# Comments on design-based and model-based abundance estimates for the RMP and other contexts

Sharon Hedley*and Mark Bravington†

# Design-based abundance estimation

The Requirements and Guidelines for Conducting Surveys and Analysing Data within the Revised Management Management Scheme IWC (2012, hereafter "the Guidelines") were written to help ensure that with adequate survey planning and design, a particular sighting survey could, if desired, be analysed by drawing on statistical design-based principles. Broadly speaking then, such an analysis would rely on there being an estimable probability that any point in the stratum of interest could be included within the strip surveyed. For any extrapolation of density beyond the surveyed strip, then it is usual (though not necessary for estimation) that the "chance" that a point is surveyed is the same throughout the stratum of interest. Specifically, The Guidelines say this: "Estimates of abundance obtained from a survey design that does not meet this criterion [uniform, close to uniform, or estimable coverage probability], will not be accepted for use in the *CLA* unless they have received prior approval from the Scientific Committee".

This paper considers various types of surveys and analyses, and if and how they meet The Guidelines. It examines whether there are circumstances under which estimates from surveys that do not meet the design-based definition (either in a strict sense, or otherwise) may be considered for use in RMP *Implementation Simulation Trials* and in actual application of the *CLA*.

In particular, model-based methods for estimation of abundance (e.g. Beekmans et al., 2010; Bravington and Hedley, 2012; Miller et al., 2013, and references therein) are now available, and though usually requiring some statistical expertise to implement, could be considered for use both for surveys which did and — more importantly — did not, meet the design-based criteria. In this report we summarise some of our own experience in use of such spatial models, and attempt to provide useful suggestions, caveats and criteria for evaluating and conducting future analyses using these methods.

## 1 The importance of estimating coverage probability

What is meant by "estimable coverage probability"? A stratum with transects that have been set with no randomization scheme applied do have *estimable* coverage probability (it is 1 along the surveyed strip and 0

*The Schoolhouse, Denhead, St Andrews, Fife KY16 8PA
†CSIRO, Hobart 7000, Australia

elsewhere), but sampling theory precludes estimation in areas of zero coverage probability. In practice, analyses usually assume uniform, or nearly uniform, coverage probability. Depending on the transect design (eg. equally-spaced parallel lines vs. zig-zags) and the complexity of the survey region, it may be possible to appeal to theoretical results regarding coverage probability, provided that as a minimum, a random location has been selected for the placement of one of the transects (even if the location of this single transect then pre-determines the locations of the remaining transect grid, as in *systematic* survey designs). However, when parallel lines are not the design-of-choice, the actual coverage probability throughout a region is less obvious, and it makes sense to check proposed designs via simulation. Software *Distance* Thomas et al. (2009), for example, provides a Windows interface for specifying typical transect designs and estimates and displays coverage probability in a linked GIS.

Despite both readily available software and the relatively simple simulations required to examine the equal coverage probability assumption, the literature reveals few instances where evidence of testing or checking this assumption has been included in the reporting of an analysis; it is far more common to include a sentence or so in the text regarding the randomization scheme alone. An example where coverage probability was explicitly illustrated is in Thomas et al. (2007), who report on design issues involving a cetacean survey within a fairly complex survey region off British Columbia. Coverage probability was estimated from many replications of a design algorithm (which specified waypoint spacing, and equally-spaced zig-zag or equally-spaced parallel lines). Coverage probability for the parallel lines design was even, and for the zig-zag design was similar, but was higher at the northern and southern boundaries of the region. Not unreasonably, Thomas et al. (2007) considered that as the affected regions were relatively small in area, then this neither significantly compromised the design nor the subsequent analysis, which proceeded assuming equal coverage probability. Perhaps the pertinent point is that they carefully looked at the design, and made sensible analytical decisions based on their examination.

In terms of assessing whether the results of an analysis are acceptable for use in *Implementation Simulation Trials* or the *CLA*, then although the Guidelines include reference to cruisetrack design and coverage probability, it is perhaps more generally the case that it is possible to scrutinise the analysis more than the design. This depends on what information on the design is presented to the Scientific Committee; three possibilities are as follows:

1. Coverage probability confirmed to be (near) uniform, and a set of transects generated from an appropriate algorithm. Note that the transect generation may be statistically randomized (from which the analyst can draw on well-established statistical survey principles for estimation, including estimation of variance), or more commonly, generated systematically, where only the initial start point of one transect has been determined in a randomized manner. See section 2.

