UK
[3] We use a single year to refer to each cruise. The vast majority of survey effort took place in the months of January and February, although the cruises actually left port during December of the previous year. Thus, our '1986 cruise' began in December 1985 and continued into February 1986. Other papers on SOWER may use split-year nomenclatures, e.g., '1985/86'.

small schools of AMW, although it may be reasonable for larger baleen whales. Indeed, double-platform protocols were introduced in CP2 deliberately to permit estimation of g_0. Nevertheless, 'standard method' estimates might have been deemed adequate (for the IWC's purposes, such as to inform the Revised Management Procedure's Catch Limit Algorithm) had there been little apparent change between CP2 and CP3 – i.e., from the perspective of assessing risk of depletion from catches, one single (presumably) negatively-biased estimate is conservative.

However, as CP3 progressed, it became clear that abundance estimates from the 'standard method' in CP3 tended to be much lower than for comparable regions in CP2, often under half (Branch, 2006).[4] Such a large decline, if real, would have major implications for management. But how much of the decline was real, and how much was just an artefact of changing conditions and oversimplified analysis? There were at least two factors that had changed, and which might have introduced substantial artefacts. First, school sizes were typically smaller in CP3, with a higher proportion of singletons; g_0 is lower for smaller schools, so the bias from ignoring g_0 in the 'standard method' might have worsened. Second, the spatial coverage of the surveys had (deliberately) been changed considerably, with more effort far from the ice out to 60°S; for AMW, proximity to ice strongly affects the distribution of school size, the density of schools, and the quality of sighting conditions. As a result, the extent of bias from the 'standard method' might have changed substantially during CP2 and CP3. Attempts to gain any further insight were unsuccessful due to the sheer complexity of SOWER data and AMW behaviour. It became clear that the only reliable way forward would be to develop more sophisticated models capable of reducing those biases. Two advances in particular were desirable:

- Estimation of g_0;
- Spatial modelling of school density and school size, to account for uneven coverage.

As explained in the next section, the SOWER CP2/CP3 data for AMWs exhibit most of the difficult features that are encountered (and often ignored) in other distance-sampling datasets for cetaceans.[5] Since it is also a very large dataset collected with standardised protocols, it can be used to test assumptions which cannot be investigated properly in smaller datasets. Thus, it provides a good testbed for developing reliable methods that can be applied more widely.

In theory, the SOWER AMW data do have the necessary prerequisites for addressing g_0 and for spatial modelling:

- Independent observers;
- Two survey modes, interspersed across the whole region:
    - 'Passing-type' mode, which should give unbiased estimates of local encounter rate;
    - 'Closing-type' mode, which should give unbiased estimates of local school size for encountered schools.

However, in reality, matters turn out to be more complicated. In this Introduction, we give a general description of relevant SOWER operations, followed by a list of the main challenges for AMW abundance estimation, and, finally, a synopsis of our own involvement in the IWC process leading towards 'agreed abundance estimates' in 2013.

Subsequent sections adopt a primarily methodological viewpoint, addressing general issues of spatial modelling and advanced distance sampling, as well as issues more specific to SOWER and AMWs. Section 2 describes the basic components of our SPLINTR (SPatial LIne TRansect) model, followed, in Section 3, by an explanation of SOWER-specific data choices and implications for model formulation. The results in Section 4 are fairly brief, concentrating on spatial and detection-function aspects, since the IWC's agreed abundance estimates

[4] By the end of CP3, the 'standard method' estimates of pan-Antarctic abundance were 786,000 (nominal CV 9.4%) in CP2, and 338,000 in CP3 (nominal CV 7.9%).

[5] With two exceptions: large school sizes such as for ETP dolphins; and long dive times that make some schools completely unavailable for detection.

are already well covered (Palka *et al*., in prep). Finally, in Section 5, we review the features that make SOWER AMW analysis so difficult; consider whether AMW detection-probability estimates for SOWER could ever be improved; comment briefly on the implications of time-trends in the estimates (Kitakado & Okamura, 2009); propose some lessons for the design of future LT/DS abundance surveys; and describe how methods developed in SPLINTR have subsequently percolated into general applications of spatial LT/DS.

Although this paper is designed to be fairly self-contained, so that it could be read by anyone interested in spatial-modelling or distance-sampling technicalities but without a specific background in SOWER, it does presuppose a knowledge of distance sampling (e.g., Buckland *et al*., 2001) and modern statistical methods (e.g., Wood, 2006). Table 1 contains a glossary of SOWER-specific and distance-sampling terms, and Table 2 summarises the mathematical symbols and notation.

1.1 A recap of SOWER CP2 and CP3 operations

An extensive description of all the SOWER surveys, and any year-to-year variations, is given in Matsuoka *et al*. (2003). Here, we summarise informally the basic CP2 and CP3 set-up and AMW biology as relevant to abundance estimation.

SOWER collected data on all whale species, but the primary focus was abundance estimation of Antarctic minke whales. These are often found as singletons, but (unlike minke whales in the North Atlantic) equally often

Table 1 – Glossary

These are the main acronyms related to survey details and distance-sampling. Standard statistical acronyms are omitted.

LT/DS	Line Transect and/or Distance Sampling.
AMW	Antarctic minke whale, *Balaenoptera bonaerensis*.
IDCR	International Decade of Cetacean Research: the programme responsible *inter alia* for the first two sets of Antarctic circumpolar cruises CP1 and CP2 until 1992.
SOWER	The Southern Ocean Whale and Ecosystem Research programme, successor to IDCR, responsible for the third set of circumpolar cruises and several years of follow-up cruises until 2010. We generally use 'SOWER' to refer to all the IDCR/SOWER cruises.
MA	'Management Area': a 60° block of longitude traditionally used by the IWC to subdivide abundance estimates. MA1 is immediately west of the Antarctic Peninsula, and the numbering runs in an easterly direction.
CP1/2/3	The three Circumpolar cruise series within IDCR/SOWER, each being one circuit of the Antarctic, intended for AMW abundance estimation. CP1 did not yield suitable data for modern DS analysis. CP2 ran from 1986–91 covering one MA per year, and CP3 from 1993–2004 covering about half an MA (30° of longitude) per year, from the ice edge out to roughly 60°S latitude. Note that there were several other IDCR/SOWER cruises that were not part of the CP series.
SSX	School Size eXperiment (SSX) surveys (also known as SSII and SSIII), part of the 2007–09 SOWER cruises.
VDT	Video Dive Time experiments, part of the 2005 SOWER cruise.
BT	Buckland-Turnock (Buckland & Turnock, 1992) approach to estimating g_0 requiring different protocols to SOWER.
IO/CL	Independent Observer mode (B-platform in use, passing-mode) and Closing-mode (without B-platform).
IDEDD	Independent Dives with Exponentially-Distributed Durations (the typical assumption behind cue-based estimates of detection probability).
ESW	Effective Strip Width, including the effect of g_0.
NESW	Nominal Effective Strip Width, without g_0: thus ESW = NESW × g_0.
TCI	Trackline Conditional Independence (for school-based, as opposed to cue-based, detection probability estimation).
HP	Hazard Probability (in the context of cue-based detection functions). Note that this is completely unrelated to 'hazard-rate models' in the context of school-based detection functions.
SPLINTR	SPatial LINe Transect – the spatial-surface (non-stratified) school-based model in this paper, consisting of the following two sub-models:
SPAMASSS	Sighting-Probability-And-Misunderestimation-And-Spatial-School-Size: combined model incorporating probabilities of sighting conditional on true school size, probabilities of recorded school size given true school size, and a spatial model for the distribution of true school size.
DOSS	Density of Schools, Spatially: spatial model for density of schools, regardless of size.
OK	Okamara and Kitakado's stratified cue-based HP model (Okamura *et al*., 2005)
IM	Integrated Model; Cooke's approach to SOWER AMWs which included both spatial modeling and cue-based detection probability (Cooke, 2008).
SPHAZ	An unfinished HP variant of SPLINTR explored in the terminal stages of the IWC's quest for AMW abundance estimates.
dsm	Density Surface Modelling: Widely-used R software for fitting detection functions and spatial smoothers (Miller *et al*., 2013; Miller, 2025), building on the ideas of DISTANCE itself.

Table 2

Main mathematical notation. Except as otherwise noted: capital letters denote random variables, with the lowercase version denoting an observed value; other lowercase letters are observed covariates of a sighting or location; and Greek letters are parameters to be estimated, or functions of such parameters. Subscripts are used as required to relate these symbols to platforms, sections of the trackline, and individual sightings.

Symbol	Meaning
$\mathbb{P};\mathbb{E};\mathbb{V};\mathbb{C}$	Probability; expectation (mean); variance; covariance.
$\mathbb{I}$	Indicator function; for any possible event e, $\mathbb{I}[e] = 1$ if e actually did happen, or 0 if not.
$A,B,C/a,b,c$	The three observation platforms (breaking the capital-letters convention).
h	Who saw the school/cue, i.e., which combination of A,B,C. Specific values of h use upper/lower case to indicate whether a platform did/did not see the school, e.g., abC means 'A & B did not, but C did'.
o	The fact that *someone* saw the school, i.e., that $h \neq \emptyset$.
y	Perpendicular distance.
x	Location (rescaled latitude and longitude).
z	Environmental covariate related to sighting conditions, e.g., Beaufort scale.
m	Survey mode (IO or CL).
g_0	Probability of seeing a school that is on the trackline.
p	Local school density, per nmi^2.
w	Strip width, i.e., twice the perpendicular truncation distance.
$\boldsymbol{\ell}$	Length of a snippet of effort.
N,n	Number of schools seen in a snippet (not abundance, nor 'sample size').
A	Total abundance within some region (Appendix B).
S	True school size (category).
S_e	Estimated school size.
anything^{ss}	Superscript SS means that 'anything' relates to spatial school size, rather than to school density or some other part of SPLINTR.

in schools of 2 or more, rarely up to 10 animals, and occasionally in larger numbers. While we use the term 'school' throughout, AMW schools are not tightly associated and may simply reflect temporary associations while feeding. Nevertheless, there is enough synchrony and adjacency in cueing to make 'schools' the most meaningful unit for distance sampling.

The surveys generally operated using two very similar vessels, the *Shonan Maru* and the *Shonan Maru No. 2*. Survey strata were defined as 'Southern' and 'Northern', and usually one vessel operated in each stratum, although during the course of a survey, each vessel would survey sometimes in the north and sometimes in the south. The southern boundary of the southern survey strata was defined by the 'estimated ice edge' – a logistical boundary beyond which the vessels (which were not ice-strengthened) could not safely navigate. The great majority of survey effort was in open water, or occasionally in low ice concentrations of up to 10%. Estimates of AMW abundance from the SOWER surveys only relate to animals in open water at the time of surveying. However, unlike the 'great whales', AMWs are also known to be commonly present in higher ice concentrations, which were not surveyed. See Herr *et al.* (2019) and references therein.

The survey transects were planned beforehand to follow a zig-zag design, intended to ensure good coverage throughout each survey. If sighting conditions became too bad, the survey would pause until better weather arrived, so that, in principle, there were no gaps in coverage. The transect locations were not strictly randomised – e.g., the start-point for the northern strata effort was almost always from a corner of the stratum – although some degree of 'randomness' was introduced by the start-point of the southern stratum effort being at the ice edge, a constantly moving feature. The boundary between northern and southern strata, determined by the latitudinal width of the southern stratum, was also affected by the ice-edge location. Despite intentions to the contrary, the realised survey coverage was sometimes very uneven. In such situations, design-based estimates reliant on equal coverage probability cannot be assumed to be unbiased.

The basic statistical and logistical challenges for distance-sampling AMW surveys have been recognised since at least the CP1 series:

1. To estimate AMW school size reliably, and to confirm species ID, it is necessary to 'close on' (i.e., get fairly near to) the school and focus on it for some time, which, for most sightings, means deviating substantially from the trackline and abandoning normal search protocols during the closure attempt;
2. However, reliable estimates of school density (encounter rate) can only be obtained if the vessel does *not* interrupt normal search protocols;
3. Detection on the trackline is not certain, especially for small schools (e.g., singletons).

The overall strategy in CP2/3 for dealing with these challenges was to use multiple sighting platforms and two different survey modes: Closing (CL) and Independent-Observer (IO). Broadly, the purpose of CL-mode was reliable estimation of mean school size (and species), while the purpose of IO-mode was unbiased estimation of g_0 and encounter rate. (It is well-known that encounter rates in CL-mode are generally biased, and the direction of bias may vary depending on the nature of the local clustering (Haw, 1991) as well as on operational details of closure.) Each vessel would typically alternate between CL and IO-mode, spending several hours in each.

There were three sighting platforms on each vessel (Fig. 1): the 'Top' Platform A in the barrel with two observers; the Independent Observer ('IO' or 'IOP') Platform B located on the mast but below the barrel, with one observer; and the Front or Upper Bridge (Platform C), with between two and six observers searching, where all sightings data were recorded. Searching, distance/angle estimation, and subsequent tracking of sightings used reticule binoculars and angle-boards. AMWs were often seen up to 2 nmi ahead of the vessel (about 10 minutes' transit), and sometimes up to 3 nmi. Larger species could be seen further away. AMW dive-times are – or at least were thought to be – usually no more than a couple of minutes, so most AMW schools should present several sighting opportunities prior to passing abeam. However, the details of AMW dive-patterns eventually turned out to be both complicated and very important for abundance estimation (see Discussion for more details).

In CL-mode, Platform B was not used, but Platforms A and C operated together with full communication. As soon as a sighting was made by either platform, the vessel immediately changed course towards the sighting, in an attempt to close on the school to confirm size and species. The confirmation attempt could typically take between 0–15 minutes and was usually successful; however, in about 15% of cases, the school was 'lost' before its size and species identification (or in the case of mixed-species schools, composition) could be confirmed.

In IO-mode, the vessel remained on its designated course regardless of any sightings, as for the typical passing-mode of distance sampling. All three platforms operated: Platforms A and B searched independently of each other. Platform C was informed of sightings made by A or B but not vice versa. If the sighting was made from either Platform A or B, then Platform C was informed and its primary job changed to locating and tracking that sighting, until either the school was judged to have passed abeam, or it was judged to have been seen by both Platforms A and B, with the 'duplicate status' being determined by the Platform C personnel. During such a time, all platforms, including Platform C, were still able to make new sightings, but the primary role of Platform C became the assessment of duplicate status. It was also possible to record a 'triplicate' sighting in a similar manner, provided that a sighting was seen first by Platform C (and subsequently by Platform A, then B, or Platform B then A). Occasionally, the duplicate status was uncertain, i.e., when two sighting-cues well-separated in space and time may or may not have been of the same school.

The protocol for recording school size in IO-mode was to record only the minimum school size that the observer(s) were collectively sure was present, i.e., not to infer unseen whales. While designed to minimise subjectivity and inter-observer variability, this leads to underestimation for which allowance needs to be made during analysis.

Although AMW abundance was the main focus of SOWER, many other species were of course encountered. Most cues seen were blows rather than body cues, and it is hard to reliably distinguish between the smaller Antarctic whale species on the basis of a distant blow. Thanks to the tracking protocols in IO-mode, most sightings within a perpendicular truncation distance of 1.5 nmi (as used in these AMW-focused analyses) were confidently identified to species, and the CL-mode data prove that, in any case, AMWs were by far the most common of the smaller species with similar cues. Some uncertainty remains about the true identity of 'like-minke sightings' in IO-mode.

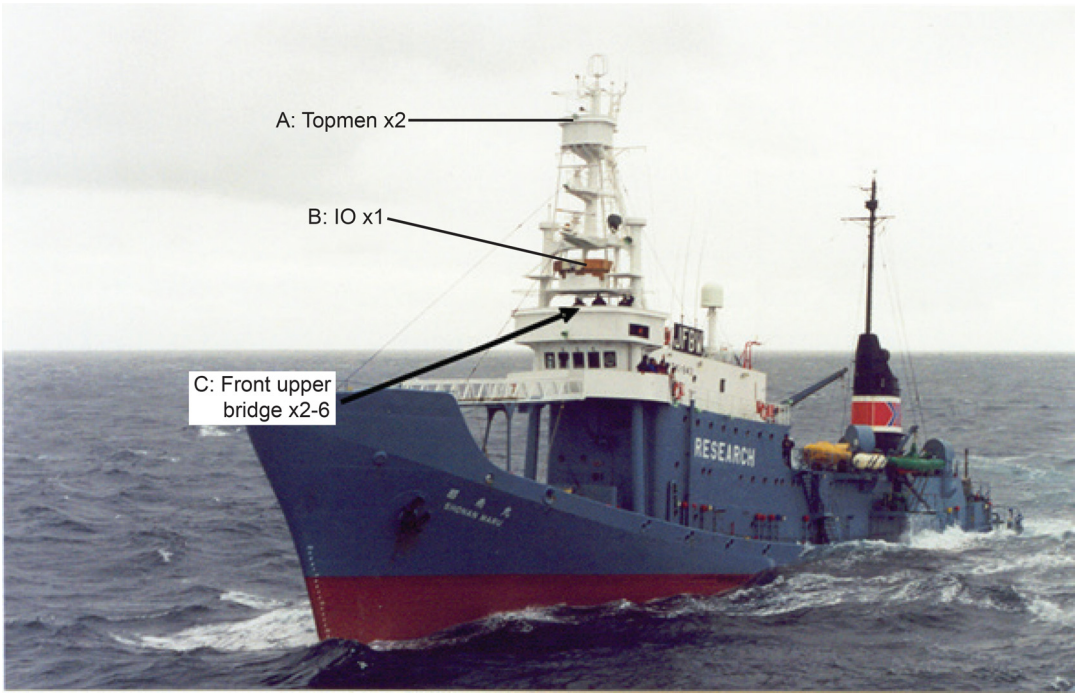

Figure 1. Platforms and observers on the *Shonan Maru*.

1.1.1 Experiments

As well as the standard survey data, all SOWER cruises conducted some effort in Experimental mode, and some years were entirely experimental. Routine experiments included biopsy sampling, photo-identification, collection of resightings data in IO-mode, and the Estimated Angle and Distance Experiment (Matsuoka *et al*., 2003). Of particular relevance to this paper, however, are the school size experiments (SSII and SSIII, collectively called SSX here) from 2007, 2008 and 2009.[6] The aim was to investigate the extent of school size uncertainty in IO-mode, when no attempt is made to close and observers are instructed to report only the number of whales that they are confident are present. In SSX, standard IO-mode protocols (including the reporting of minimum definite school size) were followed until the sighted school was judged to have passed abeam of the vessel, from which point no further sightings of the school would be recorded in normal IO-mode. At that point, a closure/confirmation attempt was made, to determine true school size. The pre-closure estimates used almost exactly the same process as in IO-mode and the post-closure estimates were ground-truthed on a per-school basis, so the SSX data provide direct and indispensible insight into IO-mode school size uncertainty. The only difference from normal IO-mode was that, in most of SSX, Platform B was not operating for logistic reasons; since most of the sighting data comes from Platforms A and C, the hope is that this did not greatly affect school size uncertainty compared to normal IO-mode.

Two further cruise experiments (Video Dive Time, VDT, and BT-mode, BT) were key to the diagnostic process and to the development of an agreed estimate, as described in Palka *et al*. (in prep). We did not use those data explicitly in SPLINTR, but we did take account of them qualitatively in the way that we developed the model.

1.2 Challenges for AMW abundance estimation

Although methods do exist to estimate g_0 from double platform line-transect data, such as the IO-mode data on the SOWER surveys (e.g., Skaug & Schweder, 1999; Laake & Borchers, 2004), there are some uniquely challenging features both of SOWER data collection protocols and the Antarctic itself, which render straightforward application of existing approaches unreliable, particularly when examining trend in abundance. Together with estimation of g_0 (and in some cases linked to its estimation), the main challenges are as follows:

[6] School-size experiments were first tried in 1984/85 and 1985/86, but because of protocol problems (Section 3.1), these early data are not considered reliable. We did not use them in any analysis.

1. The average school size varies in space (even within the designed strata) and has decreased over the decades of CP2 and CP3;
2. The density of schools varies in space (even within the designed strata) in a way that is often correlated with average school size;
3. The average sighting conditions vary in space (even within the designed strata) in a way that is correlated with both average school size and school density. Near the ice edge, conditions tend to be better, schools tend to be larger, and density is higher. For example, this means that non-covariate-based estimates of effective strip width, which are based on all observations, end up biased towards conditions close to the ice, where more sightings happen to be made due to better weather;
4. The coverage within a stratum is often very uneven despite good intentions in the survey design, because weather cannot be controlled, and the ice edge location is highly variable in some regions;
5. Similarly, existing line-transect spatial model-based approaches (e.g., Hedley & Buckland, 2004) that could be applied when design-based estimates might be considered unreasonable suffer in practice from the following issues of their own:
    a. Where survey coverage is poor near the edge of the survey region, particularly towards the corners, typical smoothers tend to extrapolate linearly. This is biologically unreasonable, and, since smoothers typically describe log-abundance rather than abundance, any increasing linear trend has a disproportionate impact on the abundance estimate, resulting in positively biased and imprecise estimates;
    b. Where there are complex survey boundaries, e.g., where a narrow peninsula of ice or land sticks out into the middle of a body of water, data from one side of the peninsula leaks across to the other side of the peninsula because the two sections of water are close by a simple distance metric ('as the crow flies') but not by a more reasonable biological distance metric ('as the whale swims');
6. g_0 and nominal[7] effective strip widths strongly depend on school size (as well as sighting conditions):
    a. In IO-mode, g_0 could be estimated from the independent platforms, except that true school size is frequently underestimated (a deliberate design feature of the protocols, to minimise subjectivity), which precludes direct estimation of g_0 conditional on true school size. Unfortunately, g_0 conditional on *recorded* size is not directly useful for abundance estimation;
    b. In Closing mode, school size is reliable, but there is no independent-platform data to estimate g_0;
7. A different combination of platforms operate in IO and Closing modes. Platform C is one-way dependent in IO-mode (so there are no data to determine whether Platform C would have seen a school initially sighted by A or B), and in Closing mode, the platforms are not independent at all (closure begins immediately upon a school being sighted – there are no data on whether the other platform would have eventually seen the school);
8. The data records do not include independent estimates of school size or identification of species by platform. Rather, they are a 'joint effort' between observers on the platforms that sighted the school;
9. Schools are somewhat clumped spatially on a small scale;
10. There are substantial measurement errors in SOWER distance and angle estimates (and to some extent in the recorded times of sightings), as determined from simultaneous duplicate sightings. As well as complicating (and causing mistakes in) duplicate identification, such errors also affect detection probability estimation and are known to lead to systematic bias in abundance estimates in some situations (Borchers *et al*., 2010).

[7] Nominal effective strip width, i.e., unadjusted for $g_0 < 1$.

In addition, there were changes over time in observer habits and experience (Mori *et al*., 2003), and in platform set-up – e.g., considerable structural modifications were made to the *Shonan Maru* platforms before the 1998/99 survey; a new IOP platform was installed on the *Shonan Maru No. 2* at the same time, and further modifications were made prior to the 1999/2000 survey (Matsuoka *et al*., 2003). We did not try to deal with these issues.

1.3 Synopsis

We became interested in this problem because of our previous (separate) work on spatial LT/DS models for whales. Crucially, SLH also had extensive field experience, including as a researcher (observer) on two SOWER cruises. In 2001, notwithstanding our own prior work, there was no generally satisfactory approach to LT/DS spatial modelling for whales. We thus had two overlapping aims: to develop a reliable AMW estimate for the IWC; and to develop spatial-model-based abundance estimation methods that could be applied more generally.

A key part of our interest in spatial modelling lay in its ability (in principle) to produce almost unbiased point estimates and variance estimates for the whole surveyed region[8] and any subregion of interest, regardless of how uneven the actual (compared to planned) coverage turned out to be.

Conventional stratified design-based LT abundance estimation appears simple in theory but, in practice, can suffer from numerous complications with respect to point and variance estimation. First, realised coverage may not correspond to intentions, creating a conundrum since the variance and bias properties are established statistically on the basis of intentions (i.e., randomised design). Second, even when realised coverage does correspond to planned, variance estimation for the whole region often entails post-stratification (*post hoc* merging of strata whose achieved sample sizes are individually too small) – a rather arbitrary process where different choices can have a big impact on variance calculations. Third, there is no coherent statistical rationale within the design-based paradigm for making point estimates within arbitrary subregions; of course, people devise *ad hoc* solutions which may seem reasonable in specific cases, but, as with post-stratification, 'adhockery' is the enemy of coherent variance estimation. Fourth, the nice-sounding properties of 'design-unbiasedness' and 'design-based variance', which arise from the randomisation step in the conventional survey design process (if indeed it really has been randomised – a condition which, if it applies at all, does not generally apply to subregions) are only repeat-sampling properties. This means that the average abundance estimate over many repeated surveys of the same region with the same population but varying randomised designs would eventually converge to the true value; and the average of their 'internal' variance estimates would converge to the true variance (i.e., between one survey's abundance estimate and the truth). However, these repeat-sampling properties are scant consolation if, as in SOWER, one has to make inferences based on a single replicate, whose actual coverage through the (sub)region-of-interest may be quite uneven.

Spatial modelling does away with those problems by making conditional probability statements (via means and variances) about the likely values of true abundance in any subregion, given whatever tracklines actually happened and whatever was seen. The process is complicated but automatic (see below for more details). In practice, the statistical tools used to implement the spatial model must be well-chosen. Late 20th Century formulations were not very robust (Section 2.1), so that a lot of work was needed to actualise the potential benefits of spatial modelling for LT/DS.

Over the course of a decade, we developed the spatial abundance estimation method SPLINTR (SPatial LINe TRansect) which attempted to deal with the issues listed above. To model detection probabilities, we used the perpendicular-distance-only school-based approach of Trackline Conditional Independence (Laake, 1999), adapted to the SOWER platform setup. We expected that TCI should do an acceptable job of estimating g_0 given SOWER sighting properties and AMW behaviour. Meanwhile, an alternative method – OK – came from another

[8] Strictly speaking, 'the surveyed region' is just a narrow strip of twice the truncation distance around the cruise track. Any useful approach to abundance estimation entails extrapolation from that 'strictly surveyed region' to some much larger region of interest. Classical (design-based) approaches require *a priori* definition of the notional region within which tracks might have been placed according to random selection. In practice, the tracks actually get placed in specific locations, and parts of the notional region may end up a long way from any track. The validity of the notional region as the unique definition of 'survey region' can be questioned, especially when (as with SOWER) the track placement cannot be fully specified beforehand for operational reasons (e.g., unknown location of the ice edge).

team of developers, who took a cue-based approach to detection probability using a Hazard-Probability (HP) model; they did not develop a spatial model, instead using a traditional (post-) stratified abundance estimate. A third approach – IM – attempted to combine cue-based detection probability with a spatial model but did not reach fruition. The IWC's agreed estimates of abundance (IWC, 2013) are a hybrid of results from OK and SPLINTR.