2. No estimated coverage probability and stated (in the text accompanying the analysis) that for example, transects "were designed using the program *Distance* Thomas et al., 2009 following the principles outlined in [The Guidelines, IWC, 2012]" but diagrams depict transects starting in corners, close to a port, etc.

3. No estimated coverage probability and the design is described in the accompanying text as above, but the diagrams depict transects drawn in an apparently *ad hoc* way, but with on the face of it, "reasonable" coverage of the survey area.

4. Furthermore, there may be additional pieces of information relevant to the realized (as opposed to the planned) tracklines, eg. failure to survey one half of a stratum owing to poor weather.

Designs such as in items 2 and 3 above contravene the spirit (if not the letter) of the Guidelines. Nothing other

than making a subjective judgment that "the tracklines look OK" (or not) can be done to review a subsequent design-based analysis, the fundamental basis of which — that of randomization and replication — is statistically flawed. One is forced to review the resulting analysis as if the (possibly non-randomized) realized transects came from a randomized (or systematic with randomized startpoint) design. This is an unnecessary dilemma which can be avoided with some degree of attention to appropriate randomization. It may, of course, be possible to use model-based estimation; guidance on when this might be appropriate is given throughout other sections of this report.

## 2 Trackline placement and variance

Obtaining reliable results from a line transect survey depends critically on good survey design. In a cetacean management context, systematic random placement of independent sampling units (e.g. transect lines) is typically preferable to a completely randomized design, because systematic designs usually provide lower variance (see, for example, Fewster et al., 2009), particularly when density varies spatially (which it invariably does). The Guidelines describe a "classical" approach to trackline placement: placing a random grid of evenly spaced parallel lines within each stratum. In this paper, we consider this type of design to be *systematic* (with the randomness being determined by the placement of, say, the leftmost transect, and the locations of others in the grid being deterministically set based on the location of the first and the spacing). As noted in the Guidelines, a widely acceptable alternative common in shipboard surveys is a modification of this systematic scheme, whereby a start point is randomly determined, and the "grid" is a set of zig-zag lines set evenly across the region of interest.

Clearly the survey design has several intertwined components — practical considerations, transect spacing, number of transects/sampling units, level of effort available; type of sampler (eg. parallel lines, zig-zag lines), coverage probability, etc. A good general overview of these is provided in Buckland et al. (2001b, chapter 7), whilst Strindberg and Buckland (2004) consider the use of zig-zag survey designs in detail. In a strictly theoretical sense, it should be noted that an unbiased estimator of variance cannot be calculated using design-based principles (since there is no replication), but usually — and for most intents and purposes — this is ignored. Analyses then proceed as if each systematic transect (sampling unit) was in fact independent. These analyses are widely accepted in the peer-reviewed literature. For wildlife management, they have a practical advantage in that the CV generated from data collected using a systematic sampling scheme is likely to be more precise than that generated from a simple random sample (because the transects are more likely to have covered the variation in underlying density across the whole region). There are various rules of thumb associated with the question of "how many transects are required...?" but as with all such rules, the answer really depends on many factors specific to the survey in question. We consider that a *minimum* of 10 transects across a survey region are needed for design-based variance estimation; as **FewBuc04** note though, the design should preferably include "more than 20". Improved precision of a survey-region-wide design-based estimator can be obtained using post-stratification of adjoining geographic blocks (in this case, at least two transects per stratum are required Fewster et al., 2009) but if estimates are required by stratum, then certainly no fewer than 10 transects per stratum would suffice. Improvement in precision can, of course, also be achieved using other (non-geographic) post-stratification schemes. Appropriate care is needed as this can lead to substantial negative bias in the variance estimator if the stratification is based on the response rather than an objective covariate.

Recently Fewster (2011) developed a "striplet" estimator of variance for systematically-design surveys. Whilst not requiring a systematic design to be treated as if it were a truly random design, this estimator is no longer

simply design-based; it draws upon model-based estimation as well. As yet, the estimator is for parallel lines designs; it does not extend to the zig-zag samplers widely used in cetacean surveys. Other research in this area includes a recent study by Barabesi and Fattorini (2013), who propose a systematic design schemes which yield equal coverage probability. These schemes do not currently accommodate the practicalities of "joined-up" transects (minimum off-transect effort) so are probably not directly relevant to the Committee's work right now, though they may be in the future.

# 3 What we mean by model-based

The choice of framework (design or model) affects the choice of estimators, especially variance estimators (see discussion of issues in Thompson (2002, chapter 10)). For example, an indefensible model would be one in which estimation relied on the assumption that whales were uniformly distributed across a stratum, and could lead to poor results. The design-based framework doesn't make any assumptions about the distribution of animals: variance is due to the sampling process.