The overall structure and spatial-modelling aspects of SPLINTR were largely settled by 2006 and refined over the next 2–3 years. These parts of the model seem to be generally satisfactory and of general applicability. Detection-probability *per se* has been more problematic, both for our school-based SPLINTR and for the cue-based OK model (Okamura & Kitakado, 2010). We continued to work on AMW detection-probability for several years, while the IWC investigated why SPLINTR and OK were giving such different estimates based on the real SOWER data, differences much larger than were seen when fitting to simulated data (Palka & Smith, 2024). Eventually, diagnostic investigations in 2011 suggested that the TCI assumption in SPLINTR might be causing appreciable negative bias (and that the HP formulation in OK might have positive bias) (IWC, 2012a). In 2012, we began developing a cue-based HP version of SPLINTR (Section 5.1.1), until further diagnostic work highlighted some intrinsic difficulties with any HP approach to AMW. At this stage, eight years after the completion of CP3 and two years after the final SOWER experimental cruise, we had progressed about as far as we could, at least without a complete rethink of detection-probability modelling that would entail major research effort without any guarantee of success.

The IWC agreed on its 'best available' AMW estimates in 2013, based on estimates from OK adjusted by overall 'factors' estimated from TCI SPLINTR, e.g., for imbalanced spatial coverage and various points of modelling/data-selection detail (see Data and Modelling Choices below for more details). Although the 'best available' estimates reflect careful modelling attention to the challenges listed in Section 1.2 (unlike the 'standard method' estimates), there are still some deficiencies, such as inconsistencies with the diving-behaviour data; CP-aggregated rather than spatially-localised adjustment for such 'factors'; and the inability to coherently compute variances for such a hybrid. Nevertheless, for the purpose of inferring what is likely to have happened to AMW abundance in the late 20th Century, the unresolvable and un-modellable uncertainties around SOWER – i.e., the proportion of 'invisible' AMWs outside surveyed open-water areas, and how much that proportion changed between years and decades – are probably now more important than any remaining methodological deficiencies. In the following sections, we concentrate on the 'classic' pre-2012 TCI version of SPLINTR, which includes all the spatial modelling.

2 TECHNICAL DETAILS OF SPLINTR

Any LT/DS spatial abundance estimate for whales needs to have at least two components: one dealing with detection-probability, and one with spatial distribution. If average school size does not vary spatially, then the spatial distribution needs only describe the density surface of 'points' (either schools or whales). However, if average school size does vary spatially, as with the SOWER surveys, then the spatial distribution conceptually needs to describe a greater level of detail: e.g., average school size as well as density of schools, or separate densities for schools of different sizes. Note that even a stratified estimate can be seen as a special case of 'a spatial model', in which density (and/or mean school size) is assumed constant within each 'stratum'.

Accordingly, SPLINTR is a three-stage model:

1. Probability of actually seeing a school, given true school size and conditions;
2. Frequency distribution of true school size, which depends on spatial location (and year);
3. Density of schools, which depends on spatial location (and year).

If true school size was known for each sighting (or for a large-enough representative subset), these three stages could be fitted in order, each stage using the results from the previous stage as 'offsets' to inform the next.[9] At least in principle, that process would be comparatively simple, in that each stage belongs to an

[9] Because of the complexity of each part, and the need for careful checking of diagnostics, we favour fitting spatial abundance models successively rather than simultaneously, as long there is no information loss. In a three-part model of this type, it is evident that the data used in later parts would carry no information to assist with fitting earlier stages, so a sequential approach is attractive.

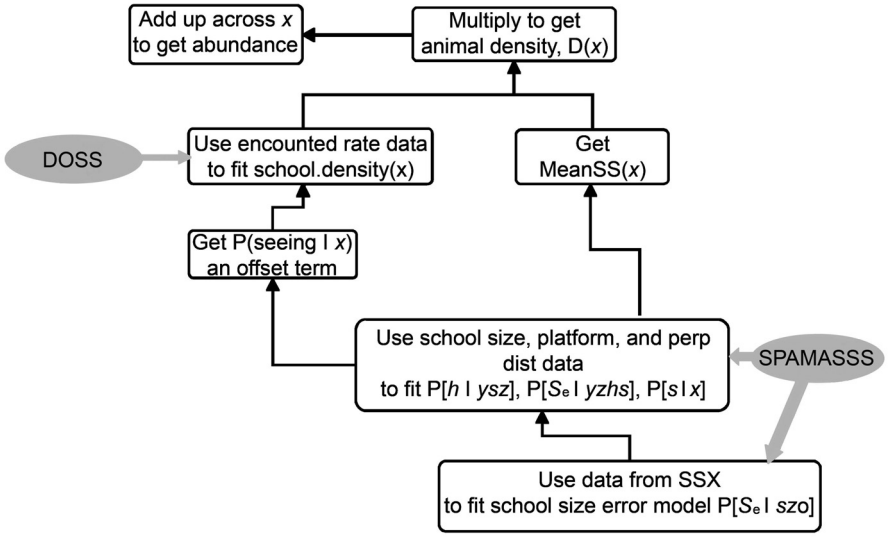

Figure 2. Schematic (but simplified) operation of SPLINTR. The two 'models', which incorporate spatial smoothers, are indicated by grey ovals. Notation is as follows: x is location; h is the combination of platforms that sees a school; o is the fact that the school was seen by at least one platform; y is perpendicular distance; z represents environmental sighting covariates ('weather'); s_e is reported school size; s is true school size. Some details have been omitted for clarity.

established statistical paradigm. In practice, it would still have been a challenging task in the mid-2000s as there was no directly suitable software for any stage, and the problem of how to propagate uncertainty (variance) across linked models had not been addressed.

For SOWER, the process is even more difficult. True school size is known for CL-mode, but g_0 cannot be estimated from CL sightings (or SSX sightings) alone because there is no independent observer, and, although a subset of IO-mode sightings do have a confirmed size, it is a highly skewed subset (for example, size-confirmation is more likely when more platforms see the school). Hence, it is impossible to estimate detection probability in a first step. Instead, Stages 1 and 2 have to be tackled simultaneously, while also dealing with school size uncertainty, for which SSX does provide direct data. That leads to a complex model, SPAMASSS (Sighting-Probability-And-Misunderestimation[10]-And-Spatial-School-Size). SPAMASSS treats certain properties of each detection as random response variables (perpendicular distance; who saw it; estimated school size) and conditions on other properties as explanatory covariates (location/year; weather conditions; survey mode). It does not take account of how many detections were made from place to place. Once SPAMASSS has been fitted, an overall detection probability can be estimated for each section of trackline, averaged across the estimated school-size distribution in that locality. For Stage 3, the actual number of schools seen in each section, without regard for their individual properties, such as perpendicular distance, can be fitted via a spatial model, DOSS (Density-Of-Schools-Spatially; Section 2.3), with the estimated overall detection probabilities as offsets.

Figure 2 illustrates how the results from SPAMASSS and DOSS are combined into abundance estimates. Both main models involve the general notion of a spatial smooth (actually one per model per survey year). When we came to fit to data, we used completely separate analyses for CP2 and CP3 data, except that *all* the SSX data were used in *each* analysis.

2.1 Smoothers for spatial abundance estimation

Both parts of SPLINTR require spatial smooths embedded into non-standard statistical models. Developing a well-behaved smoother was a major part of the SPLINTR process. Smoothers for spatial density and abundance estimation sound appealing in principle: there are good reasons to expect that they should reduce uncertainty

[10] 'Misunderestimation' was a briefly-famous malapropism coined by George W. Bush in 2000, which seemed perfectly apt for SOWER IO-mode school sizes. With hindsight, it is not quite ideal because it carries a connotation of 'mistake'. In fact, obedience to the sighting protocols demands that many school sizes should be 'misunderestimated'.

generally, characterise it more accurately, and avoid bias in the rather typical case of uneven effective coverage. However, in the early days of their application, Hedley & Buckland (2004) noted two practical problems:

1. Where survey coverage is poor near the edge of the survey region, particularly towards the corners, 'off-the-shelf' smoothers, such as thin-plate splines, tend to extrapolate linearly. However, since the model operates on a log-abundance scale for mathematical reasons, this leads to locally exponential trends in density near the edge, leading to positive bias and large estimated variance.

2. Where there is complex topology – e.g., where two bodies of water are separated by a narrow tongue of ice or land – then data from one place can 'leak' influence into predictions elsewhere in a biologically unreasonable way, because of the 'as the crow flies vs. as the whale swims' problem (Point #5 in Section 1.2).

SPLINTR incorporates a 'soap-film' smoother called `soap` (Wood *et al*., 2008), specifically designed to address the second problem but which fortuitously appears to overcome the first as well. `soap` has two parts. The first is a cyclic spline (i.e., ends joined together) laid down around the perimeter of the area to be smoothed, including tracking any complex boundaries. While its x and y-coordinates are fixed, its z-value (representing the local mean) can be changed by the spline's parameters. The perimeter spline can be thought of as a stretchy wire frame that can be bent upwards or downwards. The second is a 'soap-film' anchored to the perimeter spline which stretches across the inside of the domain. The data 'suck' the soap-film towards them. Two smoothing parameters control: (1) the wiggliness in the perimeter spline; and (2) how strongly the data suck.

By ensuring that the perimeter spline does not cross tongues of land or ice, the soap-film construction clearly prevents leakage. The perimeter spline can also be set up to provide 'smoother taming' to address the extrapolation issue. Even when the perimeter spline goes round part of the region where there are no data, its start and end values are effectively anchored by the data near the start and end, and any excursions between are penalised by the smoothing parameter. The problem encountered with, for example, thin-plate splines – namely linear extrapolation into a square corner – doesn't happen because it would require a sharp change in derivative near the corner, which would be heavily penalised. See Wood *et al*. (2008) (Fig. 4) for an example of leak-proofing and taming.

Smoother taming and leakage prevention are less important for spatial school size, because there are limits to how big/small the mean school size can become (assuming that all 'sufficiently big' groups are grouped into a single category). However, they are still convenient features to have, and there is no extra difficulty associated with using `soap` rather than some other smoother for spatial-school-size.

To ensure that distances are approximately comparable between x and y-coordinates, we transformed the longitude and latitude within each survey to (x,y) coordinates as follows:

$$y = latitude$$
$$x = (longitude - midlong) \times \cos(longitude)$$

where 'midlong' is the longitude midway between the eastern and western ends of that survey, with appropriate adjustments for the international dateline.

2.1.1 Fitting smoothers and estimating smoothness[11]

Although there have been many proposals for how to formulate and fit 'smoothers' in various applications, the basis-and-penalty approach (Kimeldorf & Wahba, 1970; Wood, 2006) has risen to the fore thanks to a compelling combination of computational stability, generality, and theoretical rigour, as well as being a natural extension of well-known statistical techniques, such as GLM and REML. A smoother – which is often, like soap, defined in terms of the implicit solution of a differential equation – gets approximated by a weighted sum of a moderate number of basis functions, each corresponding to one column of a design matrix. The weights become coefficients

[11] Notation in this section follows mainstream statistical practice rather than Distance-sampling, so X,Y,S etc., have different meanings here to the rest of the paper. Our notation is similar to Wood (2011) but differs slightly since that paper is restricted to GLM-like settings, whereas SPLINTR response variables are more complex.

β to be estimated, and the smoother simply becomes part of the 'linear predictor' (or its analogue in non-GLM settings). Smoothness is enforced by placing a Gaussian prior on the coefficients, with covariance matrix $\lambda^{-1}S^{-1}$ where the scalar λ is a 'smoothness parameter' that needs to be estimated, and S is a fixed matrix which arises from the nature of the smooth itself. The operational details of estimating λ and β are exactly equivalent to random-effects/mixed-models.

Having selected a particular type of smooth such as `soap` or thin-plate-spline, the R package `mgcv` (Wood *et al*., 2016) provides a convenient way to compute the design matrix X and penalty matrix S for any set of covariates where data were measured and/or predictions will be sought. These matrices apply regardless of what type of observations are made at these covariate values. While `mgcv` is usually used to fit observations from well-known statistical families, it can also be used just to set up X and S, which can then be incorporated into bespoke likelihood calculations and fitting algorithms, such as SPAMASSS. For example, in SPLINTR, the perimeter smooth for each year's region in SOWER is a cyclic p-spline of degree 1, with X and S obtained by calling `mgcv::smooth.construct (s( d, bs = 'cp', k, m = 1))` where d and k describe the boundary.

Of the various criteria that have been proposed for estimating λ, a REML approach corresponds naturally to the Bayesian formulation of Wahba (1990) and has shown good numerical stability (Wood, 2011). It also generalises straightforwardly to any kind of response variable, which is essential for both parts of SPLINTR. If Y is the response variable, $f(.)$ denotes the probability (density) of its argument, and θ is any other set of parameters, such as those governing detection probabilities, then the underlying Bayesian framework amounts to this log-probability (up to an additive constant):

$$\log f(Y,\beta \mid X;\lambda,\theta) = \log f(Y \mid X\beta;\theta) + \frac{1}{2}\log \| \lambda S \|_+ - \frac{1}{2}\lambda\beta^T S\beta \tag{1.1}$$

where $\|\lambda S\|_+$ is a generalised determinant (product of non-zero eigenvalues). For `soap`, there are actually two sub-smoothers, each with its own design matrix, smoothing parameter and penalty matrix. When multiple years/regions are modelled, each receives its own soap-film smoother, although the same pair of smoothing parameters might be shared by all years, thereby adjoining all the separate X's to form one giant design matrix, and adding together all the terms involving individual S's.

Since β are neither 'data' nor 'parameters' in a classical sense (they are random effects), Equation 1.1 is *not* a valid log-likelihood. Inference on λ and θ needs to be based on the marginal log-likelihood (REML) after integrating out β:

$$\log f(Y \mid X;\lambda,\theta) = \frac{1}{2}\log \| \lambda S \|_+ - \log \int_\beta \exp\left(\log f(Y \mid X\beta;\theta) - \frac{1}{2}\lambda\beta^T S\beta \right) d\beta \tag{1.2}$$

Equation 1.2 usually has no closed form, so to implement REML, a Laplace approximation can be used. For any current working value of (λ,θ), this means doing an inner maximisation over β in Equation 1.1, then computing its Hessian, then combining the inner maximum and the determinant of the Hessian in a simple formula to approximate Equation 1.2. An outer optimisation of this whole process is used to estimate (λ,θ). The most interesting quantities are usually the coefficients β, which govern the shape and level of the smoother (and thus, for example, directly determine the predicted number of schools in DOSS). So-called 'empirical Bayes' inference about β can be based on its approximate posterior distribution given the point estimates $(\hat{\lambda},\hat{\theta})$, which arises naturally from the REML formulation; see Wood (2011) for details. For non-standard response variables, as in SPLINTR, implementing this requires sophisticated Automatic Differentiation machinery; see Skaug & Fournier (2006) for the principles, and Section 2.4 for our own similar approach, which uses different software tools. Once that machinery exists, its application to any problem is fairly routine.

The description above does not depend on the distribution of the response variable Y, which needs to be captured in $f(Y|X\beta;\theta)$. There are two different soap-film smoothers in SPLINTR: one for school size, and one for school density. The response in the first is a categorical variable for each school sighting. The response for school-density is number of schools seen within a short stretch of effort. We describe the detailed treatment of these responses in the following sections.

2.2 School size and detection probability

The SPAMASSS model has three components to its log-probability (Eqn. 1.1), which we call Λs:

$$\Lambda s = \log \mathbb{P}\left[\text{data and random effects}|\text{parameters}\right] = \Lambda_{CP} + \Lambda_{SSX} + \Lambda_{\Delta Z}$$

These three components are explained in the following subsections, after a short note on how the model handles 'sighting conditions' recorded at the segment level. For tractability, we worked throughout with five categories of school size (1; 2; 3–4; 5–9; 10+), assuming that detection probabilities did not depend much on school size within each category. Mean school size within each category and year was estimated by the corresponding mean from CL-mode sightings.

2.2.1 Sighting-conditions covariates

Standard modern software for multiple-covariate distance sampling allows a general specification for sighting-conditions covariates that affect detection probability (recorded for each segment of effort, not for individual sightings). There can be several covariates in one model, and users can experiment with model selection to decide which covariates to include. Throughout SPAMASSS, we opted, instead, to allow just one pre-specified discrete-valued covariate (perhaps with vessel-specific estimates), for two reasons:

- To avoid making the computation and model-investigation more complicated;
- To allow automatic enforcement of sensible constraints via the code (e.g., that better conditions always increase detection probability).

This covariate, whatever it may be, is referred to below as Z. Most of our SPLINTR results, including our 'preferred estimates', were made with Sightability as the covariate, as a three-level factor: 2 (worst), 3, 4–5 (best). Sightability in SOWER is the Captain's assessment of a composite of effects, so potentially should have the best explanatory power of any one covariate. There is clearly some subjectivity in the definition, which could potentially change over time (and between personnel), but a generally consistent relationship was found between Sightability and other sighting-conditions covariates, such as Beaufort sea state and Weather code, at least within-boat and within-CP-series (Peel & Bravington, 2005). As part of the IWC process leading to agreed estimates, we also made some estimates using Beaufort sea state instead as a two-level factor.

2.2.2 Λ_{CP}: sighting probabilities and school size

This is the main part of the SPAMASSS model and incorporates all the per-sighting data from CP2 and CP3. It is the only part of SPLINTR that differs between the TCI and HP versions. Here, we describe the TCI version.

Each sighting is made at covariates x (location), z (sighting conditions), and m (survey mode, either IO or CL). The observations are y (perpendicular distance), s_e (estimated school size), and h (who saw the school, i.e., what combination of platforms). The four possibilities for h boil down to AB, Ab, aB and abC, where a capital means 'seen' and a lower case means 'not seen'. If either A or B sees the school, then the one-way independence effectively censors platform C.[12] In CL-mode, there is only one possibility: $h = A \cup C$ since the platforms immediately notify each other of any sighting.

In addition, we condition on the fact O which denotes that *something* was seen (e.g., that $h \neq abc$ in IO-mode). We assume that there is some true school size s for the sighting, which may or not equal S_e.

The statistical task is to model $\mathbb{P}[ys_eh \mid xzom]$ which can then be used to compute Effective Strip Width ESW (including g_0) as a function of x, z and m. Appendix A.1 shows that:

$$\mathbb{P}\left[ys_eh \mid xzom\right] = \frac{\sum_s \mathbb{P}[h \mid yszm] \times \mathbb{P}[s_e \mid yzhsm]}{\sum_s \mathbb{P}[o \mid szm] \times \mathbb{P}[s \mid x]} \tag{1.3}$$

[12] At least in perpendicular-distance-only models. In cue-based models, such as HP, there is some information for schools which C sees before A or B.

The three terms in the numerator, which are described in more detail in the following subsections, have simple interpretations, though parameterising them is not simple. Note that the denominator of Equation 1.3 requires nothing more than is already needed for the numerator (see Appendix A.1). The interpretations are:

- $\mathbb{P}[h \mid yszm]$ is a general 'distance sampling' term for probability-of-sighting. This term is complicated because of SOWER's multiplatform setup.
- $\mathbb{P}[s \mid x]$ describes how the frequency distribution of true school size varies spatially. It acts as a mixing distribution on the distance-sampling probability.
- $\mathbb{P}[s_e \mid yzhsm]$ describes school size uncertainty. It is only relevant in IO-mode. In CL-mode, $\mathbb{P}[S_e = s' \mid yzhs,m = CL] = \mathbb{I}[s' = s]$. Exactly the same term was used in Λ_{SSX} for the school size experiment data where both s and s_e are available.

2.2.2.1 DISTANCE SAMPLING AND DETECTION PROBABILITIES

The two-platform version of TCI needs to be able to compute $\mathbb{P}[h \mid y]$ for the three cases $h = AB$, $h = Ab$, $h = aB$. Specifying these 'independently' can lead to contradictions (e.g., $\hat{\mathbb{P}}[A] < \hat{\mathbb{P}}[AB]$), so a different set of fundamental functions is used, which can be specified independently and from which $\mathbb{P}[h \mid y]$ can be reconstructed without violating any logical constraints. There are several possibilities, of which the most usual seems to be $\mathbb{P}[y \mid A \cup B]$, $\mathbb{P}[A \mid By]$ and $\mathbb{P}[B \mid Ay]$ (e.g., Fewster & Pople, 2008).

SOWER is much more complicated because there are three platforms, with C being only one-way-independent, and two survey modes, with B only operational in IO-mode, and A and C operating non-independently in CL-mode. There are five cases to consider: the four IO-mode possibilities $h = AB$, $h = Ab$, $h = aB$, $h = abC$, and, for CL-mode, $h = A \cup C$.[13] One of the hardest parts of SPLINTR was to devise and parameterise fundamental functions for this variant of TCI. At least 14 types of parameter appear to be required to specify all five probabilities, and since many of those parameter-types may plausibly depend on covariates (school size and/or sighting conditions), the actual number of coefficients required may be quite large. In the end, we arrived at the set of five below. Appendix A.2 shows how $\mathbb{P}[h \mid y]$ can be reconstructed and gives further details of the parameterisation.

- A standard distance-sampling function $\mathbb{P}[y \mid A \cup B \cup C]$;
- Four conditional probabilities:
  - $\mathbb{P}[A \cup B \mid A \cup B \cup C,y]$
  - $\mathbb{P}[A \mid By]$
  - $\mathbb{P}[B \mid Ay]$
  - $\mathbb{P}[C \mid aBy]$

The last of these corresponds to a type of sighting that could never actually occur under IO-mode protocols. However, it could occur if the platforms were all independent. It is required when computing $\mathbb{P}[h \mid y]$ in accordance with the laws of probability.

2.2.2.2 SPATIAL SCHOOL SIZE

To describe how the frequency distribution of true school size[14] varies spatially, we used a smooth polytomous regression (e.g., Harrell, 2001). The 'base' frequency distribution in a given year – i.e., five probabilities, one per school-size category s, and summing to 1 – is parametrised as ζ_s^{SS}, the corresponding quantiles of a hypothetical Gaussian distribution. For the base distribution, we have $\mathbb{P}[S = s] = \Phi(\zeta_s^{SS}) - \Phi(\zeta_{s-1}^{SS})$, with the convention that

[13] For our TCI model, we did not separately treat the cases in IO-mode where C saw the school before A or B. If either A or B saw it eventually, we did not use the information on whether C saw it beforehand. For our extension to cue-based HP models, where forward-sighting distances are crucial, we did treat the C-first sightings separately.

[14] 'True' as in 'among all schools that exist nearby,' regardless of whether anyone is trying to observe them or not.

$\zeta_0^{SS} \equiv -\infty$. The base distribution is then adjusted at any desired location x (a rescaled longitude and latitude) by a year-specific soap-film smoother $\text{smoo}^{SS}(x)$:

$$\mathbb{P}[S = s \mid x] = \Phi\left(\zeta_s^{ss} - \text{smoo}^{ss}(x)\right) - \Phi\left(\zeta_{s-1}^{ss} - \text{smoo}^{ss}(x)\right) \tag{1.4}$$

The remaining ζ_s^{SS}'s are estimatable parameters that are constrained to increase with s. The smoother is implemented parametrically (as described in Section 2.1.1). One consequence of the polytomous formulation is that mean school size can never fall below 1 nor exceed the observed mean school size in the largest category. In practice, it never got anywhere close to the latter.

This polytomous formulation is a compromise between the full flexibility of independent spatial smooths for each school-size-category (not practical when SPLINTR was developed, though see Section 5), and the parsimonious, but inflexible, not-upper-bounded option of arbitrarily assuming some parametric distribution, such as the Negative-Binomial for the base distribution.

2.2.2.3 SCHOOL SIZE UNCERTAINTY IN IO-MODE

We explored numerous different formulations for school size uncertainty in the course of developing SPLINTR. While the formula enters into all the IO-mode likelihoods, there is no direct data to parametrise it (because true school size is unknown). It also underpins the SSX likelihood (discussed below), which is the main source of direct information on school size uncertainty.

For the final TCI results, we used a shifted-Binomial model involving a bias parameter υ that can depend on true school size and sighting conditions z. Specifically:

$$\left(S_e - 1 \mid s, z\right) \sim \text{Bin}\left(s - 1, \upsilon_{sz}\right) \tag{1.5}$$

This simple form deliberately omits several potential complications compared to the full equation $\mathbb{P}[S_e \mid yzhs, m = \text{IO}]$. First, it does not allow for possible school size over-estimation, something which the SOWER protocols are explicitly designed to avoid.[15] Second, there is no-dependence, even though one might expect that the more platforms see the school, the better its size will be estimated. However, since SOWER only systematically records whether Platforms A and B saw the school, whereas in practice the school size estimates were also substantially informed by Platform C, the 'real h' is not available. Third, there is no dependence on perpendicular distance y. We abandoned a more complicated y-linked formulation early on SPLINTR based on preliminary results from SSX. Subsequent analysis of complete SSX, after SPLINTR was complete, suggested that a y-dependent model might have been better after all. For further discussion, see Section 3.1.

2.2.3 Λ_{SSX}: school size uncertainty

The SSX experimental data provides ground-truth data on school size uncertainty (Section 1.1.1). IO-mode protocols, including school size estimation, were followed as usual (usually without Platform B operating). The true school size was checked by closing on the school *after* it had passed abeam (and thus after normal IO protocols would have stopped).

We used each of the 106 SSX observations for which true (i.e., post-closure) size was 2 or more. We computed the probability of the recorded pre-closure size given true school size and environmental conditions according to Equation 1.5. In other words, we took each SSX observation as a trial with a Binomial outcome, with Λ_{SSX} being the sum of the log-likelihoods.

Although the quantity of SSX data is quite low, SPLINTR will not run without it: parameters are no longer statistically identifiable. We view that as unavoidable in SOWER analysis; otherwise, there is no defensible way to check assumptions about school-size uncertainty in IO-mode; see Section 3.1. We therefore had to re-use the SSX data in both the CP2 and CP3 analyses, leading to some between-CP correlation between final estimates, which we ignored.

[15] One of the simulation scenarios (Palka & Smith, 2004; 2005) includes an appreciable proportion of overestimation, evident in the corresponding simulated 'school size experiment' datasets. For these specific simulated datasets, we used a more complicated distribution (Weibull), again with an estimated parameter bsz, which allowed for overestimation.

In principle, there is also information from SSX in the perpendicular distance conditional on true school size (as in SOWER CL-mode), but SSX contains only about 4% of the number of sightings in SOWER itself, so we did not pursue this extra complication.

2.2.4 Λ_Z: Z-change model

Most of the information on detection function parameters should come from details of the sightings themselves, regardless of the encounter rates and how these vary with environmental covariates. On a large spatial scale, there is substantial confounding between weather conditions and location, so it would be logically incorrect to include all the encounter rate data directly in the SPAMASSS (detection-function) model. Nevertheless, there is some information on overall detection probability that is embodied in relative encounter rates over short distance scales. Suppose there were twice as many sightings-per-mile in Good-weather stretches as in neighbouring Bad-weather stretches, but the SPAMASSS detection function only predicts 10% more, there would clearly be some tension between the detection-function model (or at least its point estimates) and the encounter-rate data.

Some of this tension was evident in early SPLINTR fits to real not simulated data, so we introduced the 'Z-change model' as an extra component to the SPAMASSS log-likelihood. Specifically, when sighting conditions change during a stretch of effort, we make a pairwise comparison of numbers of sightings in a small region (ca. 6 nmi total trackline) before and after the change. This is much smaller than the spatial scale of variation represented in the soap-film models, which is set indirectly by the knot spacing of 90 nmi. Hence, there is no confounding between the large-scale spatial DOSS models and the small-scale Z-change model; nor is there any double use of data. The Z-change model uses only the difference between encounters over small spatial scales, whereas the school density model only uses the total over a much larger scale.

Clearly, there is no information in a Z-change event unless at least one school is seen. If there is, then, since local density will be roughly constant within that small distance, the number seen before/after the change should follow a Binomial distribution conditional on the total number of schools seen, allowing for any difference in the before/after time interval, and, of course, for the intrinsic difference in ESW due to before/after sighting conditions. However, to allow for imprecision in recording the time of the change in conditions, and because of fine-scale clustering of schools which might, by chance, lead to more sightings on either side of the change, overdispersion might be expected. We therefore used Beta-Binomial distributions with an extra parameter for any overdispersion. Results from incorporating the Z-change model on analyses of the IWC simulated datasets were encouraging (Palka, 2009). For the real SOWER data, it improved the diagnostics in the DOSS model without worsening fits to the detection-function data *per se*. We suspect this improvement is because the underlying TCI model does not handle AMW g_0 effects perfectly (not that any available model seems to do so; see Discussion).

2.3 School density

For our 'Density of Spatial Schools' (DOSS) submodel, we only used IO-mode data. We split the track into snippets of, at most, 15 minutes (about 3 nmi transit), with shorter units whenever the environmental conditions changed or if the vessel went off-effort before the 15 minutes elapsed.[16] The spatial smoother is used to compute expected whale density ρ_i at the centre $\bar{x}_i$ of snippet i, which is treated as the average for the whole snippet. The response variable N_i is the number of schools seen in snippet i. The equations are:

$$\begin{aligned} \log\rho_i &= \sum \beta_k f_k\left(\bar{x}_i\right) \\ \mathbb{E}\left[N_i\right] &= \hat{p}_i \ell_i w_i \rho_i \end{aligned} \tag{1.6}$$

Where ℓ_i is the length of trackline in the snippet, w_i is the strip width, and $\hat{p}_i$ is the school-averaged estimated detection probability for any school within the truncation distance of the trackline near $(\bar{x}_i, \bar{y}_i)$, the snippet's

[16] In principle, smaller snippets allow slightly more precise estimates because they are more responsive to local gradients in the smoother, but there is not much point in making them too fine. For SOWER data, there is at least a 10-minute window in which any AMW sighting could be made (from furthest straight-ahead sighting distance to abeam), so 'location' of a sighting is a rather imprecise concept at this scale.

centre. The latter needs to take into account both sighting conditions z_i and the local distribution of true school size from Equation 1.4:

$$\hat{p}_i = \sum_s \hat{p}_s(z_i)\hat{\mathbb{P}}[S = s \mid \bar{x}_i] \tag{1.7}$$

Since Equation 1.6 specifies the mean, it remains to specify the statistical distribution of $N \mid \rho$. Since N consists of count data, the obvious choice would be Poisson distributions independent across snippets, but this is not adequate for SOWER (and many other LT/DS) datasets – see below for more details.

2.3.1 Fine-scale clustering

Basis-penalty smoothers, such as `soap`, can only represent density variations above some minimum spatial scale, determined indirectly as some fraction of the knot spacing (Wood, 2006). For SPLINTR, we eventually spaced the knots 90 nmi apart, equivalent to about eight hours of IO-mode survey[17] since this proved adequate to describe large-scale distribution. However, the practical experience of observers is that AMW schools are often found in clusters over much smaller spatial scales, of the order of 30–60 minutes, which is too fine to be represented well with knots at eight-hour spacing. Our initial investigations of SOWER, made without allowance for clustering, showed borderline significant residual autocorrelation due to clustering.

This fine-scale spatial autocorrelation is an issue for fitting smoothers. If ignored, it tends to lead to undersmoothing on the large spatial scales which are important for inferring abundance (Diggle & Hutchinson, 1989). Undersmoothing affects estimated variance (biased upwards) and may affect point estimates too, since it permits more extreme behaviour near boundaries. While some mechanism to handle local autocorrelation is therefore desirable, as long as it is statistically reasonable, the details are probably not critical because the impact on abundance is indirect, via the smoothing parameter.

One approach might be to aggregate the data over some spatial scale well beyond the 'half-life' of autocorrelation (by making the snippets very large) and then modelling the count data as an overdispersed non-Poisson variable independent across snippets. The overdispersion parameter is supposed to deal with the consequences of within-snippet autocorrelation. However, this entails the following not necessarily straightforward decisions and trade-offs:

- What distribution to use (Negative Binomial, Tweedie, Quasi-Poisson)?
- How big is big enough?
- Aggregation cannot apply (at least not without horrible consequences for the underlying model code) across a change in environmental conditions, so some (many) snippets will not be 'as large as wanted';
- How should the overdispersion parameter be adjusted for snippets of variable size?
- Potential sensitivity to the scale of aggregation.

Instead of trying to aggregate, we made explicit allowance for autocorrelation between snippets via a discretised approximation to the Skaug (2006) Markov-Modulated Poisson Process (MMPP). This allows for transient variations in density around the local background level which persist over a distance-scale much smaller than that represented by the spatial smoother. In each snippet i, the MMPP has a state M_i that is either 'Hi' or 'Lo'. The state is not directly observed, but it affects the expected value of N_i, the number of schools seen, relative to the background density ρ_i. N_i then follows a Poisson distribution around that expected value. At the *end* of the snippet (but not in the middle), the state may flip. This simplification recasts the MMPP as a discrete Hidden Markov Model (HMM; Baum & Petrie, 1966; Zucchini *et al*., 2021) and is a reasonable

[17] The statistical rationale for most of the GAM machinery in `mgcv` is that the number of knots is almost irrelevant provided some minimum threshold is exceeded. Beyond this point, it is better from a computational point-of-view to not use too many. The threshold is set (for line-transect-type data) by the density of sightings and tracks, and for SOWER we found no reason to space the knots more closely (i.e., no clear misfits). Fewer knots means substantially quicker computation.

approximation to a continuous-time MMPP (for which flips can happen anywhere, including in principle several times within a snippet), provided that the typical scale of clustering is larger than the snippet length. In equations, we have:

$$\begin{aligned}
&\mathbb{E}\left[N_i\right] = p_i w_i \ell_i \rho_i \\
&M_i \in \{\text{hi},\text{lo}\} \\
&\mu_i^* \triangleq \mathbb{E}\left[N_i \mid M_i = m\right] = p_i w_i \ell_i \rho_i \xi_m \Delta t_i \\
&N_i \sim \text{Po}\left(\mu_i^*\right) \\
&\mathbb{P}\left[M_{i+1} \neq M_i \mid M_i = m, \ell_i\right] = \exp\left(-\left(\mu_{\min} + \exp\kappa_m\right)\ell_i\right)
\end{aligned} \tag{1.8}$$

where the two κ parameters determine state-flipping rates, and the two ξ parameters determine how much the local density is increased when in the Hi state (by a factor ξ_{hi}) and decreased when in the Lo state (by ξ_{lo}). There is a constraint on ξ_{lo} in terms of ξ_{hi} and κ to ensure that the unconditional mean of N_i is unaffected by the MMPP. The SPLINTR code in fact allows all three free parameters to depend on location covariates (e.g., distance from ice edge), but, in practice, we did not find that elaboration to be useful. The parameter $\mu_{\min}$ is a fixed small constant which prevents the likelihood-maximising routine from getting stuck in a region of rapid state-flips (e.g., shorter than snippet length) where the parameters become unidentifiable. We chose $\mu_{\min}$ so that the probability of at least one flip between smoother knots (about eight hours of travel) could not fall below 50%. With the MMPP, the snippets are no longer independent, but the joint likelihood for the set of consecutive snippets within a stretch of effort can be easily calculated using the 'forward algorithm' for HMMs (Zucchini *et al*., 2021). At the start of the stretch, and whenever there is a substantial break in effort, those state probabilities are reset to the equilibrium determined by $\{\kappa_{\text{hi}},\kappa_{\text{lo}}\}$.

Importantly, the MMPP (or any other approach to fine-scale clustering) is only relevant to the trackline and when fitting the model, not for predicting abundance over larger spatial scales. The small-scale fluctuations it describes have zero mean and would cancel out over any large area thanks to the Law of Large Numbers (Bernoulli, 1713).

Our discretised version, where sightings are grouped in time per snippet of effort, avoids some computational awkwardness with Skaug's original formulation, which works in continuous-time (or space) based on the interval to the next sighting, and involves matrix exponentials. The code for our modified MMPP is compact and requires only about twice the computational effort of a naive Poisson likelihood which treats each snippet as independent. The MMPP should be largely robust to snippet-size (the flipping parameters have units of 'per nmi' and will rescale themselves accordingly), and there is no need to choose any distributions or make other decisions. Thus, it is arguably simpler for the user than the various overdispersion options found elsewhere in LT/DS, e.g., Miller *et al*. (2013).

2.3.2 Closing mode encounter rates

Overall, CL-mode encounter rates are not reliable for density estimation for several reasons. First, the act of closing breaks the normal search pattern and may lead to 'secondary sightings' which are hard to 'unsee' when search mode resumes. This is especially problematic when schools are clustered. The effects are too dependent on local school density for a single overall correction factor to make sense, so both OK and SPLINTR opted not to use CL-mode encounter rates for density, though CL-mode sightings are used in SPAMASSS.

However, from the start of a CL-mode transect until the first sighting (if any), the searching process should be similar to IO-mode, except without Platform B. Thus, it might be possible, in principle, to use something like 'time-to-first-sighting' in CL-mode (or the entire transect, if no sightings are made), to improve inference about local density. To do so without bias requires a good model for clustering. We hoped that the MMPP might achieve that, but we still found a discrepancy between observed and expected first-encounters in CL-mode. Thus, while MMPP should be helpful for IO-mode, it was not sufficient to correct for CL-mode encounter rate bias. The effort required to explore this further did not seem worthwhile so in the end, we abandoned CL-mode for density purposes (as did OK).

2.4 Computation prediction and variance

Computations were done separately for CP2 and CP3, except that SSX data were used in both. For each CP series, the SPAMASSS model contains a separate spatial smooth for each year (to account for variation in mean school size), plus the 100+ detection-function parameters which are shared across years. The general strategy for both SPAMASSS and DOSS is REML-style empirical Bayes estimation as per Wood (2017), using Automatic Differentiation (AD) to implement Laplace Approximations with an 'outer' problem that repeatedly maximises an 'inner' problem using a Quasi-Newton algorithm (Gill *et al*., 1981). Similar ideas were embodied in the software ADMB (Skaug & Fournier, 2006), but we opted to write our own software based around the AD tool Tapenade (Hascoët & Pascual, 2013) as we needed extra flexibility such as the ability to handle multiple simultaneous smooths. Computational software has advanced greatly during and since the development of SPLINTR, such that a rewrite of SPLINTR could take advantage of more powerful tools, such as TMB (Kristensen *et al*., 2016) or perhaps some extension of the INLA framework (Rue & Martino, 2007) to the specific observational data of SOWER along the lines of 'inlabru' (Bachl *et al*., 2019), to simplify the coding while retaining the same basic statistical structure.

Both SPAMASSS and DOSS had separate inner problems for each year, each with its own spatial smoother. That was a computationally efficient choice, since the inner maximisation is $O(p)^3$ in the number of smoothing coefficients p. With 12 years in CP3, trying to fit all the smooths simultaneously would have led to a 144-fold increase in computational burden. (A sparsity-aware framework such as TMB could avoid this problem.)

The outer parameters for SPAMASSS comprise not just the smoothing penalties and MMPP parameters but also all the detection-function and school-size-uncertainty parameters. There are over 100 outer parameters in all. That is not how 'outer problems' are usually set up – normally with only a handful of parameters – but it was essential so that all the inner problems could be handled separately. The power of AD made it feasible (even in 2009) and SPAMASSS runtime was under an hour.

Computation for DOSS was easier, with five outer parameters comprising two smoothness parameters, plus three parameters for the MMPP. Runtime was a few minutes.

Once both models had been fitted, abundance was estimated by predicting school density and mean school size at each point in a fine grid across the region-of-interest (see Fig. 3). Appendix B gives formulae for point estimates and variances, which can be computed by standard first-order ('delta method') expansions. The principles of variance calculation in SPLINTR are straightforward, but the implementation was complicated because of the linked models (see also Section 5.6.3 for post-SPLINTR developments). Note that no bootstraps or other awkward contrivances were needed. Abundance estimates between years within the same CP series are of course correlated, because they share the same detection-function parameters. Estimates in different CP series should be almost uncorrelated. The 'almost' comes from re-use of SSX in both series-specific analyses, which we ignored (i.e., we did not try to compute CP2-CP3 covariances). When interpreting each year's abundance estimate in terms of population dynamics, it is important to consider not just these model-internal variance estimates but also the 'additional variance' that comes from large-scale inter-annual movements (Kitakado & Okamura, 2009).

For fitting to the 5,400 simulated datasets, we used CSIRO's Condor (Thain *et al*., 2005) network to distribute jobs across several thousand idle Windows desktops, which allowed an overnight result and a speedup of about 1,500-fold, compared to running the jobs sequentially.

3 DATA AND MODELLING CHOICES

Palka *et al*. (in prep) documents numerous large-scale decisions about which SOWER data to include/exclude in these analyses, which were agreed by all involved in the IWC process: for example, like-minke sightings, uncertain duplicates, obscure survey mode variants (such as 'BH'), and replicated coverage of a few areas in multiple years. This section covers various more detailed aspects of SOWER data and our decisions about how best to handle them in abundance estimation. These decisions were based partly on direct experience of SOWER cruises, and partly on data analysis. All do have some impact on abundance estimation; some would be of interest only to someone developing another abundance estimation method specifically for SOWER, but a few are of more

general relevance. In some cases, OK made different decisions to SPLINTR. For the IWC-agreed estimates, the IWC's Scientific Committee decided sometimes to follow our decisions, sometimes to follow OK's, and sometimes to average the results from the two. We have noted below the cases where our preference was not adopted.

3.1 SSX

The SSX data provide direct information on the extent of school size underestimation in IO-mode. It comprises 106 sightings with true (i.e., post-closure) school size of 2 or more. There were also 63 sightings with post-closure size 1, which are not directly useful because a school-size of 1 obviously cannot be underestimated given it is observed at all. The 106 sightings are statistically powerful, despite the limited sample size, because each observation contains estimate, truth, and covariates (such as sighting distance and weather conditions) and can be analysed directly with simple statistical models. Apart from their obvious use as a validation and diagnostic tool, the SSX data can be incorporated directly into the likelihood function for all the SOWER data (Section 2.2.3).

In our view, the SSX data are indispensable for constructing a reliable AMW abundance estimate from SOWER, because school size clearly has a major effect on detection probability. Without SSX, there is no way to ground-truth the detection-probability models, which are otherwise forced to rely entirely on a complicated and delicate mixture-decomposition of detection data based on estimated school size in IO mode. About 52% of schools with true size 2 post-closure were estimated as size 1 pre-closure, so the effect of school size error is clearly important, especially for small schools where g_0 is most problematic.

The IWC discussed whether the SSX post-closure estimates were always accurate, and whether over-estimation pre-closure was likely to matter (under-estimation being a deliberate consequence of the protocol). The 2007–09 SSX followed an earlier attempt in the 1980s, which was aborted when it became clear that some topmen (Platform A observers) had begun inflating their pre-closure estimates to match post-closure numbers. Unfortunately, these 1980s data were therefore unusable. This problem did not recur in SSX: the IO protocol of reporting only the minimum number of whales that were evidently in the school was established and well understood. Nevertheless, there were a few cases where pre-closure estimates in SSX did exceed post-closure (Table 3).

In three of the 106 cases with pre-closure estimates of 2 or more, schools had lower post-closure estimates of 1. However, it is rather difficult to mistake one animal for two, and discussions with observers suggest that these 'schools' split up prior to or during the closure attempt. In other words, the pre-closure estimates were likely correct. This is also plausible for the three-into-two and five-into-four cases: even if these were genuine pre-closure over-estimates, the effect is not very important. The final case is a very large school (something very rare in SOWER) where some noise in estimation is inevitable. The rate of 'lost' schools (a school which can no longer be closed on by the time it passes abeam) in SSX also turned out to be very low. Consequently, we concluded that the SSX data justified ignoring the possibility of over-estimation in IO-mode. For the component A_{SSX}, the few cases in Table 3 were adjusted so that post-closure matched pre-closure.

Most SSX only used the A and C Platforms for logistical reasons, so results are not completely comparable with IO-mode. However, the difference in estimated school size with or without the least-effective Platform B is expected to be small. Unfortunately, there are too few SSX sightings with Platform B operational to provide much standalone information on g_0 with known school size.

Table 3

Cases in SSX where school size was apparently over-estimated prior to closure. 'Pre' and 'Post' are the school sizes, 'Count' is the number of such cases, and 'Total cases' is the total number of schools seen with that post-closure size range (either 1, or greater-than-1).

Pre	Post	Count	Total cases
2	1	3	63
3	2	1	
5	4	2	106
40	35	1	
	Totals	**7**	**169**

Table 4

Estimated effect of first-sighting distances (nmi) on school size uncertainty, summarised via predicted probability of underestimating at the distances shown when true size is 2. For this standalone analysis, we used a slightly different model to Section 2.2.3: a zero-truncated Binomial $S_{obs} \sim ZBin(S_{truth},p)$ with $\text{logitp} = \beta_0 + \beta_1 \text{fwd} + \beta_2 \text{perp}$. This model was fitted using VGAM::vglm in R to all SSX data from 2007–09 for which true school size was 2 or more. No further significant effects of S_{truth} or Sightability were found.

Perp	Fwd	$\mathbb{P}[S_{obs} = 1 \mid S_{true} = 2]$
0.10	0.10	0.68
1.40	0.10	0.81
0.10	3.10	0.32
1.40	3.10	0.49

Several covariates, both for general conditions and specific to each sighting, might be expected to affect school size uncertainty. All versions of SPLINTR included Sightability in the school-size-uncertainty model (or Beaufort, in a few sensitivity tests). Early versions also included an elaborate formulation to allow a perpendicular distance effect. Because an initial analysis of some SSX data in 2008 suggested no such effect, we abandoned that in favour of the much simpler Equation 1.5 in all subsequent SPLINTR models, including the results in Section 4.

However, subsequent re-analysis of all post-CP3 SSX data (in 2012, and for this paper), using forward as well as perpendicular distance, does in fact reveal a significant effect from perpendicular-distance (bigger distance implies more underestimation), and an even stronger effect from forward-distance effect (bigger distance implies less underestimation, presumably through more chances to see subsequent cues). Sightability (see Section 3.5) is significant when forward-distance is excluded, but not when forward-distance is included. Poor Sightability leads to smaller forward sighting distances, so the apparent effect of Sightability may simply be acting as surrogate for forward-distance. The estimated effects are substantial (Table 4; Bravington & Hedley, 2012).

Neither OK nor SPLINTR included any distance-dependence in their final models (i.e., those used as the basis for IWC's agreed estimates). Only SPLINTR included a sighting-conditions effect. Quite what these omissions mean for abundance estimation is unclear. One lesson from the SOWER process is that it is almost impossible to correctly intuit the effects of model mis-specification when 'mixing' over different possible true school sizes. It is also plausible that the probability of underestimation should depend on whether the sighting was a duplicate (since there are more chances to see a duplicate), but there are not enough data in SSX to test that.

3.2 Which distance estimate to use?

Perpendicular distance is a key covariate for distance sampling, and even random errors in perpendicular distance measurements can lead to systematic bias (not just increased variance) in abundance estimates (Chen, 1998). For schools seen by more than one platform, more than one estimate of perpendicular distance is available.[18] The question is: which to use? There are arguments in favour of 'Firstperp', 'Lastperp' and 'Highestperp' (see below for more details). The 'standard method' in DESS by default uses 'Highestperp', i.e., from A then B then C if available, regardless of the time-sequence of sightings. The rationale is that reticule-angle-based distance errors are less at larger declinations, i.e., for higher platforms.

The OK model prefers to use 'Lastperp', on the basis that distance errors in SOWER are substantial and will be smaller when the whale is closer to the vessel, i.e., at the last sighting. 'Lastperp' was chosen for the IWC's agreed estimates.

There are also two rationales for 'Firstperp'. To begin with, it is the only way to ensure that IO and CL mode distances are equivalent (in CL-mode, the only distance estimate is from the first sighting, regardless of platform). Second, if (some) whales exhibit responsive movement, then they are less likely to have moved away from/towards the trackline when they are first seen than when they are last seen. Unless special survey protocols

[18] Resights of subsequent cues by the upper bridge (Platform C) do not have distance estimates recorded. The other platforms, A and B, do not record resights at all; the protocol was that they should not be looking for them.

are adopted, responsive movement in distance-sampling generally causes some abundance-estimation bias (Glennie *et al*., 2015), which is presumably minimised (though cannot be eliminated) by using 'Firstperp'.

This raises the question of whether there is evidence of responsive movement for AMWs in SOWER. For all IO-mode delayed-duplicate sightings separated by at least 60 secs, the median of (first perp dist – second perp dist) is –0.003 nmi, and the mean is –0.0127 nmi: i.e., no difference. At first glance, this might seem to be evidence against responsive movement. However, we know that AMWs do move, responsively or otherwise. Suppose they move at random: amongst all whales whose first sighting is at a given perpendicular distance, those which happen to be moving towards the trackline are more likely to get spotted later as a delayed-duplicate[19] than those moving away. Hence, under purely random movement, the second sighting in a delayed-duplicate is likely to be closer to the trackline. The fact that this doesn't happen in the real data is therefore consistent with some net tendency (i.e., in some but not necessarily all AMW encounters) for away-from-trackline movement in SOWER.

There is no conclusive *a priori* argument in favour of any of the three choices. While simulations can provide some insight (Palka & Smith, 2024), a proper understanding of the quantitative tradeoffs between bias induced by responsive movement and by measurement error in SOWER would require a much better understanding of AMW movements than was (or is now) available. We generally preferred to use 'Firstperp' in SPLINTR (including for the results in this paper) due to concerns about responsive movement, but we also investigated the effect of changing to IWC's 'Lastperp'.

3.3 Pre-extension

In SOWER, schools are sometimes seen beyond the end of the official trackline of a leg of effort (i.e., the school is seen before going off-effort, but only passes abeam afterwards), and the sightings are still used. The effective area searched in any leg is thus somewhat larger than the official trackline length multiplied by the ESW. Pre-extension accounts for this by slightly increasing the length of each leg by an amount equal to half the average forward sighting distance.

The 'half' is an *ad hoc* correction factor designed to reflect possible reduced attention at the start and/or end of the leg. Note that it is certainly not the case that the extra effort at the end is offset by a gap at the very start. Observers do not necessarily keep their eyes shut tight until the precise instant that a leg of effort is officially deemed to begin. If they did, it would be a probability-zero event for sightings to be made instantaneously at the start of a leg, whereas there are quite a few such sightings in SOWER.

Pre-extending the effort led to a 2–3% reduction in abundance estimates. While recognising that the phenomenon is real (i.e., the area surveyed is greater than nominal transect length times ESW), the IWC opted not to use pre-extended effort for the agreed estimates, since the basis for choosing 'half' is not clear, and because pre-extension is not traditional with LT/DS estimates.

3.4 Confirmation in IO and CL mode

The SOWER dataset includes a field indicating whether each school size estimate is 'Confirmed' or not. This turns out to be a rather subtle entity, and whether or how to use it during analysis can appreciably affect the abundance estimates. The first point is that 'Confirmation' means something fundamentally different in IO-mode vs. CL-mode. In IO-mode, it means 'the group was seen within 0.3 nmi of the vessel, which carried on as normal,' whereas, in CL-mode, it means 'we chased the group and got close enough to be sure of its size'. The two processes are so different that it is impossible to make inferences from one about the other, so we consider them separately.

3.4.1 IO-mode Confirmation

If we do not use Confirmation status in IO-mode, then all IO-mode groups must be treated as of uncertain size. Thus, in principle, using IO-mode Confirmation status might seem attractive, because at least we would be sure of the size of about 15% of groups. The problem is that it makes things worse for the unconfirmed groups. The

[19] Duplicate sightings, i.e., made by more than one platform, can either be 'simultaneous' (same cue seen) or 'delayed' (presumably the same school, but separate cue seen later).

very fact of failure-to-IO-confirm is likely to be correlated with several variables in the sighting process, such as which platforms actually saw the group, where it was first seen, etc. When trying to use CL-mode data to infer something about school sizes in IO-mode, there is no way to know what these other variables would have been. Hence, there is no reliable basis for deciding which CL-mode observations to use when 'adjusting' IO-Unconfirmed school-size inferences. It is certainly not safe to use them all, and, as above, it is not reasonable to only use the CL-Unconfirmed. The alternative – i.e., trying to develop a comprehensive model for conditional relationships between IO-mode Confirmation status and perpendicular distance, sighting distance, which platforms made the sighting, etc. – seems infeasibly complicated.

Since it is perfectly possible to model the SOWER data without recourse to IO-mode confirmation status, and since the interpretation of 'IO-mode Unconfirmed' is problematic, our (eventual) decision with SPLINTR was that the safest course of action is to omit the Confirmation response variable altogether when analysing IO-mode data. This was also the IWC's choice.

3.4.2 CL-mode Confirmation

About 25% of CL-mode school sizes are Unconfirmed, the great majority with an estimated school size of 1. We cannot simply exclude these sightings, since that would be informative censoring (i.e., choosing the data based on the response values, rather than the covariate values). There are at least two options, though neither is ideal: replace the CL-unconfirmed school sizes by some kind of average (perhaps stratified, etc.) or treat the estimated values as truth. The problem with replace-by-average is that there is quite possibly something atypical about those groups for which the attempt to close is unsuccessful. After all, most closure attempts succeed; failure means that the target group is somehow 'lost', and that is quite likely linked to school size. In particular, it seems much easier to lose a solitary whale than to lose a larger group (Ensor, *pers. comm.*). The Video Dive Time (VDT) experiment at the end of SOWER is informative here. It proved very difficult to track solitary whales (only one successful record) compared to larger schools. Solitary minke whales in the Antarctic seem to be rather skittish, at least when in open water, and the reason for failure-to-close appears to be 'losing the single-whale group' much more often than 'the larger group breaks up'. In other words, the estimated-group-size is probably accurate for most groups that are not closed on, because most failures-to-confirm are solitary anyway. Note that there is no way to check this within the SOWER data itself; we have to rely on observational data (e.g., VDT) and observer/Cruise Leader experience.

For these reasons, SPLINTR treats all CL-mode school sizes as 'confirmed'. This could lead to a small underestimate of mean school size (because a few non-size-one groups will be lost during closure, and estimates of their school size based on the failed closure may be too low), but in view of the on-vessel logistics and VDT data, it seems preferable to trust recorded school sizes in CL-mode rather than replace CL-Unconfirmed school size estimates by an average over cases that are likely to be intrinsically unrepresentative.