However, in line transect sampling it is impossible to rely entirely on the design framework, because there is a fundamental source of *non-sampling* error, namely imperfect detection. There is no way round this except to use a model. The basic inferential framework can be design-based, but a model must be fitted at the detectability stage (Buckland et al. (2001b, section 10.3.)). Modelling the detection function is essentially fitting a density curve to data, and is uncontroversial in the scheme of things. The models to avoid are those that make strong assumptions about the distribution of animals over a wide area. In line transect sampling, assumptions about the distribution of animals relate to that within the narrow search-strips only.

In this report, model-based analyses are recommended in some circumstances — unless specifically stated otherwise, then a 'model-based analysis' refers to a spatial model (e.g. Miller et al., 2013; Bravington and Hedley, 2012; Johnson et al., 2010). It is important to note that a spatial model in this sense can also be relatively uncontroversial; it is simply fitting a flexible smooth model to data rather than making some assumption about, say, a particular parametric relationship between density and sea surface temperature. Section 6 details the mathematical framework for the spatial models as they relate to cetacean sightings survey data.

# 4 Analysis requirements

In this section, we consider some decisions that need to be made when analysing line transect survey data. For the design-based estimators at least, these are all well-established and certainly not new, but with the relatively recent advent of useable, reliable spatial models, we considered that it would be useful to formulate an overall map of the decision-making process when faced with survey data.

As a starting point, then a statistically-defensible design-based analysis must demonstrate that coverage probability within the survey region is (nearly) uniform, or is estimable (see Fig. 1). If coverage probability is unknown, then there is little alternative but to go down the model-based route (but see section 11); in this case, transect placement does not have to be randomized and variance does not have to rely on replication, but transects should be spaced across the survey region to obtain even coverage (ideally, spaced to minimise interpolation and extrapolation beyond the range of covariates encountered on the transects). If coverage probability is known, but unequal, then model-based estimation remains a good option. This case also opens up the option of using Horvitz-Thompson-like design-based estimators; in addition to modelling the detection probability of

each school, the inclusion probability of each school being in the sample (based on the sampler location) is also accounted for. We have no experience of this type of modelling but in a simulation study to investigate potential improvements in precision when accounting for gradients in animal density in the design (with increased coverage probability in higher densities), **Rexstad2007** found little improvement compared to traditional equal coverage probability estimators.

It is perhaps the norm for equal coverage probability to be assumed, rather than demonstrated (either by appeal to the randomization scheme and type of sampler, or by simulation). In this case, a design-based analysis should at best, be flagged as potentially biassed, and as a minimum, we recommend that the results be checked against those from a spatial model (Fig. 1). In such circumstances, if one is going to the trouble of fitting a spatial model as a check, then in our opinion, it also makes sense to use the estimates and variance from the spatial model too (they should be similar, and are not dependent on ignoring the basic sampling theory).

In a pre-defined survey region with (almost) equal coverage probability, unbiassed estimation from design-based or model-based analyses is possible. For a design-based analysis, the sampling unit (typically the transect) usually forms the basis of the variance estimator, so a sufficient number of replicates is required (15 or so, preferably more than 20) for stable estimates of precision. As suggested above, if fewer than 15 transects were surveyed, then we would recommend at least a check of the design-based results using a spatial model.

If, with equal coverage probability, there are more than 15 transects, then fully design-based estimation is feasible and widely applied in wildlife management. *Distance* software Thomas et al. (2009) and its R-packages therein offer a range of approaches to refine analyses dependent on the the data (stratification, different detection functions, options for estimating mean school size, incorporating covariates, estimating $g(0)$ etc.). It is not possible to incorporate diagnostics for all of these options in the Guidelines, but where analysis choices have been made, then it seems reasonable to provide logical justification (as well as say, appealing to statistical information criteria). An example might be, say, where school size was found (by AIC) to be a significant covariate in a detection function model, but that detectability *decreased* with increasing school size. This *could* happen (eg. if all small schools behaved conspicuously, and larger schools were secretive) but unlikely, and therefore it might be quite alright to ignore the AIC verdict! If we were reviewing an analysis, we would generally find logical arguments and commentary of this type *as useful* as a sensitivity analysis — there is no single right way do do an analysis, but there are lots of wrong ways!