The IWC, noting that neither treatment was watertight, chose to split the difference (arbitrarily) in half: i.e., to calculate two sets of otherwise-comparable estimates where only the CL-confirmation treatment differed, then to estimate the average difference across strata, then to apply half this difference to whichever set of 'preferred' estimates were selected.

3.5 Choice of 'weather' covariate

SOWER effort data include several fields related to weather or sighting conditions, including Beaufort (sea state), Sightability, Weather, Visibility, and several more.[20] 'Sightability' is a composite assessment of the suitability of current weather conditions for seeing whale cues, made by the wheelhouse officer up to 1994 or the Captain in later years, and measured on an integer 1–5 scale (5 = best). SPLINTR can only handle one weather-related covariate, so we experimented with both Beaufort and Sightability. Sightability showed consistently stronger effects on estimates of sighting probability (g_0, ESW) than Beaufort – which is precisely what Sightability is meant to do, since it integrates other factors (wind direction, glare, swell height, etc.) that affect the ability to see AMW

[20] The full list for CP2 comprises Weather, WindDirection, WindSpeed, SurfaceTemperature, AirTemp, Visibility, IceCover, Sightability, SeaState and Swell. Not all were regularly recorded.

cues. Although subjective, its inclusion in the data record was specifically for the reason that these multivariate effects are difficult to model. Unlike classical pooling-robust distance sampling, models where $g_0 < 1$ are known to be subject to negative bias when there is unmodelled heterogeneity in the detection-probability model. Such heterogeneity can come from unmeasured but real covariates that affect detection probability. Presumably, some of these are captured in Sightability but not in Beaufort.

One concern about Sightability has been that it may be more subjective, and thus less consistent across vessels and years. While developing SPLINTR, we conducted several analyses to calibrate Sightability against the full set of SOWER weather covariates, which are supposed to use more objective criteria, and to construct a 'Sightability predictor' that would use the other more objective weather-related covariates to build a more objective version of Sightability. The latter actually pointed to inconsistencies among some of the other weather-related covariates. Due to lack of time, we abandoned that approach. On the whole, Sightability did appear to be quite consistently recorded, but SPLINTR diagnostic plots in 2009 revealed some residual vessel effects in CP2 (though not CP3) when Sightability was used. Therefore, in some analyses, we reluctantly switched to using Beaufort in CP2 only. Since the desirability of reducing bias by 'modelling out' heterogeneity seems to outweigh the presence of some residual vessel effect in CP2, our preference is to use Sightability, as in Section 4. Given the large sample size of SOWER AMW sightings, bias-reduction has generally been prioritised over variance-reduction in SOWER analyses.

The IWC chose to use Beaufort instead of Sightability on the grounds that Beaufort would be recorded more consistently and less subjectively.

4 RESULTS

4.1 Detection probability

Estimates of g_0 by platform and strip width in different sightings conditions are given in Table 5. Our estimates of g_{0A} are consistent with the BT-NSP experimental results (Burt *et al*., 2012) which, when constrained to school-size 1 and either 'good' (Beaufort < = 3) or 'bad' (Beaufort 4+) conditions, are $\hat{g}_{0A}(\mathrm{SS}=1;\mathrm{good/bad}) = 0.33/0.20$. In principle, the BT-NSP estimates provide a direct independent comparison, provided that the Buckland-Turnock assumptions were met in this case (see Discussion).

4.2 Abundance

Table 6 shows estimates from our preferred version of TCI SPLINTR, i.e., using the modelling choices just described. The IWC's agreed estimates, plus estimates from one alternative run of SPLINTR (with different data/model choices), are included for comparison. Following tradition, the estimates are broken down into the six Management Areas. While convenient, these are arbitrary divisions for AMWs and do not correspond to biological populations (see Discussion).

The preferred-SPLINTR and IWC estimates are reasonably similar overall, differing by about 10% when completely aggregated. At finer scales of space and time, where each estimate is based on fewer sightings, some differences are inevitably bigger. However, the two sets of estimates make slightly different modelling choices (in the sense of Section 3), so the overall similarity actually disguises several effects which somewhat counteract each other. As an example, the 'alt SPLINTR' row shows the impact of changing three detailed modelling choices within the overall SPLINTR framework. We return to those choices in the next section. For more comments on the numbers themselves and the trends, see Section 5.4 and Palka *et al*. (in prep).

Figure 3 shows perhaps our key result: the estimated spatial distribution of AMWs by year. The general tendency for higher densities near the ice-edge is evident, but it is not universal and there is also substantial longitudinal variation. The soap-film smoother appears to be well-behaved. Only in one year (1991, MA6) does it extrapolate somewhat higher densities in a region that was not well-covered by survey effort.

There is also spatial variation in mean school size, though it is more difficult to visualise. Contour plots have not been particularly enlightening; Figure 4 shows reasonably typical examples for one Management Area from three years, but we have omitted the full set here (see Bravington & Hedley [2009] for the full set). Across all 18 CP years, the highest contour value is 4.0 in one year (MA2, Weddell Sea, 1987), with 3.5 attained in only four of the other 17 years. There is a tendency for school sizes to be higher where sightings are denser, thus usually but

Table 5

g_0 (probabilities) and ESW (effective strip half-width; units of nmi; maximum value would be 1.5 nmi, the truncation distance). Overall g_0 is calculated from $g_0 = 1 - (1 - g_{0A})(1 - g_{0B})(1 - g_{0C})$, following the TCI assumption. 'Sig' is Sightability.

	School size	1	2	3–4	5–9	10+
			CP2			
Sig2	g_{0A}	0.30	0.39	0.61	0.75	0.79
	g_{0B}	0.15	0.21	0.39	0.56	0.61
	g_{0C}	0.37	0.46	0.67	0.80	0.83
	g_0	0.63	0.74	0.92	0.98	0.99
	ESW	0.33	0.42	0.64	0.71	1.06
Sig3	g_{0A}	0.31	0.41	0.62	0.76	0.80
	g_{0B}	0.16	0.23	0.40	0.57	0.62
	g_{0C}	0.37	0.48	0.68	0.80	0.84
	g_0	0.63	0.77	0.93	0.98	0.99
	ESW	0.36	0.54	0.67	0.72	1.08
Sig4+	g_{0A}	0.31	0.63	0.69	0.77	0.81
	g_{0B}	0.16	0.41	0.48	0.58	0.63
	g_{0C}	0.38	0.69	0.74	0.81	0.84
	g_0	0.64	0.93	0.96	0.98	0.99
	ESW	0.38	0.75	0.82	1.07	1.28
			CP3			
Sig2	g_{0A}	0.12	0.37	0.51	0.57	0.62
	g_{0B}	0.06	0.20	0.30	0.36	0.41
	g_{0C}	0.15	0.43	0.57	0.63	0.68
	g_0	0.30	0.71	0.85	0.90	0.93
	ESW	0.16	0.62	0.76	0.81	0.98
Sig3	g_{0A}	0.31	0.39	0.67	0.71	0.79
	g_{0B}	0.16	0.21	0.46	0.50	0.62
	g_{0C}	0.37	0.45	0.71	0.75	0.84
	g_0	0.63	0.73	0.95	0.96	0.99
	ESW	0.41	0.64	0.85	0.91	1.33
Sig4+	g_{0A}	0.38	0.54	0.69	0.72	0.80
	g_{0B}	0.21	0.32	0.48	0.51	0.64
	g_{0C}	0.44	0.59	0.73	0.76	0.85
	g_0	0.73	0.87	0.96	0.97	0.99
	ESW	0.52	0.78	0.99	1.02	1.38

Table 6

AMW abundance estimates in 1,000s of animals by Management Area for 'survey-once open-water not restricted to common northern boundary'; see Palka *et al.* (in prep.) for translation. Total CV does not allow for inter-annual movements ('Additional Variance'; Section 5.4). Several versions of 'preferred SPLINTR estimates' have been published in the IWC's Scientific Committee papers and documents. This set follows section 3 and is similar to those in Bravington & Hedley (2010) except now using Firstperp instead of Highestperp. 'alt SPLINTR' is similar except for changing three choices into the IWC consensus choices: Firstperp → Lastperp; Sightability → Beaufort; Pre-extension → off. The 'IWC' row (which starts from OK estimates and then applies several adjustments calculated using SPLINTR) is from IWC (2012a) (Table 3).

		MA1	MA2	MA3	MA4	MA5	MA6	Total
CP2	SPLINTR	115	133	85	51	254	47	686
	CV	22%	23%	51%	18%	15%	25%	[13%]
	IWC	*86*	*130*	*93*	*55*	*300*	*56*	*720*
	alt SPLINTR	*122*	*145*	*91*	*63*	*305*	*61*	*788*
CP3	SPLINTR	66	52	64	32	157	58	430
	CV	15%	17%	13%	23%	10%	18%	[9%]
	IWC	*39*	*57*	*94*	*60*	*184*	*81*	*515*
	alt SPLINTR	*48*	*46*	*51*	*32*	*195*	*48*	*421*

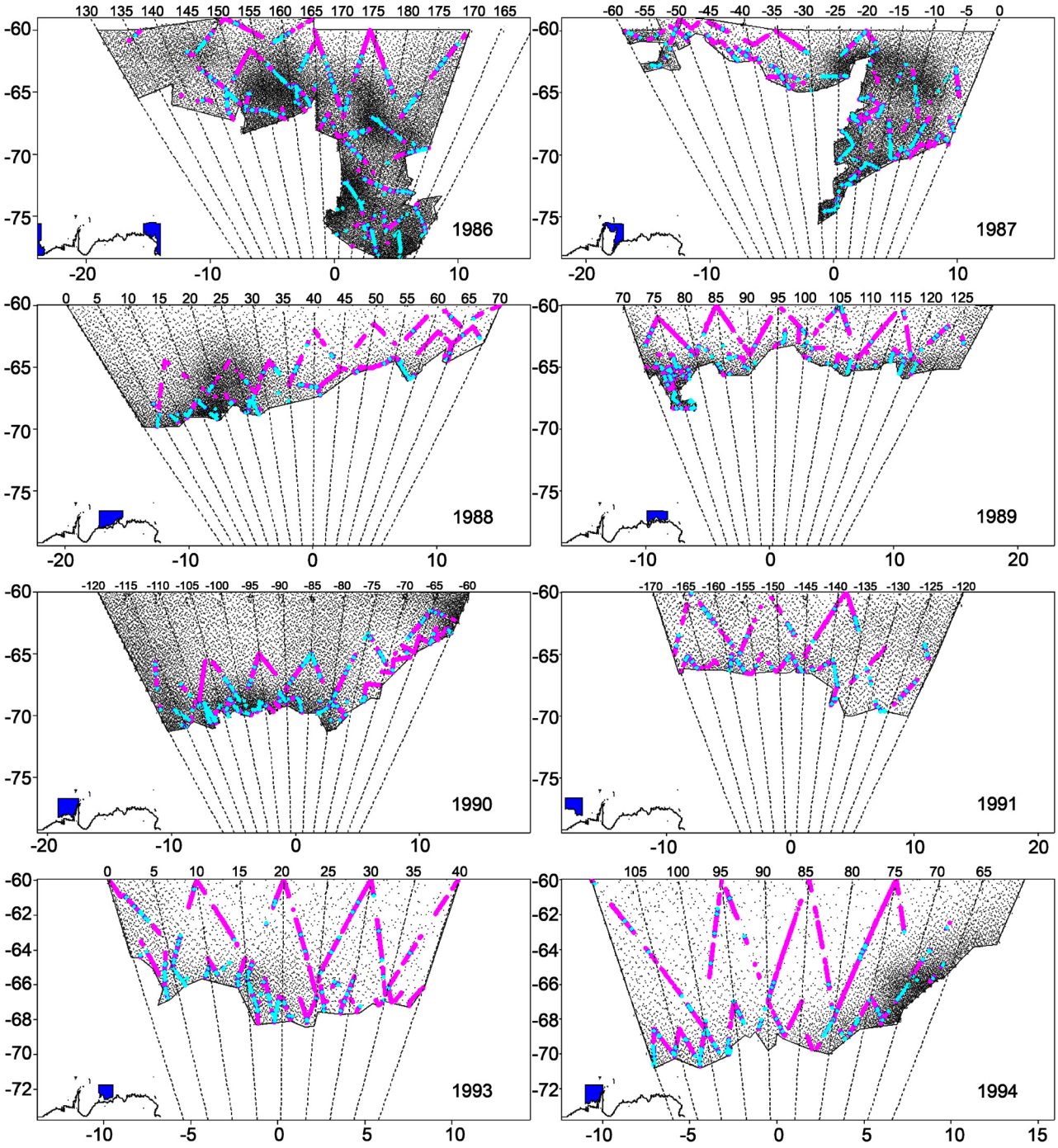

Figure 3A. Spatial distributions of AMWs during SOWER by year. One dot represents 10 whales. IO-mode effort in pink, sighting locations in turquoise. Each year's general location is shown by the blue area on the small inset maps. The main maps are equal-area sinusoidal projections around the mean longitude for the year; 1 unit on horizontal and vertical axes equals 60 nmi (one degree of latitude), and diagonal dashed lines show approximate meridians. Figure 3A: 1986–1994.

not always closer to the ice edge. When summarised by survey stratum (see Table 3 from Bravington & Hedley [2009]), all means are between 1.13 and 3.25. Our model is evidently not generating implausibly-extreme expected school sizes nor wild spatial fluctuations; the former is in fact guaranteed by the statistical construction in Section 2.2.

School-size comparisons between CP2 and CP3 are somewhat hard to interpret because of the shifts in ice-edge and changes to stratification. Nevertheless, a rough comparison of the aforementioned stratified means confirms the impression that average school size was somewhat lower in CP3, as is the case in Figure 4.

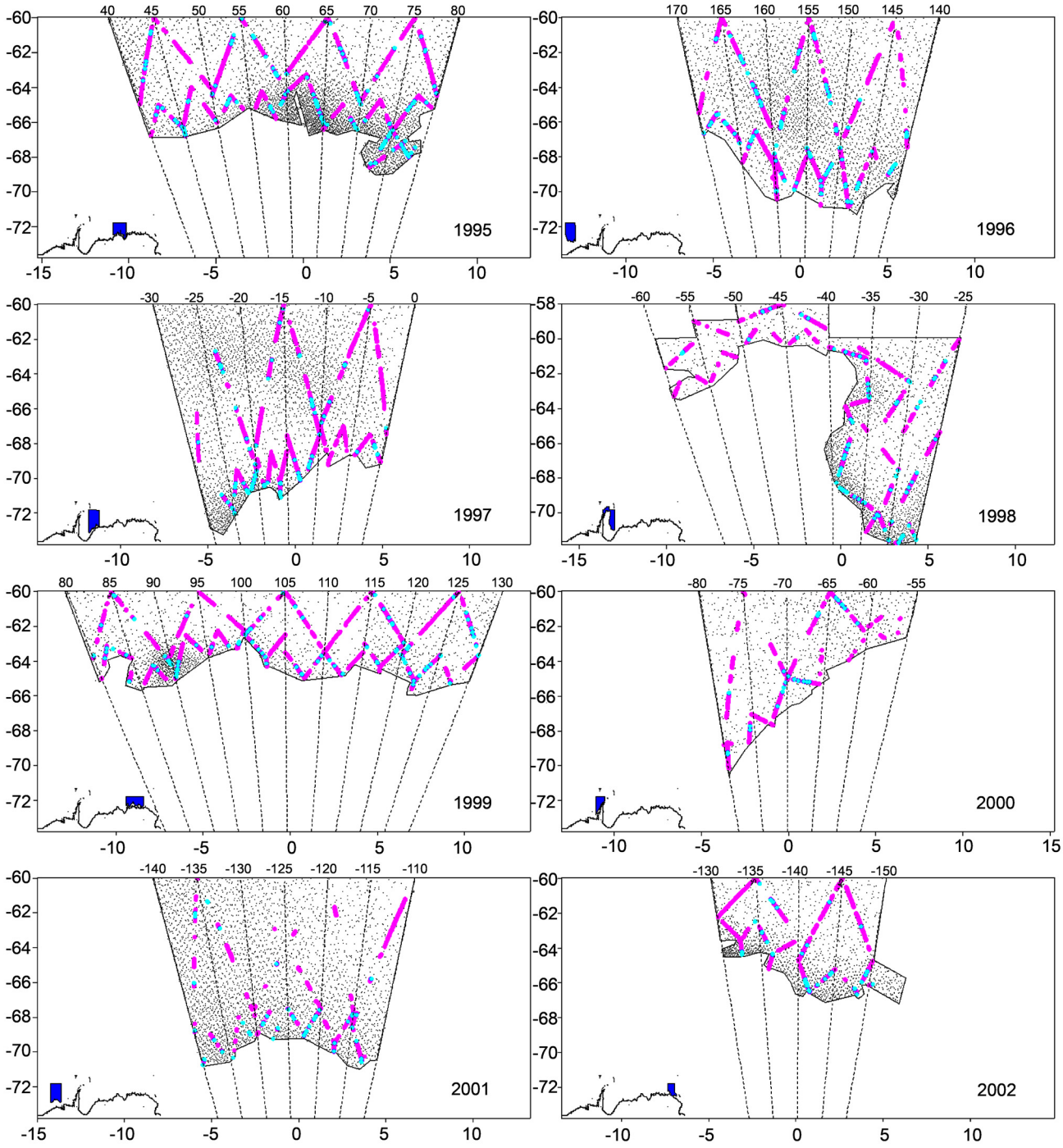

Figure 3B. Spatial distributions of AMW by year: 1995–2002.

4.3 Implications of modelling/data choices

Table 7 shows the sensitivity of SPLINTR estimates to some of the modelling choices ('factors') in Section 3, and to the impact of using a conventional stratified abundance estimate (with its potential for bias due to uneven effort distribution, etc.) instead of a spatial smoother. These are simple one-at-a-time changes from our base case, because there is little reason to expect strong interactions between the factors. When the IWC agreed on consensus abundance estimates, it opted, for pragmatic reasons, to apply adjustments at the CP level, rather than for example, MA-within-CP, even though many such adjustments (most obviously, stratified vs. spatial model) would vary spatially. The numerical sensitivities depend slightly on how comparisons are computed (e.g., one factor at a time, or cumulative; medians or means across strata, or aggregated), but the overall patterns are consistent.

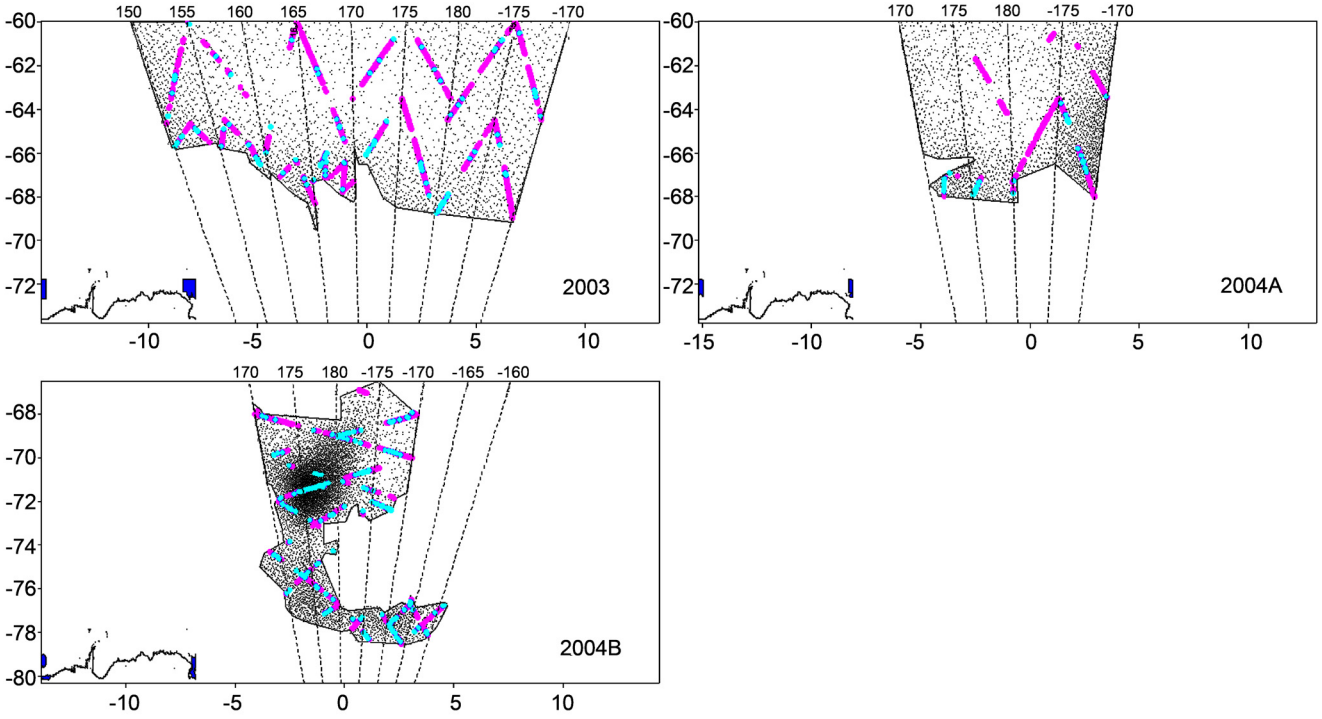

Figure 3C. Spatial distributions of AMW by year: 2003–2004.

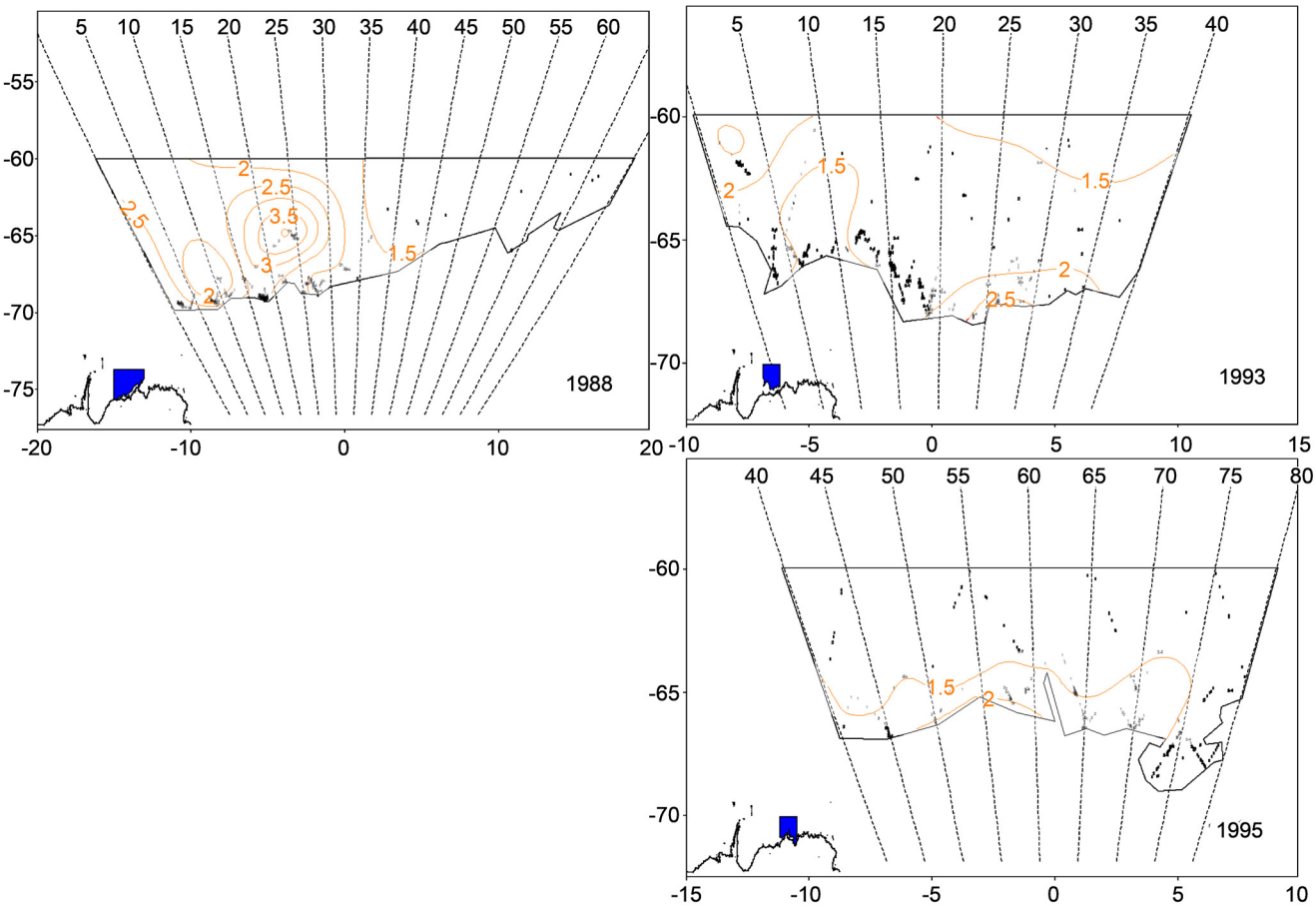

Figure 4. Spatial distribution in mean school size within MA3 for 1988 (CP2), and for 1993 and 1995 (both CP3). Contours show estimated mean school size in the population, not in observed schools. Black symbols show actual sightings. Note the difference between the ice-edge position in the western part between CP2 and CP3. The eastern longitude boundary was also different in CP3, for logistic reasons. See Figure 3 for map information.

1. Stratified estimates for AMW are biased high relative to spatially-smoothed estimates by 10–15%, more so in CP2 where coverage was less uniform.

2. Pre-extension has only a small effect, as one might hope (pre-extension increases measured effort while not changing the number of sightings, so will decrease abundance).

3. Changing from Sightability to Beaufort led to about an 8% increase in the SPLINTR estimates. This was a surprise, because we expected Sightability to be the better choice for capturing 'heterogeneity of detection probability', a phenomenon known to lead to negative bias when trackline-detection is uncertain (Borchers *et al*., 2006). However, the massive complexity of SOWER and the scope for unexpected interactions (e.g., the effect of Sightability/Beaufort on school-size uncertainty; and spatial differences between Sightability and Beaufort, potentially correlated with spatial differences in animal density) make it hard to draw firm conclusions.

4. The biggest effect in Table 7 is that changing the perpendicular-distance definition from 'Firstperp' to the IWC's preference 'Lastperp' leads to a 16% drop in SPLINTR's abundance estimates, but only for CP2; the average change for CP3 is under 1%. As noted in Section 3.2, a defensible case can be made for either definition, and the general expectation was that the choice should not matter much. Further investigation shows that most of the effect is on mean school size rather than school density, but deeper reasons must lie buried in the very complex school-size-dependent detection-probability formulae that are inescapable when modelling SOWER data (whether via school-based or cue-based models). The large size of this effect is the only really surprising result in the sensitivity analyses.