Wherever it is possible to fit a design-based model, then it is also possible to model the data with a spatial smoother, though not necessary. In sections**??**-**??**, we describe our experience with this type of modelling, and suggest diagnostics useful for reviewing spatial model output (section 10).

# 5 Example: IWC POWER survey 2010-2014

Despite best intentions, there are instances in the Scientific Committee where a large investment in survey (and analysis) time has been made, and the resulting estimates are intended for use in *Implementation Simulation Trials* or the *CLA,* but the Guidelines have not been met. It seems that the Scientific Committee is faced with a choice: stick to the "letter of the law", ie. do not accept any (design-based) estimates from such surveys; make a subjective judgment that the survey was completed in a reasonable way and, if the analysis was also reasonable, "accept" the estimates; request further analyses (which may be model-based); or perhaps some combination of these.

Since little recent information on cetacean distribution and abundance was available in the eastern North Pacific,

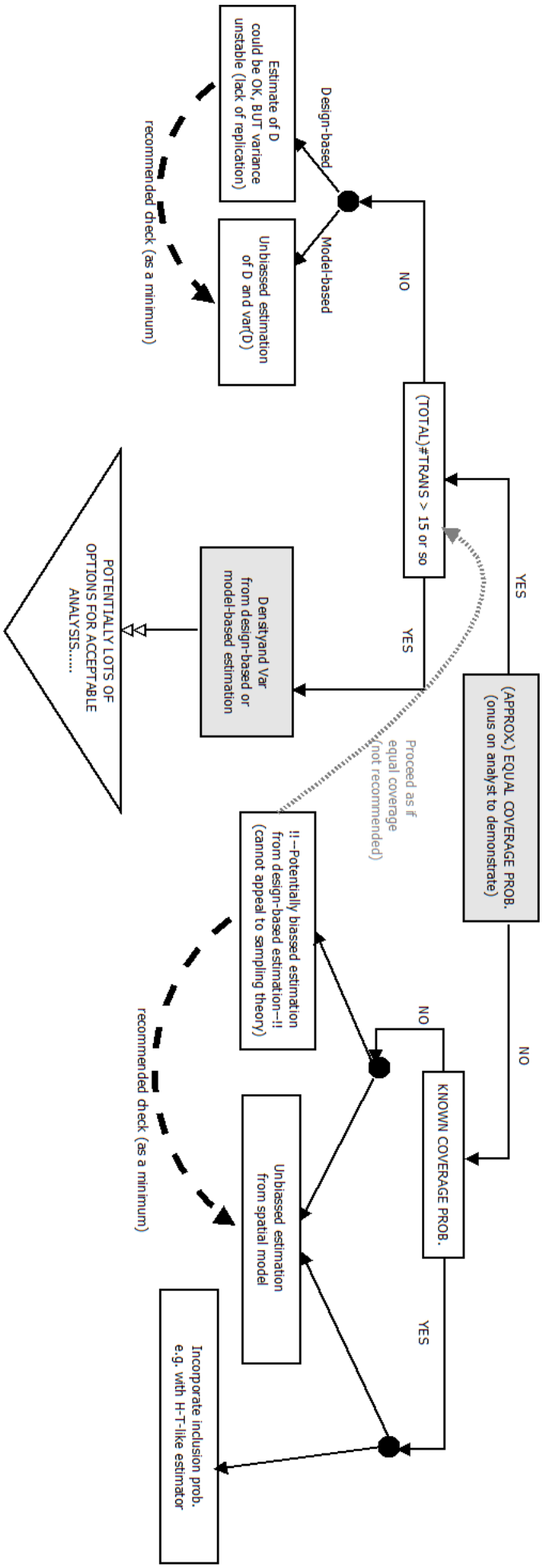


Figure 1: Flow diagram showing estimation decisions that ought to be made (and potential pitfalls) depending on survey design.