Table 7

Percent effect on abundance estimates (median across strata) of single-factor modeling choices on abundance, in each case by changing one aspect of SPLINTR away from our preferred choice to the alternative shown. 'Stratified SPLINTR' assumed stratum-specific constant densities and mean-school-sizes, without penalising the coefficients. 'prex' means 'pre-extension'.

	CP2	CP3
Firstper ⟶ Lastperp	916	0
Sig ⟶ Beauf	+8	+7
prex ⟶ no-prex	+3	+3
spatial ⟶ stratified	+15	+11

Certain other modelling choices, such as treatment of confirmation status, are hard-wired into SPLINTR which cannot be run without them. Their effects on the OK estimates were investigated in IWC, 2012a Tables 1 and 2 (Palka *et al*., in prep). Most are small, under ±4%, though OK's own preference to take IO-mode confirmation status at face value (whereas we deliberately did not use it; Section 3.4.1) increased their estimates by about 10%. The two different treatments of CL-Unconfirmed school sizes (Section 3.4.2) led to a smaller difference, about 4%.

The most important modelling choice for SOWER AMWs is how to handle g_0. This turns out to have a bigger impact than any of the factors above, but it is so fundamental to model construction for AMW abundance estimation that it cannot really be described as 'sensitivity analysis'. We return to this in the Discussion.

4.4 Diagnostics and simulation performance

The SPLINTR fits were satisfactory in terms of all the distance-sampling diagnostics used by the IWC to evaluate SOWER AMW models (omitting cue-based diagnostics such as forward-distance, which of course cannot be tested in a TCI model). Figure 5 shows one example; a full set can be found in Bravington & Hedley (2009).

SPLINTR's performance on simulated datasets was also generally satisfactory (Palka, 2010; Palka & Smith, 2024). In fact, in the more complicated scenarios, SPLINTR was roughly unbiased overall. Our interpretation is

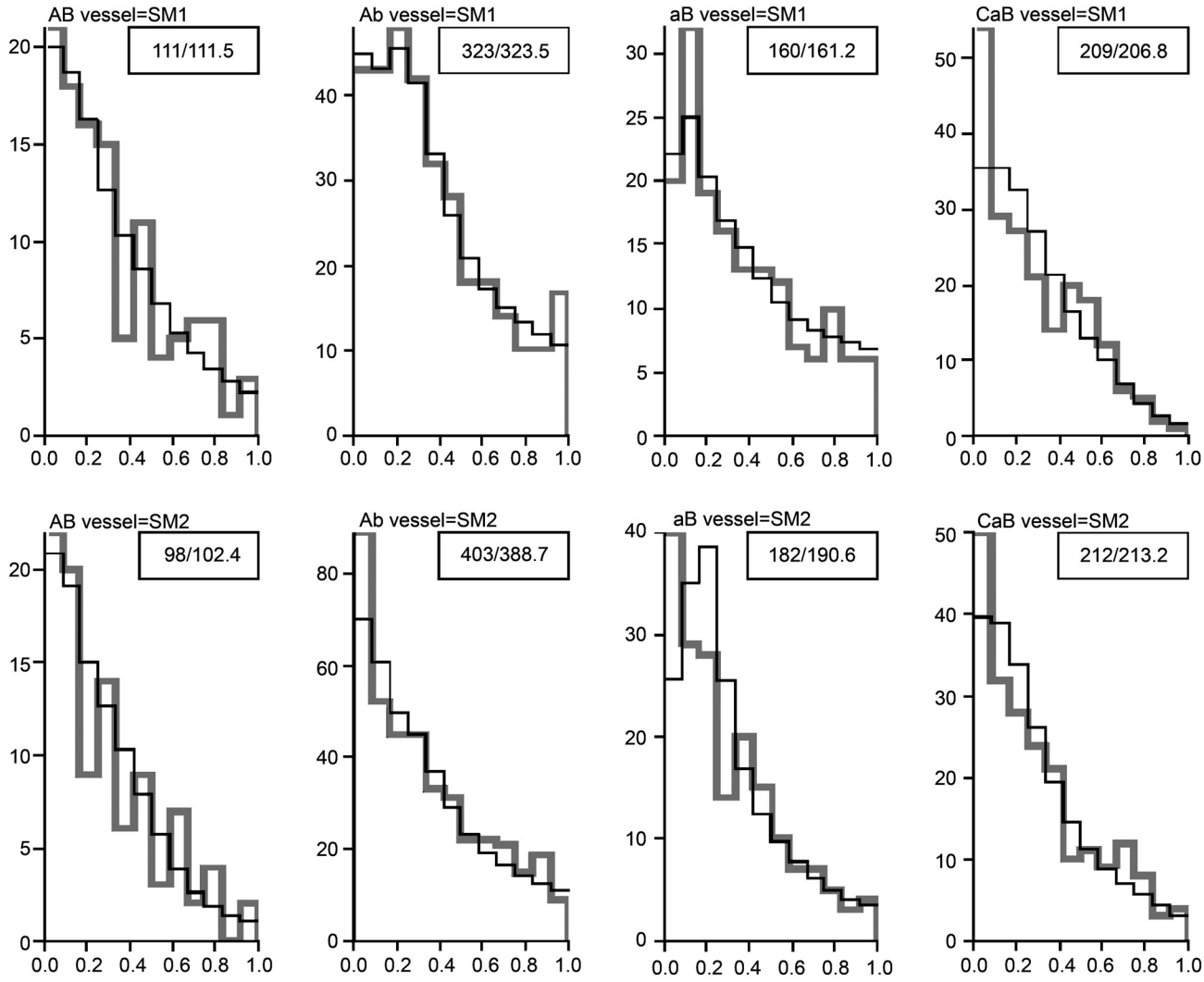

Figure 5. Example of SPLINTR's diagnostic plots for SOWER data. This particular one shows the observed and expected (from fitting SPAMASSS) breakdowns of IO-mode AMW sightings in CP3, jointly by platform-combination and perpendicular distance (expressed as a fraction of the truncation distance of 1.5 nmi), conditional on vessel. Numbers in boxes are observed/expected totals (i.e., sum of the bar heights) within each graph.

that this comes mostly from a rather fortuitous cancellation between an appreciable negative bias of 10–15% due to TCI (an imperfect way to handle g_0), and a positive bias of similar magnitude due to ignoring distance/angle measurement error. These more complicated scenarios were presumably the most realistic for AMWs in SOWER, although they do not fully capture all important nuances, particularly with respect to diving behaviour (Section 5.1).

The differences between SPLINTR and OK estimates for these simulated datasets were substantially less than the difference for the real SOWER data. As noted during the IWC's Scientific Committee in 2010 (SC62): 'the Subcommittee was now in a position where with one set of estimates alone, neither (OK's nor SPLINTR's) performance in the simulations and the diagnostics would raise sufficient concerns to fail to accept the estimates, but the fact that the (OK and SPLINTR) estimates themselves were so different was problematic' (IWC, 2011).

We found one other diagnostic tool to be very useful in the DOSS step: comparing observed and expected numbers of sightings post-stratified by covariates in various ways, e.g., by weather-related covariates or by subregions of the area. We developed code that allowed totals to be compared for any desired subset, which greatly facilitated model checking. This idea has been incorporated into beta-versions of the `dsm` spatial-modelling software (5.6). This DOSS diagnostic led us to make two important modifications that substantially improved the final model:

1. Grouping by Sightability showed a DOSS misfit in early versions of SPLINTR, even though detection probabilities had been estimated using Sightability as a covariate. Basically, the relative overall detection probabilities were not very accurately estimated just from SPAMASSS, which may reflect TCI problems or general model mis-specification. This motivated development of the Z-change model (Section 2.2.4) to look at relative encounter rates over fairly short distances but different Sightability. The Z-change model did not harm the SPAMASSS fit and diagnostics but did alleviate the DOSS misfit.
2. For the Ross Sea (MA5) in 2004, our first spatial model led to a significant overall excess of predicted sightings. On inspection, there was a sharp jump in the fitted spatial distribution— something that smooth models find hard to handle, by definition. In this particular year, sea ice initially prevented the vessels (and probably most whales) from entering the Ross Sea itself, before the ice parted to allow access in the second part of the survey. Fitting two separate spatial models (which is particularly simple when soap-film smoothers are used) alleviated the lack-of-fit. However, it remains an open question whether double-counting of whales has inflated abundance estimates in the Ross Sea in CP3.

5 DISCUSSION

The SOWER AMW dataset is unique in LT/DS for its huge sample sizes and largely consistent protocols designed to accommodate difficult phenomena, such as school-size-bias and uncertain detection on the trackline. It presents huge opportunities for refined analysis, but also huge challenges. Our own involvement with SOWER lasted for over 15 years, including two cruises (SLH). When we began working together, in about 2002, we drew up a list of the main methodological challenges with SOWER (see Section 1.2). All these challenges apply in part to other cetacean LT/DS surveys, so a big part of our motivation was the chance to develop general-purpose methodological solutions. Perhaps the most enticing was spatial modelling. Although a few spatial models for line-transect data already existed by 2000, some had led to nonsensical fits, and there was no satisfactory way to deal with overall variance estimation (i.e., including uncertainty in the detection function). On paper, spatial modelling seemed like the hardest part of the project, but in fact it was probably the 'smoothest' part (largely thanks to the involvement of Simon Wood), and a solution was more-or-less in place by 2006.

In practice, more of our time overall went into dealing with other aspects of SOWER:

- Construction of an appropriate detection function when detection-on-the-trackline is not certain, when school size is measured with error, and when three semi-independent platforms were operating in two distinctly different modes;
- Data cleaning;
- Investigating the difference in abundance estimates from different methods and reaching a reasonable consensus estimate.

The process of reaching agreed estimates and undertaking the investigations required to give some confidence in those estimates (at least relative to the 'standard method') took 10 years. The final results could not have been obtained without several experimental cruises (SSX over several years; BT; and VDT). Although the length of this process caused some frustration within the IWC, we see it as a largely unavoidable, positive exercise to develop our understanding of these and other LT/DS data. It was particularly valuable to have several competing methods (in the end, just OK and SPLINTR). Either method on its own seemed to pass all the diagnostic checks, at least until about 2012, when some further checks were added and the VDT was re-examined, at which point both methods showed some problems. Hence, if just one method had been available, presumably its results would have been agreed much earlier, giving quite different agreed estimates.[21]

Notwithstanding these efforts, we believe that Point #1 above is an area where further methodological developments might lead to improved abundance estimates (not necessarily for SOWER itself, but at least for cetacean line-transect abundance estimation in general); see Section 5.1.

[21] In fact, the first 'serious' SPLINTR estimates (Bravington & Hedley, 2009) were only 5–10% different from the final agreed estimates, but those SPLINTR CP3 numbers were inflated by a misfit to the Ross Sea data.

Given the need to present defensible estimates within a limited timeframe, the IWC-agreed abundance estimates are a reasonable summary of what has been learnt about AMW abundance during CP2 and CP3, taking account of what we now understand about the numerous factors that make abundance estimation in SOWER unusually difficult. Although one might debate the details, it is hard to see how to substantially improve the agreed numbers using only the ingredients of the 2012 OK and SPLINTR models. However, the IWC-agreed estimates were obtained only by some awkward post-hoc hybridisation (Wells, 1896) using CP-wide correction factors (i.e., the same correction was applied to each MA-estimate within a CP series), which has some adverse consequences. For example, there cannot be any coherent way to estimate variance for specific sub-regions, partly because there is no underlying spatial model, partly because the (estimated) correction factors are themselves correlated with the abundance estimates, and also because some of the correction factors (e.g., bias from not using spatial models) should, in principle, be applied at a local rather than circumpolar basis. In the rest of this Discussion, we comment on the challenges of the SOWER AMW dataset, the lessons for LT/DS surveys in future, and on the positive outcomes from the development of SPLINTR for spatial modelling of other datasets.

5.1 Detection-probability models and the effect on abundance estimates

The SPLINTR abundance estimates in Table 6 are fairly close to the IWC-agreed versions: about 10% lower overall. Modelling choices can increase or decrease that difference by a few percentage points. For example, in the 'alt SPLINTR' row of that table, with modelling choices closer to the IWC's preferences, the overall difference is under 5%. Such differences are unremarkable given the simulation results (Palka & Smith, 2024).

However, a year before the IWC's estimates were finally agreed, there was still a difference of some 30–40% between the estimates from SPLINTR and OK, even using the most-comparable modelling choices (IWC, 2012b). It became clear that most of this difference must ultimately be attributable to g_0-related issues, but the size of difference was far larger than could be explained from simulation results. Hence, something must have been missing in at least one of three places: the simulations, SPLINTR, and/or OK. In the end, it transpired that all three were sensitive to flawed assumptions about diving behaviour. This realisation led to substantial adjustments in the OK estimates, giving results close enough to SPLINTR to justify the final consensus estimates.

The obvious first place to look for a problem is the TCI assumption in SPLINTR, which is known to lead to negative bias in abundance estimates when, as with whale blows, cues are discrete. If each school (including singletons) makes a large number of cues within the forward-sighting time window (even if each cue is individually unlikely to be seen), then not much bias is likely to result from assuming TCI. On the other hand, if the number of potentially-visible cues is highly variable between one school and another of the same size, then there is substantial unmodelled heterogeneity between schools, and TCI-based abundance estimates will be negatively biased. (In contrast, modelled heterogeneity – e.g., through school size and sighting conditions – should not be a substantial source of bias.)

When we planned SPLINTR, we chose TCI for two reasons:

1. The forward-sighting window in SOWER, at least 8 minutes for schools of size 1,[22] is quite long relative to Antarctic minke whale dive intervals, which we expected to average around 90 secs (based on Joyce *et al*., 1988, and informally on evidence from North Atlantic minke and other baleen whales). On that basis, most solitary AMWs on the trackline would make at least four surfacings during the forward-sighting window. It did not seem likely that the window would be sharply 'peaked'; so precisely where the three, four or five sightable cues were made relative to the vessel, seemed unlikely to introduce much heterogeneity. Bigger schools with more cues should be even less heterogenous.

2. Two-platform TCI models are relatively simple computationally (Borchers *et al*., 2006), compared to cue-based models such as Skaug & Schweder (1999). For cue-based models, any type of diving behaviour other than IDEDD (independent dives of exponentially-distributed duration) or extremely long dives (Okamura *et al*., 2012) raises huge mathematical and computational difficulties, and, to our knowledge, still has not been successfully attempted.

[22] For school size 1 in CL-mode, the 75th percentile of first sighting distance (across Platforms A and C) is 7.2 mins.

As to the latter, the 'relative simplicity' of TCI turned out to be illusory for SOWER, where there are three platforms and two different modes. Two-platform TCI is simple because it requires only two estimable conditional functions, in addition to the overall perpendicular-distance function $\mathbb{P}[y \mid A \cup B \cup C]$. But for the three-platform two-mode version, and after considerable effort to find an appropriate mathematical decomposition, it appears that four conditional functions are necessary (Appendix A.2). In some sense, the TCI model is overparameterised relative to a cue-based model when more than two platforms are involved, because there is no way to tell TCI about commonalities that should apply (at the cue level) across sighting conditions or school sizes. The code that computes the log-likelihood for the 2012 IDEDD cue-based SPHAZ variant of SPLINTR is in fact shorter than the code for SPLINTR itself. However, considerable preparatory effort was needed to get the cue-based data into shape for SPHAZ (to prevent whales from moving backwards in time, etc.), and concerns regarding data cleanliness remain, particularly for methods beyond the distance-sampling type of analyses earlier envisaged. In addition, SPLINTR runs 3–4 times faster than SPHAZ because of the extra computations and data used in a cue-based formulation. Overall, TCI does seem somewhat simpler than IDEDD cue-based for SOWER AMW (and it is not obvious which one is actually better in practice), mainly because of the data requirements, though it is not as simple as we had hoped.

As to the former and aside from our informal thinking, simulation studies (Palka & Smith, 2024) did seem to suggest that the bias from TCI would be fairly small. Note that the diving patterns assumed in the simulations were not disclosed to model developers while they were still developing their methods.

The VDT data, examined in detail in 2012 (and not available when the SOWER simulations were first developed), differed somewhat from prior expectations, with longer dives up to six minutes, punctuated by several short dives, although (1) dive patterns did vary between schools of the same size, and (2) it proved very difficult to successfully track solitary animals, so that the school-size-1 dive patterns which *were* observed may be unrepresentative precisely because they *could* be observed. (Subsequently, Friedlaender *et al*. (2014) VHF-tracked two AMWs in close-to-ice-edge waters, and the two animals showed different patterns to each other; one dive of over nine minutes was recorded.) The longer-shorter pattern suggests that TCI bias could be more substantial than we expected. If many solitary animals will make only one group of cues during the forward-sighting window, then the overall chance of a solitary animal being seen (for any platform) varies considerably depending on whether its cue-group occurs say four minutes ahead of the vessel (0.8 nmi) or two minutes ahead (0.4 nmi), or indeed if there are two cue-groups at seven minutes and one minute ahead.

By the same token, though, the VDT are incompatible with the IDEDD model generally assumed by cue-based detection-probability models, including OK. A key point in SOWER diagnostics was the discovery in 2012 of a strong misfit between observed and expected simultaneous duplicates in the cue-based OK model (61 observed vs. 107 predicted in CP2; 58 vs. 108 in CP3; IWC, 2012b, Section 5.3.2, 'Diagnostics'). This led to examination of the estimated mean dive-times in OK, which were clearly inconsistent with VDT. The solution adopted was to use the OK model with mean dive-time estimates constrained to match values taken directly from the VDT. While this seems to be the most reasonable short-term solution, its implications for detection-probability estimates are not clear, for several reasons:

1. Sample size of dive-times is very limited for singleton schools – just two or three such schools – and, as noted above, that sample may be biased.

2. An IDEDD model still has to assume independent-exponential dives and bases its overall sighting probability per school on that assumption. A regular pattern of one long dive of more-or-less fixed duration followed by several short dives generating a group of cues (not that all schools will follow the pattern) is in one sense less heterogeneous than would be inferred from an IDEDD framework with the same mean dive-time, because the number of cue-groups in the (SOWER) forward-sighting window cannot vary beyond 1–2 (whereas an independent-exponential model predicts that there will often be 0 or 3 or even more groups). On the other hand, unless the probability of cue-detection is fairly uniform across the forward-sighting window, additional heterogeneity will be generated because of variability in *where* the cue-group(s) occur. Added to this is the complication of not actually knowing the true school size for many IO sightings on which these models have to be based. Thus, it seems to us quite

unclear whether an IDEDD model with constrained mean-dive-times will over or under-correct for heterogeneity due to surfacing patterns.

3. Certainly, a TCI model for SOWER (with say an eight minute forward-sighting window, but with cue detection probability varying within that window) seems likely to have some negative bias for schools of size 1 or 2, given frequent longer dives of the order of five minutes, but it is not obvious that the extent of bias can be predicted from an IDEDD-type approach.

Cue-based models have been successfully used for minke whales in the NASS North Atlantic surveys, and the IDEDD approach was validated there against simulations based on fairly extensive datasets of dive-patterns (Skaug *et al*., 2004). Why do cue-based approaches seem to be so much more difficult for the SOWER AMW dataset? There are several reasons:

1. The forward-sighting window is much longer in SOWER (with binoculars) than in NASS (with naked-eye searching), so more cues are available per school. On the positive side, this means SOWER might possibly have enough information to infer relevant diving behaviour 'internally' (i.e., without additional experimental data), but, as noted above, this information may be misleading when cues are clustered. On the negative side, the long sighting window means that details of the surfacing pattern over longer periods become important. For example, if the window was so short that only one cue would 'fit' and thus all cues in the window would be equally sightable, then only the *average rate* of cue generation would affect detection probability, rather than the details of the entire dive pattern.

2. Duplicates and long dive times: if a single AMW-like cue is seen seven minutes ahead of the vessel, and another cue is seen later just one minute ahead, is it the same whale or not? A six-minute dive is possible and, of course, gives ample time for substantial whale movement. Such sightings are flagged as 'possible duplicates' but, in effect, are treated as separate schools in the agreed estimates where only 'definite duplicates' are treated as such, leading to some positive bias. With three platforms operating, possible triplicates should also be considered. These may comprise a true duplicate and a singleton, or various other possibilities, and are hard to disentangle in the data. Unlike NASS (where the forward window is shorter so that this issue may be of less concern), there are no explicit resighting data in SOWER to allow automated duplicate-identification *post hoc*. The status of possible duplicates is an issue for any SOWER abundance estimate but may pose a particular problem for cue-based methods because of 'aliasing' between duplicate status and long dives.

3. SOWER has limited, and possibly non-representative, external data on cueing patterns; note that these had to be obtained by dedicated experiments (VDT) after the conclusion of the main CP surveys.

4. Substantial rounding errors in timing, especially in CP2 where times were rounded to the minute. An AMW can certainly surface more than once within 60 secs and small errors in reporting of time could split one cue into different minute-intervals. Though this can perhaps be 'modelled away' in principle by building in some allowance for distance/time error (as OK have done), it is not clear how much information would be left for inference in a more-nuanced cue-based model.

5. School size is measured with error in IO-mode. In particular, cues that appear to come from a school of size 1 may really be from a size-2 school (which likely constitute about 15% overall of *reported* size-1s in IO-mode), and the precise cueing pattern of the school may affect the estimate of its size. Further, the probability of school size error is strongly linked to sighting distance (Section 3.1), so that schools seen only once are more likely to be misidentified as size-1. We return to the school size issue in Section 5.2.

Another source of likely negative bias in our estimates is neglect of measurement error in distance and angle, for which allowance is made in OK but not in SPLINTR. Other studies (e.g., Chen, 1998) indicate that such error generally leads to negative bias in distance-sampling-based abundance estimates. That said, measurement error is already incorporated into the SOWER simulation studies (and, unlike with diving behaviour, there is some reasonable

information within SOWER itself to parameterise such errors), and SPLINTR's bias on simulated data with measurement errors was not that large (Palka & Smith, 2024). Hence, differing treatment of measurement error is insufficient to explain the large observed difference between OK and SPLINTR estimates from the real data.

The simulations used to test OK, SPLINTR and IM were very useful in leading to better models. However, they evidently did not address all the important features of SOWER, nor could they be expected to, given what was known when the simulations were created. In particular, the subtle interactions of dive-patterns, school sizes and observational setup in SOWER cannot be fully understood without additional data collection and in-depth analysis of the kind that comes only through developing an abundance estimation model. In our view, these nuances are still not fully understood quantitatively, although qualitatively, much more is now known than when the simulations were devised.

5.1.1 SPHAZ

In order to further investigate possible TCI bias and following some tentative analyses of forward-sighting data, we tried to develop a cue-based IDEDD version of SPLINTR, called SPHAZ. It used the same setup as SPLINTR, except for the detection-probability term. Rather than computing hazard-probability integrals directly as per OK (Okamura & Kitakado, 2010), we discretised the locations of possible sightings onto a grid in forward and perpendicular-distance space (which intrinsically makes some allowance for measurement error) and computed a discrete version of the integral by simple sums and products, as schools 'travelled along' the grid towards the vessel. SPHAZ converged successfully; its abundance estimates were substantially higher than SPLINTR (and, in fact, higher than OK, after allowance for stratification and other differing model choices). It is computationally possible to fit, within about four hours, a spatial smooth to multi-year data with varying school sizes in combination with a cue-based detection-probability model. However, some of the parameter estimates did not make sense, e.g., the direction of Sightability effects. With SPLINTR, we had been careful to parameterise the model to prevent nonsensical estimates, but there was not enough time to devote the same level of attention to SPHAZ.

In the end, we did not develop SPHAZ further. After 10 years of intense work on SOWER abundance estimates, we started to wonder whether the IDEDD formulation made sense for SOWER. If not, there was no compelling reason to do the work required to fully implement and test an IDEDD-based 'spatialised OK' or 'hazardised SPLINTR' model, each of which would likely have had a different set of flaws.

5.1.2 Other options for modelling detection probability

In our view, neither TCI nor IDEDD-based OK can provide fully satisfactory treatment of g_0 (or consequently of abundance) for SOWER AMWs, primarily because of dive behaviour. Any serious attempt to improve estimates would have to engage fully with non-IDEDD diving, as well as all the other SOWER complications we have described. Since 2012, there have been several papers published on non-IDEDD detection-probability models, using hidden-(semi)-Markov frameworks (Langrock & Zucchini, 2011; Okamura *et al*., 2012; Borchers *et al*., 2013; Langrock *et al*., 2013; Borchers & Langrock, 2015). For example, Langrock *et al*. (2013) fit a Markov-Modulated Poisson Process to external dive-pattern data,[23] and combine that with the two-dimensional locations of sightings relative to the vessel. In principle, for some surveys, one might contemplate inferring dive patterns using 'internal' data on resights between multiple platforms. However, we are sceptical about the prospects for that with SOWER data: the protocols were not designed with resighting as an objective and could only record at most two or three resights; there is uncertainty about duplicate ID amongst the longer dives; and AMW dive duration can exceed the typical forward-sighting window altogether. Thus, plenty of external data along the lines of the VDT would seem to be essential. Unfortunately, the VDT experience was that such data are hard to obtain visually for AMWs in open-water, at least for the singletons where such data would be most valuable. The obvious alternative is tagging, as used successfully on AMWs by Friedlaender *et al*. (2014). But these deployments were made close to the ice edge on AMWs that behaved placidly. It is not clear whether tag deployment would be possible with 'skittish' solitary AMWs far from the ice.

[23] This application of MMPP is completely unconnected from the use of MMPPs to handle small-scale spatial clustering of schools (Section 2.3.1), although the mathematical framework is similar.

5.2 Uncertain school size

Perhaps the single most difficult feature of the SOWER AMW dataset for an analyst is that there are no IO-type data where true school size is known.[24] If an unbiased subset of such sightings had been available, we would have implemented a simpler two-stage model, first fitting detection-functions conditional on true school size, then feeding the estimated model parameters into spatial models for school-size and school-density, using variance-propagation (Section 5.6) to carry the uncertainty from the first stage into the second. Although SPLINTR itself is a two-stage model, the split is in the 'wrong place': the SPAMASSS/DOSS split is a computational convenience which should not affect estimates much, but it is not the fundamental simplification that would be possible if some known-school-size IO-type data had been collected.

About 15% of reported size-1 schools in IO-mode were actually size-2, based on the SSX data, so the effect is non-trivial. Detection-function estimation and, more importantly, diagnostic checking become much harder and less reliable when school size is uncertain. The problem is that detection probability clearly ought to depend on true school size (rather than reported), but the prior distribution of true school size will vary spatially, and requires some kind of model, whether stratified or smooth. To compare, for example, observed and expected perpendicular distances for schools of reported size 1, the expected numbers need to be generated from a mixture distribution, a weighted sum using inferred distributions of true school size, which depends on the spatial school size model. This means it is never exactly clear what is really being checked: the school size model or the detection function model? The concern remains that misfits in both models might conceivably cancel to leave good-looking diagnostics but not necessarily cancel for the abundance estimates.

The general idea behind the protocols in CP2 and CP3 was that:

- IO-mode would be used to estimate g_0, detection probability and encounter rate;
- CL-mode would be used to estimate mean school size, allowing for the different overall detectabilities by school size as inferred from IO-mode data.

While this sounds superficially reasonable, it overlooks the subtle complexities, induced by not knowing the true school size when trying to estimate g_0 and the entire detection function. Although these were not readily apparent when the protocols were first designed, Joyce *et al.*'s (1988) review of the first 10 years of IDCR surveys does provide considerable further insight, in terms of both operational logistics and survey design.

One lesson for the future is that it is highly desirable, from the perspective of getting reliable abundance estimates in a reasonable time frame, to use protocols that allow detection-function estimation conditional on *known* school size for a subset of the data. This is not always easy – for example, closing on every sighting to confirm school size leads to serious violations of LT/DS assumptions, unless schools are widely spaced – which is precisely why IO and CL-mode are separated in SOWER. However, it would be possible, at least in principle, with fairly minor modifications of SOWER protocols:

- Using Closing-when-Abeam (as used successfully in post-CP3 SSX cruises) rather than Closing-when-First-Seen (as in CL-mode). The full sighting history as per IO-mode is recorded, but the school size uncertainty is eliminated because closure happens only after the history.
- Ensuring that Closing-when-Abeam has all three platforms in operation, so that independent-platform analyses can be conducted, conditional on the true school size. (The post-CP3 SSX used Closing-when-Abeam but mostly without Platform B (the Independent Observer or IOP), for manpower reasons. That is adequate for estimating school size error, but not for conditional modelling of the detection probability given true size.[25])

With the benefit of hindsight, it seems that having some ideal-type data (i.e., known school size under IO protocols) might have cut several years from the process of agreeing abundance estimates – and generated greater confidence in those estimates – notwithstanding any logistical difficulties involved.

[24] Actually, some IO sightings do have 'confirmed' school size, but these are by construction not typical (Section 3.4.1). A small subset of the post-CP3 SSX data was collected in this desirable way, but not enough to be a basis for abundance estimation.

[25] Hypothetical future SOWER-like surveys might not use three platforms, in which case this point would not apply.

5.3 Data cleaning and dataset complexity

To prepare the SOWER data for 'standard method' abundance estimation (as implemented in the DESS software system; Strindberg & Burt, 2004), the IWC already had substantial data-cleaning protocols in place during CP2 and CP3. The original plans for SOWER led to a database and data-cleaning protocols designed to answer one question with one method. However, as is so often the case, the initial results (a huge drop in minke abundance estimates between CP2 and CP3) generated different questions, e.g.: was this an artefact of imperfect analysis? What was going on with g_0? What about school size changes? To address these questions, it was necessary to develop more sophisticated analyses, such as spatial modelling and cue-based detection functions, which have more stringent data-quality requirements. The development of these methods progressively revealed numerous inconsistencies in aspects of the SOWER dataset which had not previously been investigated.

It is inevitable that such errors will occur in real datasets; the question is what to do about them. During the post-CP3 phase of SOWER, there was no centralised IWC mechanism (nor funding) in place to proactively maintain or upgrade the database for analyses beyond the 'standard method', so it was left to individual developers to come up with fixes suitable for their individual models as and when inconsistencies were found. Unsurprisingly, different developers took different decisions under time pressure; fixing or working around newly-uncovered data errors was a time-consuming task. The resulting discrepancies among datasets (and among related seemingly-minor modelling choices, such as which perpendicular distance estimate to use) was one more unwelcome complication in the lengthy process of exploring differences between estimates. One lesson might be to build in a long-term component of data-maintenance and liaison with developers, even after initial data collection is complete. Since the decision 'to clean or not to clean' can be quite subtle – e.g., what is 'dirt' to one analysis might be precisely the focus of a different analysis – there can be value to making available different 'official' versions of a dataset, ranging from 'raw' to 'squeaky clean but under specific assumptions'.

Because of its large sample size, fairly consistent protocols, and biologically/methodologically challenging aspects, SOWER may still be the most valuable dataset for testing general-purpose developments in LT/DS. Nevertheless, the SOWER dataset is extremely complicated, particularly without a learned guide and extensive time (e.g., Strindberg & Burt, 2004). Any statistician interested in tackling the methodological questions raised by SOWER would need to invest a great deal of time taking decisions about essentially peripheral aspects of the data. We suspect this complexity deterred several statisticians who might otherwise have been interested in contributing to methodological development in SOWER modelling. Some of this is unavoidable and simply must be confronted head-on, but some aspects are connected with data-cleaning and may – at least with hindsight – have been avoidable.

5.4 Interpretation of abundance estimates

Our aim here is to improve the reliability of abundance estimates, rather than interpreting these estimates in terms of population-dynamics. Palka *et al*. (in prep) gives a detailed discussion of the IWC's deliberations. However, the 30–40% drop overall from CP2 to CP3 in Table 6 is so significant (well beyond the statistical uncertainty captured by 'Total' CVs) that some comment is required. We note that all estimation methods have shown similar patterns in their version of the Table. In some versions, e.g., the 'alt SPLINTR' in Table 6, abundance estimates for one or two MAs actually increase between CP2 and CP3, but the overall pattern is always downwards.

It should be remembered that, while the six-MA breakdown is a convenient choice for reporting, it is arbitrary and artificial for AMWs. JARPA genetic data from MAs 3, 4 and 5 show two separate populations within that Hemisphere (presumably extending into the other Hemisphere, where further populations may exist) with substantial year-to-year variability in the boundary between populations (Pastene & Goto, 2016). 'Discovery' tags show that individual minke whales have sometimes moved more than 100° of longitude between years. It is conceivable that separate biological populations might show different long-term trends. While this does not preclude sensible discussions about an 'overall trend', it is not meaningful to talk about trends at any finer scale than population. The MAs are individually too small to represent meaningful persistent population units, but we do not have a clear idea of what larger regions might be appropriate.

There are basically four *a priori* plausible reasons why one CP-set of SOWER AMW estimates could differ substantially from another (see Branch [2007] for some additional implausible ones):

1. Artefacts of analysis, due to changes in survey protocols and/or animal behaviour (e.g., larger/smaller groups) that are not appropriately allowed for in the estimation model;
2. Systematic trends in the proportion of AMWs in high-ice concentrations, thus outside the surveyable regions;
3. Random shifts in AMW distribution from year to year, so that one CP was 'luckier' with where the whales happened to be each year;
4. Changes in true abundance.

In reality, these reasons will always be true to some extent. No estimation method is perfect, and the extent of its imperfections change; environmental conditions vary, both in the long term ('trend') and short term ('random'), and whales move in response; and population numbers are never completely stable. However, some are likely more important than others. How much can now be said about their relative importance?

The upshot of the IWC's work on abundance estimation methods, including this paper, is that Item #1 seems to have been largely addressed. There are some remaining defects: for example, concerning diving behaviour and $g_0(\mathrm{SS}=1)$ (for cue-based as well as TCI-based methods); the possibility of double-counting in some years, e.g., 2004 Ross Sea; and the surprisingly large impact of choosing between 'Firstperp' and 'Lastperp' in SPLINTR (16% in CP2; < 1% in CP3). But it does not seem likely that such defects could explain away the drop. Thus, we believe that the data show substantially fewer AMWs in the regions and times surveyed during CP3, compared to CP2. What this means for abundance is another question.

Items #2 and #3 both concern shifts in spatial distribution. Item #2 relates to negative bias, in particular systematic trend in that bias, and its extent cannot be inferred from vessel-based survey data. The SOWER survey limit of '10% ice cover' is for operational safety and does not correspond to the preferences of AMWs which, unlike larger baleen whales, certainly do frequent icier waters including polynyas.[26] The highest AMW densities encountered in SOWER were often at or near the survey ice limit. It would be surprising if the unsurveyed AMW proportion-in-higher-ice-cover (PIHIC) stayed exactly constant over time. If there is any trend, the key question is whether PIHIC was ever high enough for any change to matter much. Since PIHIC was certainly non-zero during CP2, and we are discussing at least a 30% drop in Table 6, the only way that Item #3 could make an important contribution would be if, in the early 2000s, PIHIC had risen to at least 30%. While the general expectation seems to be that AMW PIHIC is not 'very' large, there is currently no reliable way to quantify PIHIC, either now or in the past. This PIHIC item remains a wild-card in interpreting changes in AMW abundance estimates, though its overall effect is that all the estimates are negatively biased.

The random inter-annual movements in Item #3 (certainly longitudinal, and perhaps also into/out-of higher-ice-cover) play a different role. They are not a source of bias and would eventually become apparent with respect to trend if enough CP sets were completed. However, inter-annual movements have a considerable impact on the uncertainty of aggregate abundance estimates in any single CP, and thus on the uncertainty about trend from a small number of CPs. Fortunately, the associated 'Additional Variance' (AV) can, in principle, be quantified using SOWER or other survey data by looking at variability in longitudinally-stratified survey estimates across years.

As to Item #4, abundance *per se*: the 18-year duration of CP2/3 gives plenty of opportunity for abundance to change appreciably in response to unmeasurable changes in the environment (e.g., abundance changes for prey, predator and competitor species), even though constrained by mammalian population dynamics. It is worth pointing out that post-1986 whaling could not be responsible for a big drop in abundance during CP2/3, since catches were at least an order-of-magnitude lower than natural deaths.[27] For AMWs, there are other data sources besides abundance estimates which should be less susceptible to short or long-term spatial effects, such as

[26] Polynyas are open-water zones surrounded by thicker ice. They were inaccessible to SOWER, but not always to AMWs.

[27] This is clear even using very conservative assumptions: low abundance (e.g., 400,000) and high average lifespan (e.g., 40, corresponding to a natural mortality rate of 2.5%), which would imply at least 10,000 natural mortalities per year.

changes in age-composition and age-at-maturity. These might provide some help in disambiguating spatial effects from genuine changes in abundance, via population-dynamics modelling (Punt *et al.*, 2014), although their statistical power for AMWs seems rather limited.

Population-dynamics reconstructions (Punt *et al.*, 2014) might be of some help here, taking account of other data besides the abundance estimates. Nevertheless, for AMW, they can only cover half of the Antarctic (i.e., where JARPA survey data are also available); and, because of their intrinsic time-adjustable parameters, e.g., for carrying capacity, they will never strongly contradict the main signal about abundance trends, which continues to come directly from SOWER estimates.

The appropriate way to investigate Items #3 and #4 (plus any contribution from #2) through SOWER alone is jointly via an AV model (Additional Variance; e.g., Skaug *et al*., 2004 for historical versions), as implemented for AMWs by Kitakado & Okamura (2009). The inputs to an AV model are abundance estimates and (co)variances as per Table 6 from each longitudinal block (e.g., 30°); the parameters include trend terms, as well as noise terms for annual variability in true abundance within any longitudinal block. Kitakado & Okamura (2009) found considerable AV for SOWER AMW; thus, the CVs in the 'Total' column of Table 6 need to be substantially increased to capture the uncertainty about true abundance (Palka *et al*., in prep).

While the primary goal of an AV model is to estimate the AV, the model also needs to include trend parameters in order to estimate AV properly, so these trend estimates and their precision can be examined as well. Modern statistical tools make it fairly easy to fit AV models, but the challenge for AMW is perhaps in deciding exactly which models to look at. For example, 'trend' might be one overall parameter, or a separate parameter for each distinct population (if their ranges were known in advance), or an overall term plus regional random effects which might be independent or spatially correlated, and so on. (The AV models in Kitakado & Okamura [2009] do include the single-overall-trend version, but the authors report only the associated AV estimate, not the actual trend estimate or its standard error.) Any such AV model could explain the observed abundance estimates, simply by adjusting the Additional Variance parameter itself; doing so has only a limited effect on goodness-of-fit. With so few replicates, the CP2/3 data have little statistical power for choosing between different AV models. This decision should therefore be guided by *a priori* biological knowledge about movements and population structure and by the questions that are really of interest. 'Biological knowledge' is not much help for AMWs: we know there is more than one AMW population, that their boundaries shift, and that the MAs are individually too small to represent meaningful units, but that is not enough to uniquely specify an appropriate AV model. Thus, it becomes important to carefully formulate a question-of-interest – e.g., how best to summarise trend(s) in a statistically-measurable way? Is the main focus just on characterising uncertainty in absolute abundance in a certain year? Such issues are beyond the scope of this paper; see Palka *et al*. (in prep).

5.5 Considerations for future absolute abundance surveys in AMW-like settings

One goal of SOWER was to obtain absolute abundance estimates of AMWs, and the survey protocols of CP2 and CP3 were set up with that in mind. After one circum-Antarctic iteration (CP1), it was clear that the abundance was, by any measure, large. By the end of CP3, the focus had evolved into whether there had been a large change in abundance. That question – which arose retrospectively – raises all sorts of complications connected with inter-annual shifts in whale distribution, which it would have been difficult for any conceivable survey to address in such a small number of iterations. However, from the narrow perspective of assessing how many whales were in the survey area, there are technical lessons that to be drawn for future whale abundance surveys, not restricted to AMWs.

The first observation is that, for AMWs under SOWER sighting protocols, the development and testing of detection-function models – even simplified approximations like TCI and IDEDD cue-based – was a long and difficult exercise. Presumably that would apply *a fortiori* to any modification that tried to fix the difficulties with TCI and IDEDD. The review of early SOWER (then IDCR surveys) by Joyce *et al*. (1988), based on 10 years of Southern Ocean surveys and over 2,700 AMW sightings, is a salutary read for anyone contemplating distance-sampling for cetacean abundance. It shows that the IWC's Scientific Committee was, by then, well aware of many of the complications that could arise when estimating absolute abundance of AMWs. A number of experiments had been tried prior to 1988 – some sensible, some less so – but in the end, many issues were left unresolved,

as subsequent surveys, right up until the end of CP3, focused on consistency of protocol and achieving spatial coverage. In our opinion, these unresolved issues contributed substantially to the delay in agreeing 'final' SOWER AMW abundance estimates after the end of CP3. Notwithstanding the crucial experimental data gathered post-SOWER, the 10-year gap between 'end of CP3' and 'agreed estimates' put considerable strain on many analysts who had run out of 'official' time to work on or review SOWER. As already noted, one difficulty was the lack of a long-term data-cleaning mechanism that could respond to changing demands of analysis.

Related to this is the difficulty of devising adequate diagnostic tests for such a complex survey. For AMWs, the IWC went through a diligent diagnostic – and simulation-checking process for all models, perhaps more so than for any other cetacean survey. This process was certainly useful for development but ultimately did not resolve the large discrepancy between the OK and SPLINTR estimates. Nor could it resolve, for example, the complicated interaction between school-size uncertainty, perpendicular distance and detection probability, because the underlying data were not available. Furthermore, there were no diagnostic problems with using 'Firstperp' (or 'Highestperp') perpendicular distance, even though this choice led to a substantial 16% shift in abundance estimates for CP2 compared to the IWC's eventual choice of 'Lastperp'.

The IWC's collective process to 'agree a number' (more accurately, to agree on a method to produce an abundance estimate) was diligent, lengthy and difficult. Nevertheless, it was only at the last moment that the remaining researchers collectively noticed a key problem, in the simultaneous-duplicate diagnostic for hazard-probability cue-based models, leading to adjustments in the OK model which ultimately gave estimates that differed from the most-comparable SPLINTR numbers by less than 20%, more-or-less within the expected margin (Table 6). The point here is not the specifics of the diagnostic but rather its magnitude: a misfit between predicted and observed of about 110 vs. 60 in each CP series. The difference is highly significant ($\mathrm{p-value} \sim 10^{-6}$), the effect size is large, and so are the implications for AMW abundance. In one sense, this is a success for the collective process: a substantial and unforeseen problem was eventually detected, and subsequently fixed (though the fix is not perfect; see Section 5.1). But at the same time, it is a cautionary tale for LT/DS surveys in general. Simulations did not pick up the fundamental diving-behaviour issue, presumably because, when the simulations were set up, there was not enough external data for proper ground-truthing. To obtain such a clear discrepancy required a total sample size of several thousand sightings in each CP series, which took 6–12 years to collect. A more typical survey might have one-tenth the number of sightings, such that a predicted::observed discrepancy of 10::6 would not alert much suspicion.

The lesson here is that diagnostics and simulation-testing are certainly important and *can* be useful; but their power is limited, and they cannot be relied on to fix everything. This is especially true when survey protocols necessitate complicated models (like SOWER), and also in surveys with only moderate sample sizes (unlike SOWER).

Our overall impression remains that, when $g_0 < 1$, it is best to bypass the need for complicated models by using survey protocols that allow simple Buckland-Turnock-type estimators, and by ensuring that a sufficient sample of known-school-size sightings are obtained under the same protocols to allow direct estimates of detection probability conditional on true school size (e.g., closing-when-abeam). BT avoids both the complexities of cue-based estimators amid non-IDEDD dive behaviour, etc., and the hard-to-quantify bias from TCI. It may not require external dive data, unless dives often exceed the forward-sighting range of the tracker platform. Despite this, BT itself has certain logistical requirements which are not necessarily easy to meet in practice: in particular, the requirements for adequate forward-sighting spread in the tracker platform, and adequate separation between the field-of-view of the tracker from that of the primary observer. For example, it is not clear how far these requirements were actually met in the 'BT-NSP' experimental data, which led to independent estimates of g_{0A} that were, if taken at face value, encouragingly consistent with our TCI-based estimates (Section 4.1).

Any continuation of SOWER into 'CP4' would have had to re-design its protocols and logistics, to avoid facing exactly the same interpretational problems when the next set of abundance estimates were due. The 'BT version 2' experiments (IWC, 2008; Burt *et al*., 2012) tried to do that, with the platforms and protocols changed to the following:

- Platform A (the 'Top') acted as the 'Tracking' platform, with two observers searching a narrower field of view (60° either side of the trackline) with 7 × 50 binoculars;

- Platform B (the 'IOP') acted as the 'Primary' platform, with two observers searching with naked eyes;
- The role of observers on Platform C (the Upper Bridge) was to assist with tracking sightings made by Platform A and to help determine duplicate status.

In principle, the trade-off is sacrificing overall number of sightings for much-simpler and more-robust analysis. These trials showed some promise, but it remains unclear whether a protocol consistent with BT assumptions could in fact be implemented for AMW surveys using *Shonan Maru*-like vessels.

Of course, beyond any technical issues of surveying, the elephant in the AMW-abundance room is the spatial distribution of AMWs, and how it relates to the regions actually surveyed (Section 5.4). It is clear, both from SOWER and JARPA data, that AMWs shift their distributions longitudinally from year to year. There is at least some hope of inferring the associated 'additional variance' by conducting repeat surveys. However, if the shifts are large, the inescapable price of reliable abundance estimates might be many more replicate surveys, and a reduced ability to infer trends 'quickly'. More intractable is the PIHIC issue (proportion of AMWs inaccessible to survey due to ice cover). It might be possible somehow to obtain recent estimates of that proportion, e.g., via aerial surveys or satellite tagging. If the proportion turns out to be substantial, better high-ice coverage would need to be designed into any future AMW survey. Without any information about PIHIC, any further AMW surveys would be a gamble.

Compared to Joyce *et al*. (1988) at the start of CP2, we now have much better understanding of the issues around AMW abundance estimation and its intrinsic limitations. Any similar surveys in future would need to be explicit about objectives before embarking on design, and perhaps about collecting more background information. Certain questions (for example, about local trends) may be unanswerable from short non-synoptic data series, even with the best protocols and high resourcing. How important it is to answer all such questions quickly and definitely is a separate issue. For example, when developing the RMP and AWMP for management of whale catches, the IWC deliberately stepped away from expecting rapid binary outcomes (e.g., 'statistically significant' or 'above/below threshold') in favour of gradual, long-term responses to a growing dataset, on the multi-decadal timescale of whale population dynamics.

5.6 Legacy of SPLINTR

Developing SPLINTR entailed a number of theoretical and computational advances relevant not just to cetacean abundance estimation but also to smooth-surface modelling in general. During the early years, we worked closely with Simon Wood (developer of the widely-used `mgcv` software for GAMs), such that several current features of `mgcv` grew directly out of these discussions. No model devised for SOWER AMWs is ever likely to transfer directly to any other setting, due to the unique features of SOWER sighting protocols and the impossibility of completely disentangling school-size-uncertainty (a general issue) from detection-function-fitting (often a survey-specific issue). Nevertheless, elements of SPLINTR are highly transferable, as was always our intention. In particular, the `dsm` software (Miller *et al*., 2013; Miller, 2025) which provides general-purpose smooth models for distance-sampling-based abundance estimation, has incorporated a number of SPLINTR developments, and, as discussed below, we are working with its developers to include more in future. Our objective is to have a set of tools for LT/DS datasets where:

- A subset of sightings is available with all relevant covariates known (including school size, unlike SOWER), so that various detection-probability models can be explored and (at least provisionally) fitted with reasonable precision;
- Spatial modelling (including school-size as well as school-density) is done as a separate step, but correctly propagating the uncertainty associated with detection-probability.

We see the first dot-point as a high priority for designing DS studies, based on the SOWER AMW experience. Assuming the propagation of uncertainty can be handled appropriately (which it now can; see below), we continue to prefer two-stage models for spatial DS to all-in-one models. Detection-probability models can be complicated and may require extensive exploratory analysis. Thus, it is highly desirable to avoid complicating

the detection-probability step by dragging spatial models into it. By partly decoupling these two parts of abundance estimation, analysts would be able to experiment freely with different approaches to detection-probability, or with different spatial models, without needing to dive deeply into C code to handle both at once.

The main features of SPLINTR that are influencing ongoing developments in LT/DS, are described in the next four subsections.

5.6.1 Spatial modelling

The basis-penalty random-effect formulation of GAMs for smooth models (Wood *et al*., 2016; embodied in the R package `mgcv`) was first used for line-transect abundance estimation in SPLINTR, and is now well-established in LT/DS modelling, including `dsm`. While other types of smooth models are also used in non-`dsm` line-transect analyses, the basis-penalty approach is attractive because:

- It builds on a widely-used well-tested general-purpose piece of software (`mgcv`) which is theoretically and computationally sound;
- It is flexible – different smoothers can be tried without needing to modify the underlying code;
- It extends straightforwardly to variance propagation in two-stage models (Section 5.6.3).

Aside from the basis-penalty framework, there is also the issue of which type of smoother to use within that framework. `mgcv` offers many options, but not all smoothers are equally suitable for all applications. Soap-film smoothers (Wood *et al*., 2008), devised specifically to deal with irregular physical boundaries encountered in some parts of the Antarctic and then incorporated into `mgcv`, have become fairly popular for cetacean LT/DS work. One reason for their popularity is their ability to easily tame smoothers to ameliorate bizarre extrapolations, an accidental byproduct of their mathematical construction. The theoretical basis of soap-film smoothers is less clear when parts of the boundary are open (or represent a boundary to vessels but not whales, as with some of the 'ice edges' in SOWER). Soap-films may well not be the last word, but for now they remain one of the best available smoothers for LT/DS work.

5.6.2 Clustering

Practical experience in many line-transect surveys, including SOWER, is that sightings can be clustered in time (and thus space) at a scale finer than spatial models can easily handle. If ignored, this tends to lead to under-smoothed spatial models with unruly behaviour near the boundary, thus undoing some of the good work of 'smoother taming' (Section 2.1). SPLINTR addressed clustering in two ways, both of which have general applicability. First, at the level of individual segments, the number of schools seen can now be described by a Tweedie rather than Poisson distribution. Computationally tricky to implement but simple to use, Tweedie distributions were first implemented in `mgcv` in response to our requests (see Wood & Fasiolo, 2017); they are now a standard part of `dsm`. In our experience, Tweedies (in all sorts of situations, not just LT/DS) often seem more effective at handling 'overdispersion' (one manifestation of clustering) than the more familiar Negative Binomial alternative.

Clustering may also manifest itself as autocorrelation between counts in successive snippets of effort; for example, in SPLINTR, we used 15-minute snippets corresponding to about 3 nmi, but field experience in SOWER suggests that clusters can be rather larger than that. For SPLINTR, we implemented a discretised version of Skaug's (2006) Markov-Modulated Poisson Process, which explicitly allows for fine-scale autocorrelation, and we used this in place of a Tweedie distribution because of its insensitivity to snippet size: it bypasses the difficult question in spatial LT of how big segments should be. The amount of extra code was relatively small and implementing the clustering-robust MMPP turned out to be one of the simpler aspects of SPLINTR. There are ongoing discussions about incorporating such an MMPP into future versions of `dsm`.

5.6.3 Variance propagation

SPLINTR had to link two distinct stages of model (SPAMASSS and DOSS). We devised a statistically-valid general-purpose method to propagate uncertainty (variance) from the first stage into additional uncertainty for the

second stage (Appendix B). However, our method did have some drawbacks: it was computationally complex and did not allow the second-stage data to tweak the first-stage estimates, even when a little tweaking would have substantially improved the second-stage fit without markedly worsening the first-stage. This particular problem was mitigated by adding the 'Z-change model' into the first-stage fit (Section 2.2.4), but the Z-change solution is not easily transferable.

Since SPLINTR, we have developed a better algorithm for variance propagation that allows any detection-probability model to be linked to any spatial model fitted with GAM-like formulations (Bravington *et al*., 2021). It is a general-purpose tool that bypasses the many problems of bootstrapping, automates the second-stage tweak, and avoids the computational complexity of Appendix B. The version now in `dsm` also copes with simultaneous smooth modelling of school-size alongside school-density, as described in the next section.

5.6.4 School size

SPLINTR handled variation in true (as opposed to observed) school size by using two conceptually-distinct smoothers: one to describe mean school size, and one to describe school density. There was also a rather complicated nonparametric model to describe variability around the mean. (Other approaches to SOWER AMWs, such as OK, assumed a specific parametric distribution for school size variability around the mean – an assumption which we were uncomfortable with.) Our approach seemed to work but was somewhat cumbersome and not readily transferable. The fundamental problem was that true school size was unavailable for detection-function modelling from IO-mode data, meaning that modelling the detection-function and spatial mean-school-size had to be done simultaneously, making for great complexity in SPAMASSS. But even if true school size had been available (at least for a substantial subset of sightings), we would still have had a three-stage model: detection-function conditional on true school size; spatial school size; and spatial density. While that would have been substantially simpler than SPLINTR, a three-stage model is still awkward.

Since SPLINTR, matters have been simplified by the advent of 'factor-smooth interactions' in `mgcv` (Wood *et al*., 2016) which allow a single family of smooth models to parsimoniously describe spatial variation in several different school-size-classes at once. The size-specific density surfaces can vary from each other (so that mean school size changes spatially), but the extent of that variation is controlled automatically by a data-driven smoothing parameter. Given that framework, there is no need for a separate smoother to describe mean-school-size, nor for a parametric/nonparametric model of school size variability. It turns out that this fits neatly into the variance-propagation algorithm just described (Bravington *et al*., 2021), and the two are integrated into recent versions of `dsm`.

The current density-by-school-size model in `dsm` assumes that school size is measured without much error for all sightings, which is not generally true for LT/DS surveys (including SOWER) where school size varies. However, it is possible in principle to extend our approach to include some sightings with uncertain school size, provided that the survey also collects enough other sightings of known school size and comparable search protocols, for estimating detection probabilities conditioned on true school size (e.g., like SOWER SSX but with all three platforms operating). Based on our SOWER experience, we would strongly recommend designing surveys to ensure such data are available.