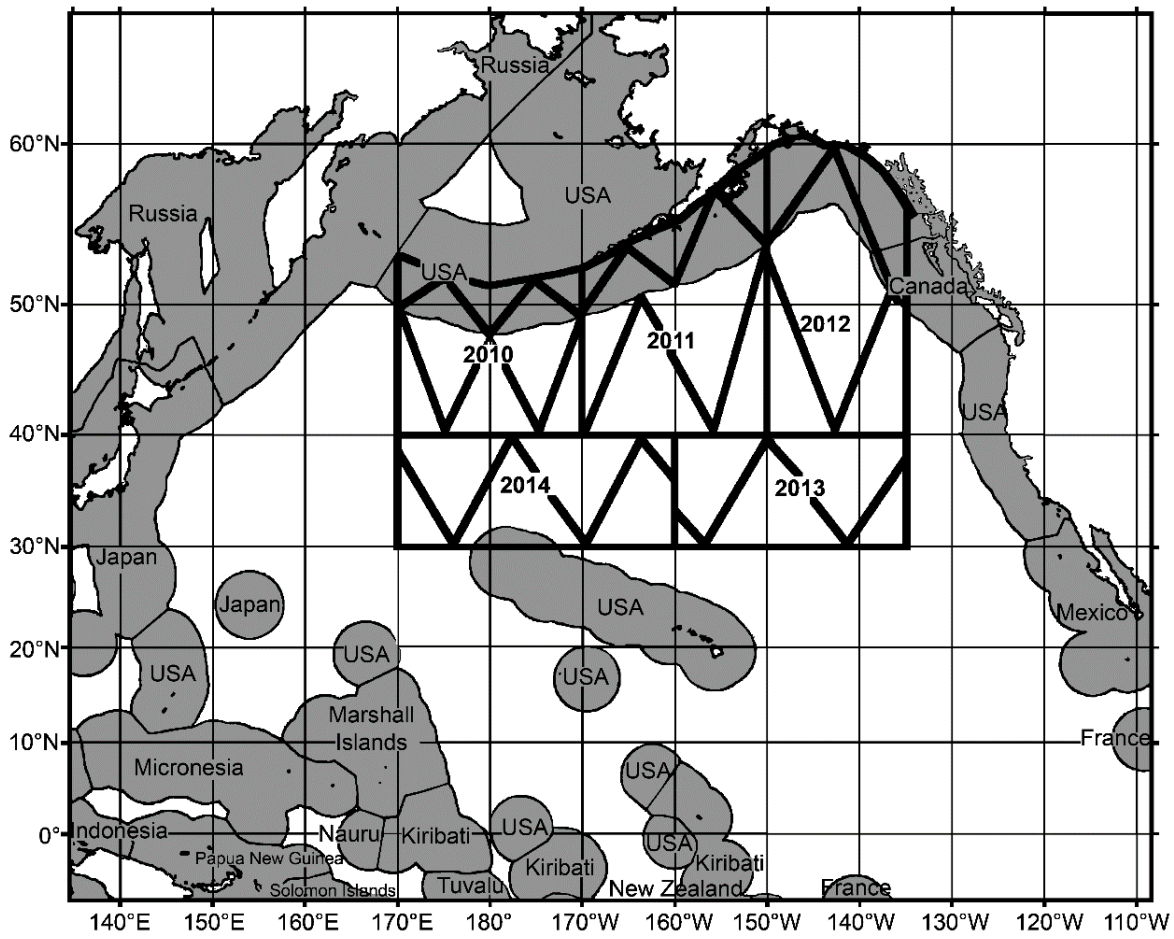


Figure 2: Research areas and tracklines for the IWC-POWER cruises 2010-14. Reproduced from Fig. 1 of IWC (2014).

a general objective of the short-term phase of the POWER programme was to provide reasonable sightings survey coverage across the area, and broadly update and increase current information, to inform planning of the medium- and long-term phases of the programme. Specifically the emphasis was *not* on obtaining the best abundance estimates possible. As can be seen in Figure 2 (reproduced from Fig.1 of IWC, 2014), the (planned) trackline coverage achieves this objective in a reasonable way.

At SC65b, a preliminary analysis of sei whale data from the first three IWC-POWER surveys (2010-12) is presented (Hakamada and Matsuoka, 2012). For illustrative purposes only, we suppose this analysis were to be reviewed with a view to "acceptance of the estimates" as per the Guidelines. We do not make comments on the fine details of the analysis (as that is for the Scientific Committee) and have chosen this survey as we were ourselves involved in planning the survey (so partly responsible for any deficiencies!) A simple checkpoint list might in this case be:

• does the analysis assume uniform coverage probability, and if so is this reliably demonstrated?

• what is the sampling unit for variance estimation, and is this appropriate?

• was the coverage severely affected by weather conditions?

As seen in Fig. 3, the planned tracklines in 2010 end in the circled region — in this case the corner of the westernmost edge of the survey that intersects the inter-stratum boundary. With a randomized design this could just possibly have happened, but we consider it unlikely.