ACKNOWLEDGEMENTS

Most of this work was done while MVB was employed at CSIRO (Commonwealth Scientific and Industrial Research Organization, Australia), and SLH was an Honorary Research Fellow at CREEM (Centre for Research into Ecological and Environmental Modelling, University of St Andrews, Scotland). Partial funding was provided by the Australian Antarctic Division, CSIRO and the IWC, although the majority of the work was completed in the authors' personal time. Simon Wood gave extensive assistance with spatial smoothing methods. David Peel helped with coding. Doug Butterworth provided both a critical eye and sufficient encouragement for us to continue to consider the effort worthwhile, and it is to him that we owe the greatest thanks for getting us over the line. The late Paul Ensor, Cruise Leader on numerous cruises and a Researcher on others, was always willing to provide a view on what, in his unique experience, warranted some analytical attention. With such a complex dataset, this was an

invaluable bridge between the practicalities of surveying in a challenging environment and the accuracy and reliability of the data. We thank the extreme patience of others, including our Subcommittee Chair, Debi Palka, the IWC's former Head of Science, Greg Donovan, and our friend and driver behind this Special Issue, the late John Bannister. Our fellow analysts, Justin Cooke, Toshihide Kitakado and Hiroshi Okamura, along with reviewer Hans Skaug, contributed to many insightful discussions.

REFERENCES

Bachl, F.E., Lindgren, F., Borchers, D.L., & Illian, J.B. (2019). inlabru: An R package for Bayesian spatial modelling from ecological survey data. *Methods Ecol. Evol.* 10: 760–766. [Available at: *https://doi.org/doi:10.1111/2041-210X.13168*]

Baum, L.E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat. 37(6): 1554–1563.

Bernoulli, J. (1713). Ars Conjectandi, Opus Posthumum: Vols. Fratrum – Werke 3 (pp. 107–286). Basileae, Impensis Thurnisiorum.

Borchers, D.L., Laake, J.L., Southwell, C., & Paxton, C.G.M. (2006). Accommodating unmodeled heterogeneity in double-observer distance sampling surveys. *Biometrics* 62(2): 372–378. [Available at: *https://doi.org/10.1111/j.1541-0420.2005.00493.x*]

Borchers, D.L., & Langrock, R. (2015). Double-observer line transect surveys with Markov-modulated Poisson process models for animal availability. *Biometrics* 71(4): 1060–1069. [Available at: *https://doi.org/10.1111/biom.12341*]

Borchers, D.L., Marques, T., Gunnlaugsson, T., & Jupp, P. (2010). Estimating distance sampling detection functions when distances are measured with errors. *JABES* 15(3): 346–361.

Borchers, D.L., Zucchini, W., Heide-Jørgensen, M.P., Cañadas, A., & Langrock, R. (2013). Using Hidden Markov Models to deal with availability bias on line transect surveys. *Biometrics* 69(3): 703–713. [Available at: *https://doi.org/10.1111/biom.12049*]

Branch, T.A. (2006). Abundance estimates for Antarctic minke whales from three completed circumpolar sets of surveys, 1978/79 to 2003/04. SC/58/IA/34 presented to the IWC's Scientific Committee, St Kitts, 2006. [Available from the IWC Publications Team]

Branch, T.A. (2007). Possible reasons for the appreciable decrease in abundance estimates for Antarctic minke whales from the IDCR/SOWER surveys between the second and third circumpolar sets of cruises. SC/59/IA/7 presented to the IWC's Scientific Committee, Alaska, 2007. [Available from the IWC Publications Team]

Bravington, M.V., & Hedley, S.L. (2009). Antarctic minke whale abundance estimates from the second and third circumpolar IDCR/SOWER surveys using the SPLINTR model.SC/61/IA/14 presented to the IWC's Scientific Committee, Madeira, 2009. [Available from the IWC Publications Team]

Bravington, M.V., & Hedley, S.L. (2010). Antarctic minke whale abundance from the SPLINTR model: Some 'reference' dataset results and 'preferred' estimates from the second and third circumpolar IDCR/SOWER surveys. SC/62/IA/12rev presented to the IWC's Scientific Committee, Agadir, 2010. [Available from the IWC Publications Team]

Bravington, M.V., & Hedley, S.L. (2012). Abundance estimates of Antarctic minke whales from the IWC IDCR/SOWER surveys, 1986–2002. SC/64/IA/13b presented to the IWC's Scientific Committee, Panama, 2012. [Available from the IWC Publications Team]

Bravington, M.V., Miller, D.L., & Hedley, S.L. (2021). Variance propagation for density surface models. *JABES* 26: 306–323. [Available at: *https://doi.org/10.1007/s13253-021-00438-2*]

Bravington, M.V., Peel, D., & Hedley, S.L. (2006). More abundance estimates for Antarctic minke whales, using different methods. SC/58/IA/15) presented to the IWC's Scientific Committee, St Kitts, 2006. [Available from the IWC Publications Team]

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L., & Thomas, L. (2001). *Introduction to Distance Sampling*. Oxford University Press.

Buckland, S.T., & Turnock, B.J. (1992). A robust line transect method. *Biometrics* 48(3): 901–909. [Available at: *https://doi.org/10.2307/2532356*]

Burt, M.L., Borchers, D.L., & Ensor, P. (2012). Trackline detection probability of Antarctic minke whales: Analysis of the BT model experiments conducted on the IWC-SOWER cruises 2005/6–2007/8. *J. Cetacean Res. Manage.* 12(3): 307–316. [Available at: *https://doi.org/10.47536/jcrm.v12i3.560*]

Chen, S.X. (1998). Measurement errors in line transect surveys. *Biometrics* 54(3): 899–908.

Cooke, J. (2008). A integrated method for analysis of IDCR/SOWER data and TRANSIM simulated data sets. SC/F08/08 presented to the IWC Scientific Committee's Workshop on Minke Whale Abundance Estimates Using IWC/SOWER Data, Seattle, WA, USA, 2008. [Available from the IWC Publications Team]

Diggle, P.J., & Hutchinson, M.F. (1989). On spline smoothing with autocorrelated errors. *Australian J. Stat*. 31(1): 166–182. [Available at: *https://doi.org/10.1111/j.1467-842X.1989.tb00510.x*]

Fewster, R.M., & Pople, A.R. (2008). A comparison of mark-recapture distance-sampling methods applied to aerial surveys of eastern grey kangaroos. *Wildl. Res.* 35(4): 320–330. [Available at: *https://doi.org/10.1071/WR07078*]

Friedlaender, A.S., Goldbogen, J.A., Nowacek, D.P., Read, A.J., Johnston, D., & Gales, N. (2014). Feeding rates and under-ice foraging strategies of the smallest lunge filter feeder, the Antarctic minke whale (*Balaenoptera bonaerensis*). *J. Exp. Biol.* 217(16): 2851–2854. [Available at: *https://doi.org/10.1242/jeb.106682*]

Gill, P.E., Murray, W., & Wright, M.H. (1981). *Practical optimization*. London: Academic Press.

Glennie, R., Buckland, S.T., & Thomas, L. (2015). The effect of animal movement on line transect estimates of abundance. *PLoS ONE* 10. [Available at: *https://doi.org/10.1371/journal.pone.0121333*]

Harrell, F.E. (2001). *Regression M*modelling *Strategies: With applications to linear models, logistic regression, and survival analysis* (2nd ed.). Springer, New York.

Hascoët, L., & Pascual, V. (2013). The Tapenade Automatic Differentiation tool: Principles, Model, and Specification. *ACM Trans. Math. Softw.* 39(3). [Available at: *http://dx.doi.org/10.1145/2450153.2450158*]

Haw, M.D. (1991). An investigation into the differences in minke whale school density estimates from passing mode and closing mode survey in IDCR Antarctic assessment cruises. Rep. Int. Whal. Comm. 41: 313–330.
Hedley, S.L., & Buckland, S.T. (2004). Spatial models for line transect sampling. *JABES* 9: 181–199.
International Whaling Commission (2008). Report of the Scientific Committee (SC60): Report of the planning meeting for the 2007/08 IWC/SOWER cruise and future cruises. *J. Cetacean Res. Manage.* (Suppl.) 11.
International Whaling Commission (2011). Report of the Scientific Committee (SC62): Subcommittee on In-Depth Assessments. *J. Cetacean Res. Manage*. (Suppl.) 12: 185–202.
International Whaling Commission (2012a). Report of the Intersessional IA Workshop on Estimating Abundance of Antarctic Minke Whales. *J. Cetacean Res. Manage.* (Suppl.) 13: 363–368.
International Whaling Commission (2012b). Report of the Scientific Committee (SC63): Subcommittee on In-Depth Assessments. *J. Cetacean Res. Manage.* (Suppl.) 13: 175–191.
International Whaling Commission. (2013). Report of the Scientific Committee (SC63): Subcommittee on In-Depth Assessments. *J. Cetacean Res. Manage.* (Suppl.), 14: 195–203.
Joyce, G.G., Kasamatsu, F., Rowlett, R., & Tsunoda, L. (1988). The IWC/IDCR Southern Hemisphere minke whale assessment cruises: The first ten years. SC/40/O/16 presented to the IWC's Scientific Committee, Auckland-San Diego, 1988
Kimeldorf, G.S., & Wahba, G. (1970). A correspondence Between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat*. 41(2): 495–502. [Available at: *https://doi.org/10.1214/aoms/1177697089*]
Kitakado, T., & Okamura, H. (2009). Estimation of additional variance for Antarctic minke whales based on the abundance estimates from the revised OK method. SC/61/IA/08 presented to the IWC's Scientific Committee, Madeira, 2009. [Available from the IWC Publications Team]
Kristensen, K., Nielsen, A., Berg, C., Skaug, H.J., & Bell, B. (2016). TMB: Automatic Differentiation and Laplace Approximation. *J. Stat. Softw*. 70(5): 1–21. [Available at: *https://doi.org/10.18637/jss.v070.i05*]
Laake, J.L. (1999). Distance sampling with independent observers: Reducing bias from heterogeneity by weakening the conditional independence assumption. In: G.W. Garner, S.C. Amstrup, J.L. Laake, B.F.J. Manly, L.L. McDonald, & D.G. Robertson (eds.), *Marine mammal survey and assessment methods* (pp. 137–148). Balkema.
Laake, J.L., & Borchers, D.L. (2004). Methods for incomplete detection at distance zero. In: S.T. Buckland, D.R. Anderson, K.P. Burnham, J.L. Laake, D.L. Borchers, & L. Thomas (eds.), *Advanced distance sampling* (pp. 108–189). Oxford University Press.
Langrock, R., Borchers, D.L., & Skaug, H.J. (2013). Markov-modulated nonhomogeneous Poisson processes for modelling detections in surveys of marine mammal abundance. *J. Am. Stat. Assoc*. 108. [Available at: *https://doi.org/10.1080/01621459.2013.797356*]
Langrock, R., & Zucchini, W. (2011). Hidden Markov models with arbitrary state dwell-time distributions. *Comput. Stat. Data Anal*. 55(1): 715–724.
Matsuoka, K., Ensor, P., Hakamada, T., Shimada, H., Nishiwaki, S., Kasamatsu, F., & Kato, H. (2003). Overview of minke whale sightings surveys conducted on IWC/IDCR and SOWER Antarctic cruises from 1978/79 to 2000/01. *J. Cetacean Res. Manage* 5(2): 173–201. [Available at: https://doi.org/10.47536/jcrm.v5i2.817]
Miller, D.L., Burt, M.L., Rexstad, E.A., & Thomas, L. (2013). Spatial models for distance sampling data: Recent developments and future directions. *Methods Ecol. Evol.* 4(11): 1001–1010. [Available at: *https://doi.org/10.1111/2041-210X.12105*]
Miller, D.L. (2025) dsm: Density Surface Modelling of Distance Sampling Data. R package v2.3.4. [Available at: *https://github.com/DistanceDevelopment/dsm*]
Mori, M., Butterworth, D.S., Brandao, A., Rademeyer, R.A., Okamura, H., & Matsuda, H. (2003). Observer experience and Antarctic minke whale sighting ability in IWC/IDCR-SOWER surveys. *J. Cetacean Res. Manag*. 5(1): 1–11. [Available at: *https://doi.org/10.47536/jcrm.v5i1.820*]
Okamura, H., Kitakado, T., & Mori, M. (2005). An improved method for line transect sampling in Antarctic minke whale surveys. *J. Cetacean Res. Manage*. 7(2): 97–107. [Available at: *https://doi.org/10.47536/jcrm.v7i2.742*]
Okamura, H., & Kitakado, T. (2010). Abundance estimates of Antarctic minke whales from the historical IDCR/SOWER survey data using the OK method. SC/62/IA/3 presented to the IWC Scientific Committee, Agadir, Morocco, 2010. [Available from the IWC Publications Team]
Okamura, H., Minamikawa, S., Skaug, H.J., & Kishiro, T. (2012). Abundance estimation of long-diving animals using line transect methods. *Biometrics* 68(2): 504–513.
Palka, D.L. (2009). Preliminary results of OK, BHWP, integrated and standard analytical methods when applied to simulated data. SC/A09/AE9 presented to the IWC's Scientific Committee, St Andrews, UK, 2009. [Available from the IWC Publications Team]
Palka, D.L. (2010). Comparison of results from OK, SPLINTR, Integrated and standard analytical methods when applied to simulated data, 2004–2008. SC/62/IA/14 presented to the IWC's Scientific Committee, Agadir, 2010. [Available from the IWC Publications Team]
Palka, D.L., & Smith, D.W. (2004). Update of specifications of data simulating the IDCR/SOWER surveys – 2004. SC/56/IA6 presented to the IWC's Scientific Committee, Sorrento, 2004. [Available from the IWC Publications Team]
Palka, D.L., & Smith, D.W. (2005). Description of 2005 simulations of the IWC/SOWER Southern Hemisphere minke whale abundance surveys. SC/57/IA2 presented to the IWC's Scientific Committee, Ulsan, 2005. [Available from the IWC Publications Team]
Palka, D.L., Butterworth, D., Bravington, M., Cooke, J., Hedley, S., Kitakado, T., & Okamura, H. (in prep.). A historical overview of the estimation of abundance for Antarctic minke whales from the IWC's IDCR/SOWER surveys.
Pastene, L.A., & Goto, M. (2016). Genetic characterization and population genetic structure of the Antarctic minke whale *Balaenoptera bonaerensis* in the Indo-Pacific region of the Southern Ocean. Fisheries Sci. 82(6): 873–886. [Available at: *https://doi.org/10.1007/s12562-016-1025-5*]
Peel, D., & Bravington, M.V. (2005). Consistency checks on some IDCR/SOWER environmental variables. SC/57/IA3 presented to the IWC's Scientific Committee, Ulsan, 2005. [Available from the IWC Publications Team]

Rue, H., & Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. J. Stat. Plan. Inference 137(10): 3177–3192. [Available at: *https://doi.org/10.1016/j.jspi.2006.07.016*]

Skaug, H.J. (2006). Markov modulated Poisson processes for clustered line transect data. Environ. Ecol. Stat. 13: 199–211.

Skaug, H.J., & Fournier, D. (2006). Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Comput. Stat. Data Anal*. 51: 699–709.

Skaug, H.J., Øien, N., Schweder, T., & Bøthun, G. (2004). Abundance of minke whales (*Balaenoptera acutorostrata*) in the Northeastern Atlantic. *Can. J. Fish. Aquat. Sci.* 61: 870–886. [Available at: *https://doi.org/10.1139/f04-020*]

Skaug, H.J., & Schweder, T. (1999). Hazard models for line transect surveys with independent observers. 55: 29–36.

Strindberg, S., & Burt, M.L. (2004). IWC Database-Estimation Software System (DESS) User Manual. University of St Andrews, Scotland.

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: The Condor experience. *Concurr. Comput. Pract. Exp.* 17(2–4): 323–356.

Wahba, G. (1990). Spline models for observational data. Society for Industrial & Applied Mathematics. [Available at: *https://doi.org/10.1137/1.9781611970128*]

Wells, H.G. (1896). *The Island of Doctor Moreau*. Penguin, 2005.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction in R*. Chapman & Hall.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. Royal Stat. Soc. B* 73(1): 3–36. [Available at: *https://doi.org/10.1111/j.1467-9868.2010.00749.x*]

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman & Hall.

Wood, S.N., Bravington, M.V., & Hedley, S.L. (2008). Soap film smoothing. *J. Royal Stat. Soc. B* 70, 931–955. [Available at: *https://doi.org/10.1111/j.1467-9868.2008.00665.x*]

Wood, S.N., & Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics* 73(4): 1071–1081. [Available at: *https://doi.org/10.1111/biom.12666*]

Wood, S.N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *J. Am. Stat. Assoc*. 111(516): 1548–1563. [Available at: *https://doi.org/10.1080/01621459.2016.1180986*]

Zucchini, W., MacDonald, I.L., & Langrock, R. (2021). *Hidden Markov models for time series: An introduction using R* (2nd ed.) CRC Press.

Supplementary Material

A Derivation of sighting probabilities

A.1 Overall structure

Here we address the breakdown of the overall sighting probability— equation (1.3) in the main document— into school-size-error, detection-probability, and spatial-school-size-distribution. For brevity we omit the survey mode m, which should appear on the RHS of every conditioning bar.

$$\begin{aligned}
\mathbb{P}\left[ys_eh|xzo\right] &= \sum\nolimits_s \mathbb{P}\left[ys_eh|\left\{x\right\}zos\right]\mathbb{P}\left[s|xzo\right] \\
\mathbb{P}\left[s|xzo\right] &= \frac{\mathbb{P}\left[o|s\left\{x\right\}z\right]\mathbb{P}\left[s|\left\{z\right\}x\right]}{\mathbb{P}\left[o|xz\right]} \\
&= \frac{\mathbb{P}\left[o|sz\right]\mathbb{P}\left[s|x\right]}{\sum_{s'}\mathbb{P}\left[o|sz\right]\mathbb{P}\left[s|x\right]} \\
\mathbb{P}\left[ys_eh|zos\right] &= \mathbb{P}\left[s_e|yzh\left\{o\right\}s\right]\mathbb{P}\left[hy|ozs\right] \\
\mathbb{P}\left[hy|ozs\right] &= \frac{\mathbb{P}\left[o|hyzs\right]\mathbb{P}\left[hy|sz\right]}{\mathbb{P}\left[o|sz\right]} \\
&= \frac{1\times\mathbb{P}\left[h|ysz\right]\mathbb{P}\left[y|sz\right]}{\mathbb{P}\left[o|sz\right]} \\
&= \frac{\mathbb{P}\left[h|ysz\right]}{\mathbb{P}\left[o|sz\right]}
\end{aligned}$$

Combining everything:

$$\begin{aligned}
\mathbb{P}\left[ys_eh|xzo\right] &= \sum\nolimits_s \mathbb{P}\left[s_e|yzhs\right]\frac{\mathbb{P}\left[h|ysz\right]}{\mathbb{P}\left[o|sz\right]}\frac{\mathbb{P}\left[o|sz\right]\mathbb{P}\left[s|x\right]}{\sum_{s'}\mathbb{P}\left[o|s'z\right]\mathbb{P}\left[s'|x\right]} \\
&= \frac{\sum_s \mathbb{P}\left[s_e|yzhs\right]\mathbb{P}\left[h|ysz\right]\mathbb{P}\left[s|x\right]}{\sum_s \mathbb{P}\left[o|sz\right]\mathbb{P}\left[s|x\right]}
\end{aligned} \tag{A1}$$

For the denominator, note that $\mathbb{P}\left[o|sz\right] = \sum_h \mathbb{P}\left[h|sz\right] = \sum_h \int \mathbb{P}\left[h|ysz\right]dy$ where the integral over y runs from the trackline to the truncation distance.

In Closing mode, $s = s_e$ by assumption, and the two active platforms A and C are not distinguished, so the formula reduces to

$$\mathbb{P}\left[ys|A\cup C\right] = \frac{\mathbb{P}\left[A\cup C|ysz\right]\mathbb{P}\left[s|x\right]}{\sum_s \mathbb{P}\left[A\cup C|sz\right]\mathbb{P}\left[s|x\right]}$$

A.2 TCI for SOWER platforms

The two-platform version of TCI requires three functions to specify $\mathbb{P}\left[h|ysz\right]$ completely. (For brevity, we will just write $\mathbb{P}\left[h|y\right]$ here, but conditioning of all functions on s and z is implied throughout this section.) There are several possible decompositions, of which the most usual seems to be $\mathbb{P}\left[y|A\cup B\right]$, $\mathbb{P}\left[A|By\right]$, and $\mathbb{P}\left[B|Ay\right]$ (e.g. Fewster and Pople, 2008).

SOWER is much more complicated because there are three platforms with C being only one-way-independent, and two survey modes, with B operating only in IO-mode, and A and C operating non-independently in CL-mode. One of the hardest parts of all SPLINTR was to devise a basic set of functions that allowed reconstruction of all the $\mathbb{P}\left[h|ysz\right]$ in a way that is guaranteed consistent with the laws of probability, and that can be expected to have sensible underlying shapes. In the end, the following set of five functions is sufficient

- A standard distance-sampling function $\mathbb{P}\left[y|A\cup B\cup C\right]$
- Four conditional probabilities:$\mathbb{P}\left[A\cup B|A\cup B\cup C,y\right]$; $\mathbb{P}\left[A|By\right]$; $\mathbb{P}\left[B|Ay\right]$; $\mathbb{P}\left[C|aBy\right]$

The first four functions all have clear interpretations and could be estimated directly from data in obvious ways if school size was known exactly. However, the fifth one, $\mathbb{P}\left[C|aBy\right]$, is an imaginary construct that exists only for reasons of

mathematical accountancy. There can never be any *direct* way to estimate it from SOWER data, since in practice if B sees the school then C is censored; it describes something that could only be observed under a hypothetical setup where platform C was fully independent[A1].

Note that the platform-specific g_{0A}, g_{0B}, g_{0C} are the intercepts of the last three functions. By the TCI assumption,

$$g_0\left(A \cup B \cup C | m = \text{IO}\right) = 1 - \left(1 - g_{0A}\right)\left(1 - g_{0B}\right)\left(1 - g_{0C}\right) \tag{A2}$$

The next section shows how these five functions can be combined to calculate detection-probabilities for the various platform combinations. Section A.2.2 explains how the functions are parameterised in SPLINTR, including a way to avoid parametrizing $\mathbb{P}\left[C|aBy\right]$ explicitly.

A.2.1 Derivation of $\mathbb{P}\left[h|y\right]$

Under the TCI assumption, the $y = 0$ intercepts of the four conditional probability curves are:

$$\begin{aligned}
\mathbb{P}\left[A \cup B | A \cup B \cup C, y = 0\right] &= \frac{\mathbb{P}\left[A \cup B \cup C | A \cup B, y = 0\right] \times \mathbb{P}\left[A \cup B | y = 0\right]}{\mathbb{P}\left[A \cup B \cup C | A \cup B, y = 0\right]} = \frac{1 \times \left(1 - (1 - g_{0A})(1 - g_{0B})\right)}{1 - (1 - g_{0A})(1 - g_{0B})(1 - g_{0C})} \\
\mathbb{P}\left[C | aB, y = 0\right] &= g_{0C} \\
\mathbb{P}\left[A | B, y = 0\right] &= g_{0A} \\
\mathbb{P}\left[B | A, y = 0\right] &= g_{0B}
\end{aligned}$$

In IO mode, the four quantities needed to be able to calculate (for the different possible h's) are

$$\mathbb{P}\left[AB|y\right] = \mathbb{P}\left[A|y\right]\mathbb{P}\left[B|Ay\right] \tag{A3}$$

$$\mathbb{P}\left[Ab|y\right] = \mathbb{P}\left[A|y\right]\left(1 - \mathbb{P}\left[B|Ay\right]\right) \tag{A4}$$

$$\mathbb{P}\left[aB|y\right] = \mathbb{P}\left[B|y\right]\left(1 - \mathbb{P}\left[A|By\right]\right) \tag{A5}$$

$$\mathbb{P}\left[Cab|y\right] = \mathbb{P}\left[A \cup B \cup C|y\right] - \mathbb{P}\left[A \cup B|y\right] \tag{A6}$$

In addition, we need $\mathbb{P}\left[A \cup B \cup C\right]$ for the overall sighting probability and the denominator of (A1). We have immediately that

$$\begin{aligned}
\mathbb{P}\left[A \cup B \cup C|y\right] &= \mathbb{P}\left[y|A \cup B \cup C\right] \times \frac{\mathbb{P}\left[A \cup B \cup C\right]}{\mathbb{P}\left[y\right]} \\
&= \frac{\text{haz}_{A \cup B \cup C}\left(y\right)}{\int_0^{w/2} \text{haz}_{A \cup B \cup C}\left(y'\right) dy'} \times \frac{\mathbb{P}\left[A \cup B \cup C\right]}{1}
\end{aligned}$$

where $\text{haz}\left(y\right)$ could actually be any standard distance-sampling function with intercept 1; for convenience, we chose a hazard-rate version, defined below in section A.2.2. The integral simply normalizes $\mathbb{P}\left[y|A \cup B \cup C\right]$ since we know that y must be somewhere within the strip given that the sighting was made at all.

At $y = 0$, we know that $\mathbb{P}\left[A \cup B \cup C|y = 0\right] = g_0\left(A \cup B \cup C\right)$ and $\text{haz}_{A \cup B \cup C}\left(0\right) = 1$, so that

$$\begin{aligned}
g_0\left(A \cup B \cup C\right) &= \frac{\mathbb{P}\left[A \cup B \cup C\right]}{\int_0^{w/2} \text{haz}_{A \cup B \cup C}\left(y'\right) dy'} \\
&\implies \mathbb{P}\left[A \cup B \cup C|y\right] = g_0\left(A \cup B \cup C\right) \text{haz}_{A \cup B \cup C}\left(y\right)
\end{aligned} \tag{A7}$$

$$\begin{aligned}
\mathbb{P}\left[A \cup B|y\right] &= \mathbb{P}\left[A \cup B|A \cup B \cup C, y\right]\mathbb{P}\left[A \cup B \cup C|y\right] + \mathbb{P}\left[A \cup B|abc, y\right]\mathbb{P}\left[abc|y\right] \\
&= \mathbb{P}\left[A \cup B|A \cup B \cup C, y\right]\mathbb{P}\left[A \cup B \cup C|y\right] + 0 \\
&= g_0\left(A \cup B \cup C\right)\text{haz}_{A \cup B \cup C}\left(y\right)\mathbb{P}\left[A \cup B|A \cup B \cup C, y\right]
\end{aligned} \tag{A8}$$

[A1] A lower limit could be estimated from cases where C sees the school before B, but not a particularly useful one.

which sorts out equation (A6). To sort out the other equations, we require $\mathbb{P}[A|y]$ and $\mathbb{P}[B|y]$.

$$\begin{aligned}\mathbb{P}[A\cup B|y] &= \mathbb{P}[A|y] + \mathbb{P}[aB|y]\\ &= 1\times\mathbb{P}[A|y] + (1-\mathbb{P}[A|By])\times\mathbb{P}[B|y]\\ \mathbb{P}[A\cup B|y] &= \mathbb{P}[B|y] + \mathbb{P}[Ab|y]\\ &= 1\times\mathbb{P}[B|y] + (1-\mathbb{P}[B|Ay])\times\mathbb{P}[A|y]\\ &= (1-\mathbb{P}[B|Ay])\times\mathbb{P}[A|y] + 1\times\mathbb{P}[B|y]\end{aligned} \tag{A9}$$

giving two simultaneous linear equations for $\mathbb{P}[A|y]$ and $\mathbb{P}[B|y]$ in terms of the known quantities $\mathbb{P}[A\cup B|y]$, $\mathbb{P}[A|By]$ and $\mathbb{P}[B|Ay]$. The solution is

$$\begin{aligned}\begin{bmatrix}\mathbb{P}[A|y]\\ \mathbb{P}[B|y]\end{bmatrix} &= \frac{1}{1-(1-\mathbb{P}[B|Ay])(1-\mathbb{P}[A|By])}\begin{bmatrix}1-(1-\mathbb{P}[A|By])\\ 1-(1-\mathbb{P}[B|Ay])\end{bmatrix}\mathbb{P}[A\cup B|y]\\ &= \frac{\mathbb{P}[A\cup B|y]}{\mathbb{P}[B|Ay]+\mathbb{P}[A|By]-\mathbb{P}[B|Ay]\,\mathbb{P}[A|By]}\begin{bmatrix}\mathbb{P}[A|By]\\ \mathbb{P}[B|Ay]\end{bmatrix}\end{aligned} \tag{A10}$$

For CL-mode, the fact-of-observation o in the denominator of (A1) actually means $\mathbb{P}[A\cup C]$. We therefore need $\mathbb{P}[A\cup C]$ and $\mathbb{P}[A\cup C|y]$. For the latter, the same reasoning as in (A8) gives this:

$$\begin{aligned}&\mathbb{P}[A\cup C|y]\\ &= \mathbb{P}[A\cup C|aBy]\,\mathbb{P}[aB|y] + \mathbb{P}[A\cup C|aby]\,\mathbb{P}[ab|y] + \mathbb{P}[A\cup C|Ay]\,\mathbb{P}[A|y]\\ &= \mathbb{P}[C|aBy]\,\mathbb{P}[aB|y] + \mathbb{P}[A\cup B\cup C|aby]\,\mathbb{P}[ab|y] + \mathbb{P}[A|y]\\ &= \mathbb{P}[C|aBy]\,\mathbb{P}[aB|y] + \frac{\mathbb{P}[ab|A\cup B\cup Cy]\,\mathbb{P}[A\cup B\cup C|y]}{\mathbb{P}[ab|y]}\mathbb{P}[ab|y] + \mathbb{P}[A|y]\\ &= \mathbb{P}[C|aBy]\,\mathbb{P}[aB|y] + \mathbb{P}[ab|A\cup B\cup Cy]\,\mathbb{P}[A\cup B\cup C|y] + \mathbb{P}[A|y]\\ &= \mathbb{P}[C|aBy]\,\mathbb{P}[B|y]\,(1-\mathbb{P}[A|By]) + (1-\mathbb{P}[A\cup B|A\cup B\cup Cy])\,\mathbb{P}[A\cup B\cup C|y] + \mathbb{P}[A|y]\end{aligned} \tag{A11}$$

Finally, $\mathbb{P}[A\cup C]$ can be calculated by numerical integration over y.

This formulation in terms of $\mathbb{P}[C|aBy]$ can lead to plateaus or, worse, to bumps in $\mathbb{P}[A\cup C|y]$, but they do not seem very severe. In essence, using three parameters to describe $\mathbb{P}[C|aBy]$, which manifests itself only through subtle differences between two curves, is excessively ambitious given the available data. Consequently, some overfitting tends to happen, as the model 'tries' to hit accidental features of the distance function for CL-mode. However, the formulation limits the damage that can be done; the aB events are fairly rare to begin with, so even if all of them would have been seen by C, they would not contribute much to $\mathbb{P}[A\cup C|y]$. This is clear from equation (A11), where the two rightmost terms are already fixed by the other probability models, and where $\mathbb{P}[A|y]$ will tend to dominate $\mathbb{P}[aB|y]$ by the nature of the sighting process.

There might appear to be more obvious ways to get to $\mathbb{P}[A\cup C|y]$ instead of via the abstruse $\mathbb{P}[C|aBy]$. However, more direct parameterisations seem prone to generating contradictions such as $\mathbb{P}[A\cup C|A\cup B\cup C, y] < \mathbb{P}[A|A\cup B\cup C, y]$. Using $\mathbb{P}[C|aBy]$ ensures consistency with the laws of probability.

A.2.2 Parameterisation of detection probability ingredients

$\mathbb{P}[y|A\cup B\cup C]$ is a standard detection function. In SPLINTR it is parameterised as a hazard-rate model[A2]. That is:

$$\mathbb{P}[y|A\cup B\cup C] \propto \mathrm{haz}\,(y;\sigma_{A\cup B\cup C}, b_{A\cup B\cup C}) \tag{A12}$$

where $\mathrm{haz}\,(y;\sigma,b) \triangleq 1-\exp\left(-(y/\sigma)^{-b}\right)$. For brevity we normally write $\mathrm{haz}_{A\cup B\cup C}\,(y)$, omitting the parameters b and σ, which can themselves depend on s and z

As to the four conditional-probability ingredients $\mathbb{P}[A|By]$ etc., it is not obvious *a priori* whether they should increase or decrease with y, nor what their limits should be at the trackline and the truncation distance. Empirical inspection of

[A2]Not to be confused with hazard-probability models for cue-based detection probability.

the data (subject to the complication of uncertain school size) gives moderately clear suggestions for which way each one goes, except for $\mathbb{P}[C|aBy]$ which can only be inferred indirectly from the difference between the shapes of the distance functions for IO and CL mode. They are all represented in TCI SPLINTR as scaled and shifted hazard-rate functions; they start flat near $y = 0$ and then rise (or fall) to another asymptote somewhere between 0 and 1. It therefore takes three further parameters apart from the g_0's to describe each of the four curves: the asymptote as $y \to \infty$, the inflection point when the curve changes from leaving one asymptote to approaching another, and the abruptness of that change. Specifically, each curve has a shape parameter b^{CP}, a scale parameter s^{CP}, and a far-distance asymptote g_∞^{CP} (a descriptive but somewhat misleading name). In total, 17 fundamental parameters are required to describe all five distance-sampling functions. Each of these fundamental parameters can potentially vary with s and/or z, and the SPLINTR code allows various forms of dependence to be specified (see next section).

An alternative parameterisation which avoids extra parameters for the hard-to-estimate ingredient $\mathbb{P}[C|aBy]$, is to instead make the assumption that

$$\mathrm{logit}\mathbb{P}[C|aBy] - \mathrm{logit}\mathbb{P}[C|aby] = \mathrm{logit}\mathbb{P}[A|By] - \mathrm{logit}\mathbb{P}[A|by]\,.$$

The implication is that the event 'B sees the school' has the same effect (on the logit scale) on $\mathbb{P}[C|ay]$, for which we have little data, as it does on $\mathbb{P}[A|y]$, for which we have good data. The assumption is hard to test, but in practice this more parsimonious version (with 14 rather than 17 fundamental parameters) did not seem worsen the fits appreciably, so we adopted it for our preferred estimates.

A.2.3 Bivariate-increasing independence

TCI SPLINTR allows considerable flexibility in how the 14-or-17 fundamental detection probability parameters are driven by the covariates s (school size) and z ('sighting conditions', perhaps stratified by vessel): no dependence, dependence on either one, or dependence on both. Sometimes common sense suggests that certain parameters should increase with one or both covariates; for example, the scale parameters of a detection function should probably increase both with increasing true school size and with increasing Sightability. Specifying a strictly additive dependence via an '$s + z$' model might achieve this, but may not be flexible enough; on the other hand, because of limited data, a full-interaction model via '$s * z$' may give nonsensical estimates where, for example, scale decreases with school size at some Sightability values. Instead, we developed a way to formulate a bivariate discrete-factor model so that the response must obey the common-sense bivariate-increasing property, but is otherwise free to take any set of values. The number of parameters is the same as for a full-interaction model, but many of the parameters have limited ability to affect the response variable; one slight downside is that there is very little information to estimate some of these limited-unimportant parameters, which can lead to minor numerical difficulties in computing Hessians.

B Variance calculations

After fitting the SPAMASSS and DOSS models to one entire CP series, it is typically necessary to estimate the variances and covariances of abundance estimates for, say, each Management Area. This can be done using familiar statistical and mathematical tools and approximations: Taylor expansions, the 'delta method', the Law of Total Variance (LOTV), Laplace approximation, the Implicit Function Theorem for derivatives, and Automatic Differentiation (AD). However, the application of those tools to SPLINTR is complex, because of the multi-stage (SPAMASSS and DOSS) and multi-year models (which share some but not all parameters across years), and because of the nested optimizations required to properly fit penalized-likelihood smoothers via REML. The presentation here concentrates on the mathematical steps, taking for simplicity as aggregated a viewpoint as possible; the computer code ends up looking rather different, for reasons of efficiency.

Several different abundance estimates are made, corresponding to different regions and/or years, but all estimated from the same SPLINTR modelling process. Each abundance element is constructed from values (local densities of animals) at a number of spatial grid cells, which are combined into an area-weighted sum. We can specify the entire problem as

$$A(\beta, \gamma) = W^\top \cdot \left(g\left(X^\top \beta\right) \times h\left(Z^\top \gamma\right)\right) \tag{B1}$$

where

- A is the *vector* of all desired abundances from that pair of SPAMASSS/DOSS models, including all years and subregions. Typically, A will have a small number of elements, say one per MA.
- β is a vector of *all* the smoother coefficients (i.e. from all years concatenated) for the DOSS model (or models; see below), and γ similarly for the spatial-school-size model within SPAMASSS. Typically, β and γ might each have ~ 100 elements per year.
- $g()$ and $h()$ are 'vector scalar-wise' functions, i.e. that map individual elements $\mathbb{R} \to \mathbb{R}$. g gives a number-density, of schools per unit area; h returns a mean school size. Typically $g = \exp$, whereas h is more complicated.
- X and Z are design matrices. Each row $X_{i\cdot}$ corresponds to one grid cell; element X_{ik} applies to the k^{th} smoother coefficient in that cell.
- W is a weightings vector, with W_{ij} showing the area of the j^{th} grid cell in abundance element i.

Also, let θ be the set of all sighting-and-size-related (SPAMASSS) parameters, including γ but also other purely sighting-related parameters here denoted by ϕ (e.g. to do with g_0), so that $\theta \triangleq (\phi, \gamma)$. While ϕ itself is not connected to A, the *estimate* of ϕ does affect the *estimates* of β and γ, and thus of A.

At a technical level, the DOSS 'model' is actually implemented as separate models for each year, although they do share a common smoothing parameter which is estimated collectively. However, there is no mathematical reason why all years (within a CP series) could not be fitted simultaneously; the decision to implement fitting via separate models was a computational one, to avoid the y^3 time cost (y=number of years here) that would apply from concatenating all the smoother coefficients into one huge vector (in the absence of sparsity tools for AD and optimisation). Similar considerations apply to the spatial-school-size components of the SPAMASSS model, which are fitted as separate year-by-year models—though all share the same smoothing parameter λ^{SPAM}, and the same sighting parameters.

Now let s be all the per-**s**ighting data (i.e. the input data for SPAMASSS), and r be all the sighting-**r**ate data (i.e. the input data for DOSS). The MLE from SPAMASSS is $\hat{\theta}(s)$, which we will generally write just as $\hat{\theta}$; we will also need the DOSS MLE $\hat{\beta}(\theta, r)$ or just $\hat{\beta}(\theta)$, which depends on θ and may or may not be evaluated at $\theta = \hat{\theta}$, depending on context. Our aim is to compute a pseudo-Bayesian quantity $\mathbb{V}[A|s,r]$, i.e. variance conditional on the observed data. Following widespread modern practice in asymptotic statistics, we will approximate that using MLE-related quantities and vague unstated assumptions about uniform priors etc. The first step is to apply LOTV and some approximations:

$$\begin{aligned}
\mathbb{V}[A|s,r] &= \mathbb{E}_{\theta|s}[\mathbb{V}[A|\theta,r]] + \mathbb{V}_{\theta|s}[\mathbb{E}[A|\theta,r]] \\
&\approx \mathbb{V}\left[A|\hat{\theta},r\right] + \mathbb{V}_{\theta|s}\left[A\left(\hat{\beta}(\theta;r),\theta\right)\right] \\
\mathbb{V}\left[A|\hat{\theta},r\right] &\approx \left.\frac{dA}{d\beta}\right|^{\top}_{\hat{\beta}(\theta)} \cdot \mathbb{V}\left[\beta|\hat{\theta},r\right] \cdot \left.\frac{dA}{d\beta}\right|_{\hat{\beta}(\theta),\hat{\theta}}
\end{aligned}$$

The left-hand term $\mathbb{E}_{\theta|s}[\cdot]$ on the top line is approximated by 'zeroth-order expansion': i.e. for general $f()$, Y, and z, $\mathbb{E}[f(Y)|z] \approx f(\mathbb{E}[Y|z]) \approx f\left(\hat{Y}(z)\right)$. That leads to the expression in the third line. For the remaining term on the right hand end of the second line, $\mathbb{V}_{\theta|s}[\cdot]$, the abundance $A\left(\hat{\beta}(\theta),\theta\right)$ has become a function of θ alone, which we can write as a first-order Taylor-expansion around $\hat{\theta}$:

$$\begin{aligned}
A\left(\hat{\beta}(\theta),\theta\right) &\approx A\left(\hat{\beta}\left(\hat{\theta}\right),\hat{\theta}\right) + \left[\left.\frac{dA}{d\beta}\right|_{\hat{\beta}(\hat{\theta}),\hat{\theta}} \cdot \left.\frac{d\hat{\beta}(\theta,r)}{d\theta}\right|_{\hat{\theta}} + \left(0,\ \left.\frac{dA}{d\gamma}\right|_{\hat{\beta}(\hat{\theta}),\hat{\theta}}\right)\right]^{\top}\left(\theta - \hat{\theta}\right) \qquad \text{(B2)} \\
&= \hat{A}(r,s) + Q^{\top}\left(\theta - \hat{\theta}\right) \\
Q &\triangleq \left.\frac{dA}{d\beta}\right|_{\hat{\beta}(\hat{\theta}),\hat{\theta}} \cdot \left.\frac{d\hat{\beta}(\theta,r)}{d\theta}\right|_{\hat{\theta}} + \left(0,\ \left.\frac{dA}{d\gamma}\right|_{\hat{\beta}(\hat{\theta}),\hat{\theta}}\right)
\end{aligned}$$

where $\hat{A}(r,s) \triangleq A\left(\hat{\beta}\left(\hat{\theta}\right),\hat{\gamma}\right)$ does not include any free parameters since it is evaluated at the MLEs. Because of the linear form of the approximation (B2), once Q is available we have the straightforward expression

$$\mathbb{V}_{\theta|s}\left[A\left(\hat{\beta}(\theta;r),\theta\right)\right] \approx Q^{\top}\mathbb{V}[\theta|s]\,Q \qquad \text{(B3)}$$

As to Q, first note that the reason for the zero in $(0,\, dA/d\gamma)$ is that the only direct dependence of A on θ is via γ,which comes at the end of θ. Next, the terms $dA/d\beta$ and $dA/d\gamma$ can readily be obtained by AD of (B1). The remaining term in Q is $d\hat{\beta}\left(\theta, r\right)/d\theta\Big|_{\hat{\theta}}$, i.e. how much a small change in θ near $\hat{\theta}$ would affect the point estimate of β given θ. Although there is no closed-form expression for $\hat{\beta}$ in terms of θ, the derivative can be obtained exactly (not approximately) via the Implicit Function Theorem:

$$\begin{aligned}
&\hat{\beta}\left(\theta\right) \text{ s.t. } \left.\frac{d\Lambda^{\mathrm{DOSS}}}{d\beta}\right|_{\hat{\beta}(\theta),\theta} = 0 \,\forall\, \theta \\
&\implies \frac{d}{d\theta}\left(\left.\frac{d\Lambda^{\mathrm{DOSS}}}{d\beta}\right|_{\hat{\beta}(\theta),\theta}\right) = 0 \,\forall\, \theta \\
&\implies \left.\frac{d^2\Lambda^{\mathrm{DOSS}}}{d\theta d\beta}\right|_{\hat{\beta}(\hat{\theta}),\hat{\theta}} + \left[\left.\frac{d^2\Lambda^{\mathrm{DOSS}}}{d\beta^2}\right|_{\hat{\beta}(\hat{\theta}),\hat{\theta}}\right] \times \left.\frac{d\hat{\beta}}{d\theta}\right|_{\hat{\theta}} = 0 \\
&\implies \left.\frac{d\hat{\beta}}{d\theta}\right|_{\hat{\theta}} = -\left[\left.\frac{d^2\Lambda^{\mathrm{DOSS}}}{d\beta^2}\right|_{\hat{\beta}(\hat{\theta}),\hat{\theta}}\right]^{-1}\left[\left.\frac{d^2\Lambda^{\mathrm{DOSS}}}{d\theta d\beta}\right|_{\hat{\beta}(\hat{\theta}),\hat{\theta}}\right]
\end{aligned} \tag{B4}$$

The parameter θ enters the DOSS log-likelihood Λ^{DOSS} via the offset term in the linear predictor, whereby

$$\log \mathbb{E}\left[R_i|\beta\right] = \mathrm{off}_i\left(\theta\right) + X_i^{\top}\beta$$

and off_i captures the effects of weather, school size frequency distribution, etc. in grid cell i. Hence the term $d^2\Lambda^{\mathrm{DOSS}}/d\theta d\beta$ is readily computed by AD. The other term in (B4), $d^2\Lambda^{\mathrm{DOSS}}/d\beta^2$, is the familiar Hessian from fitting the DOSS model(s) evaluated at the estimated smoothing parameter $\hat{\lambda}^{\mathrm{DOSS}}$; here we follow the fairly widespread practice of ignoring the impact of smoothing parameter uncertainty, which seems typically to be small in random-effects models with plenty of data.

The only remaining term to be computed is $\mathbb{V}\left[\theta|s\right]$. This looks superficially like it should just be 'an inverse Hessian', but the problem is that the SPAMASSS 'likelihood' for θ is actually a Laplace approximation, whose computation for any θ entails an 'inner' optimization over γ after fixing ϕ and the smoothing parameter λ^{SPAM}. Structurally, the (>100-dimension) parameter ϕ together with λ^{SPAM} plays the same hyperparameter role in SPAMASSS that λ^{DOSS} on its own does for DOSS; but in SPAMASSS it is clearly not acceptable to ignore the uncertainty about hyperparameters (e.g. those relating to g_0). Here again we can deploy LOTV:

$$\begin{aligned}
\mathbb{V}\left[\theta|s\right] &= \mathbb{E}_{\phi|s}\left[\mathbb{V}\left[\left(\phi,\gamma\right)|\phi,s\right]\right] + \mathbb{V}_{\phi|s}\left[\mathbb{E}\left[\left(\phi,\gamma\right)|\phi,s\right]\right] \\
\mathbb{E}_{\phi|s}\left[\mathbb{V}\left[\left(\phi,\gamma\right)|\phi,s\right]\right] &= \mathbb{E}_{\phi|s}\left[\begin{pmatrix}0 & 0\\ 0 & \mathbb{V}\left[\gamma|\phi,s\right]\end{pmatrix}\middle|\,\phi,s\right] \approx \begin{pmatrix}0 & 0\\ 0 & \mathbb{V}\left[\gamma|\hat{\phi},s\right]\end{pmatrix}
\end{aligned}$$

where the zeros arise since by definition $\mathbb{V}\left[\phi|\phi\right] \equiv 0$ etc. The term $\mathbb{V}\left[\gamma|\hat{\phi},s\right]$ is the negative inverse Hessian from the inner penalized log-likelihood used in fitting γ once ϕ is specified. The next term to grapple with is $\mathbb{V}_{\phi|s}\left[\mathbb{E}\left[\left(\phi,\gamma\right)|\phi,s\right]\right]$. Again, we start by approximating $\mathbb{E}\left[\gamma|\phi,s\right] \approx \hat{\gamma}\left(\phi,s\right)$ and then using a first-order Taylor expansion:

$$\begin{aligned}
&\mathbb{E}\left[\gamma|\phi,s\right] \approx \hat{\gamma}\left(\phi,s\right) \approx \hat{\gamma}\left(\hat{\phi},s\right) + \left.\frac{d\hat{\gamma}}{d\phi}\right|_{\hat{\phi}} \times \left(\phi - \hat{\phi}\right) \\
&\implies \mathbb{V}_{\phi|s}\left[\mathbb{E}\left[\left(\phi,\gamma\right)|\phi,s\right]\right] \approx \left.\frac{d\hat{\gamma}}{d\phi}\right|_{\hat{\phi}}^{\top} \cdot \mathbb{V}\left[\phi|s\right] \cdot \left.\frac{d\hat{\gamma}}{d\phi}\right|_{\hat{\phi}}
\end{aligned}$$

The term $d\hat{\gamma}/d\phi$ indicates how much the point estimate of γ will change in response to local variations of ϕ around its point estimate. Once again, the Implicit Function Theorem comes to our aid. We know that $\hat{\gamma}$ is the maximizer of the

SPAMASSS penalized log-likelihood; so, by similar arguments to (B4), we have

$$\hat{\gamma}(\phi) \text{ s.t. } \left.\frac{d\Lambda^{\text{SPAM}}}{d\gamma}\right|_{\hat{\gamma}(\phi),\phi} = 0 \,\forall\, \phi$$

$$\implies \left|\frac{d\hat{\gamma}}{d\phi}\right|_{\hat{\phi}} = -\left[\left.\frac{d^2\Lambda^{\text{SPAM}}}{d\gamma^2}\right|_{\hat{\gamma}(\hat{\phi}),\hat{\phi}}\right]^{-1}\left[\left.\frac{d^2\Lambda^{\text{SPAM}}}{d\phi d\gamma}\right|_{\hat{\gamma}(\hat{\phi}),\hat{\phi}}\right]$$

The absolutely final term is $\mathbb{V}[\phi|s]$. This can be approximated by the negative inverse Hessian of the Laplace approximation used to find $\hat{\phi}$, for which

$$\hat{\phi} \triangleq \text{argmax}\tilde{\Lambda}^{\text{SPAM}}(\phi|s)$$

$$\tilde{\Lambda}^{\text{SPAM}}(\phi|s) \triangleq \max_{\gamma} \tilde{\tilde{\Lambda}}^{\text{SPAM}}(\phi,\gamma)$$

$$= \tilde{\tilde{\Lambda}}^{\text{SPAM}}(\phi,\hat{\gamma}(\phi)) \text{ where } \hat{\gamma}(\phi) \triangleq \arg\max_{\gamma} \tilde{\tilde{\Lambda}}^{\text{SPAM}}(\phi,\gamma)$$

$$\tilde{\tilde{\Lambda}}^{\text{SPAM}}(\phi,\gamma) \triangleq \Lambda^{\text{SPAM}}(\phi,\gamma|s) - \frac{1}{2}\log\left\|\left.\frac{d^2\Lambda^{\text{SPAM}}}{d\gamma^2}\right|_{\gamma,\phi}\right\| \tag{B5}$$

and $\Lambda^{\text{SPAM}}(\phi,\gamma)$ is the penalized log-likelihood for SPAMASSS. The two-argument function $\tilde{\tilde{\Lambda}}^{\text{SPAM}}(\phi,\gamma)$ and its derivatives are fairly straightfoward to compute; the term $d^2\Lambda^{\text{SPAM}}/d\gamma^2$ can be obtained by AD, and its determinant can be evaluated by Cholesky decomposition which, as a simple non-iterative algorithm, is itself directly suitable for further layers of AD. The Laplace Approximation $\tilde{\Lambda}^{\text{SPAM}}$ is a one-argument version, and its Hessian is not trivial because of the dependence of $\hat{\gamma}(\phi)$ on ϕ, but a little further work with the Implicit Function Theorem, along similar lines to before, does lead to a computable solution. Overall, four layers of AD (two forward, two reverse) are required, plus one inner optimisation per year to find $\hat{\gamma}(\phi)$.

Although the entire formulation above is in one sense rather simple mathematically— nothing more than a set of derivatives and matrix multiplications— there is a considerable challenge in implementing all the steps via AD in a computationally-efficient way. In particular, it is important for reasons of speed to avoid actually computing entire matrices such as $d^2\Lambda^{\text{DOSS}}/d\theta d\beta$ wherever possible. With care, the above expressions can be written so that the only required quantities are *directional* forward-mode derivatives (of the form $(d/dt)f(x+ty)$ for scalar $f()$ and t, vector x, and known vector direction y), as well as reverse-mode derivatives (of the form $(d/dx)f(x)$).

Since A will typically have few elements, and the first derivatives of each element with respect to β and γ are weighted sums of restricted subsets of β and γ, the number of computations— and the computational time— can be substantially restricted by diligent attention to the order of calculation; the algebra above, which basically reduces to a series of matrix multiplications, disguises the fact that certain terms can be far more efficiently evaluated by AD working right-to-left rather than left-to-right, for example. The details are omitted in the interests of brevity and sanity, but implementation of the entire variance algorithm is certainly one way to learn a great deal about practical use of Automatic Differentiation.

When coded as efficiently as possible, the time required for variance estimates was similar to the time required to actually fit the SPAMASSS and DOSS models. The only problem encountered with the above process (apart from the colossal challenge of getting the whole thing to work in the first place) was that the SPAMASSS Hessian $d^2\tilde{\Lambda}/d\theta^2$ would sometimes fail to be negative-definite. In such cases it is important to find the wrongly-signed eigenvalues and look at their corresponding eigenvectors, to find out which combinations of parameters are not being estimated reliably. Invariably across different model runs, the (few) problem eigenvectors corresponded to obscure parameters: for example, 'inner' parameters for the internal structure of bivariate-increasing dependencies (section A.2.3) whose scope to affect observable probabilities is intrinsically limited by the 'edge' parameters; or parameters describing the obscure-but-mathematically-necessary $\mathbb{P}[C|aBy]$ (section A.2). There were no instances of inestimability that involved more important/obvious parameters. Since the inestimable combinations in this application turn out not to matter practically, a defensible and pragmatic solution is simply to treat those inestimable combinations as fixed (i.e. perfectly known) at their point estimates, essentially setting the wrongly-signed eigenvalues to the correctly-signed infinity. The resulting $\mathbb{V}[\phi|s]$ is technically a pseudoinverse rather than true inverse of the Hessian.

References

Fewster, RM and AR Pople (2008). "A comparison of mark-recapture distance-sampling methods applied to aerial surveys of eastern grey kangaroos". In: *Wildlife Research* 35.4, pp. 320–330. DOI: 10.1071/WR07078. URL: https://doi.org/10.1071/WR07078.