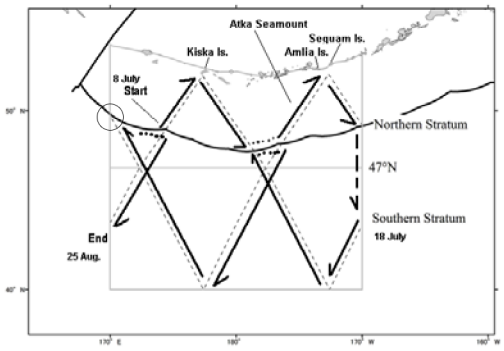


Figure 3: Planned tracklines for 2010 IWC-POWER survey. Reproduced from Hakamada and Matsuoka (2012). Circle added to show tracklines appearing to start/end in a "corner".

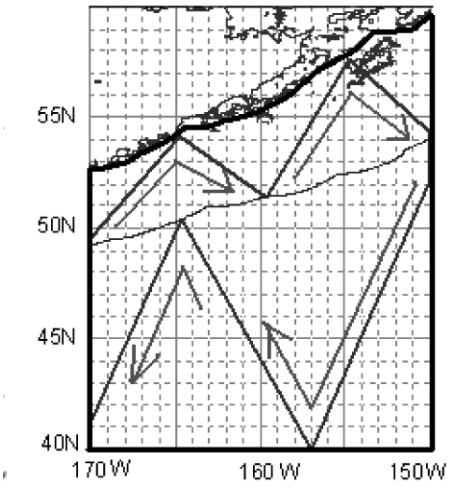


Figure 4: Planned tracklines for 2011 IWC-POWER survey. Reproduced from Hakamada and Matsuoka (2012).


Figure 5: Planned tracklines for 2012 IWC-POWER survey. Reproduced from Hakamada and Matsuoka (2012).

In 2011, the tracklines do not start in a corner, though they almost do (Fig. 4). Visually, with the amount of effort presumably available, then the tracklines appear to cover the region reasonably well, but it appears that some areas of the study — perhaps for pragmatic reasons — may have had zero coverage probability (eg. in all corners of the region but the lower left one), so it's not obvious that the design gives (nearly) uniform coverage probability in each stratum. In our view, then if this analysis were to be considered for use in *Implementation Simulation Trials* or the *CLA*, at least some further details on the randomization scheme used to design the tracklines would be required, and preferably an illustration of any variation in coverage probability.

In 2012, the planned design does not quite cover the longitudinal range of the survey area (Fig. 5), though other than this, the lines cover the region in a pragmatic way for the amount of search effort available. As in 2010, one trackline happens to finish up exactly at the westernmost point on the interstratum boundary — we consider this unlikely to have happened by chance — and thus the underpinning sampling principle of a design-based analysis is undermined. Given these apparent departures from a randomized design as intended in the Guidelines, and no obvious way to estimate the coverage probability post-survey, then we consider that a model-based framework is needed to either analyse these data reliably, or at least, to check the design-based results. Furthermore, for each survey, there are fewer than 10 transects, so conventional transect-based estimation of variance is likely to be unstable.

Regarding "effort lost due to weather", for these surveys the diagrams depict a reasonably high percentage of the planned tracklines being searched. Of course, for careful reviewing purposes, some consideration should be given to the weather conditions suitable for surveying the species for which abundance is being estimated (to avoid problems related to confounding between two types of zeroes: "no whales seen because they weren't there" and "no whales seen because they weren't able to be seen").

# Spatial modelling

## 6 What is "model-based abundance estimation"?

This section describes the framework for our discussion "model-based abundance estimation", which we also refer to as "spatial abundance estimation". It should be straightforward for statisticians, who nowadays can be expected to be familiar with GAMs, but is deliberately somewhat technical to reduce the danger of misinterpretation. In this initial description, and in most of the document below, we omit references to school size, which is addressed separately in section 9.

- The aim is to estimate total abundance inside some region of interest (called "the Region" from now on), based on tracklines of sighting survey effort that may fall inside or outside the Region. It is also important to assess the uncertainty (the CV) in as stable and unbiassed a way as possible, at least if the true CV is small; if the CV is large, then it doesn't matter so much whether the estimated CV is slightly larger or smaller, because the abundance estimate will be effectively ignored for management purposes.

- For analysis purposes, the trackline is divided into numerous small snippets of effort, each with fixed covariates that might affect detection probability and an observed number of encounters (possibly zero).

- Detection probability parameters are estimated, presumably as a function of covariates, either prior to fitting the spatial model or perhaps simultaneously. This can be done in any suitable away: MCDS, MRDS, cue-based modelling, etc.

- The estimated parameters are then used to compute the average detection probability for each snippet, given the corresponding covariates.

- The expected number of encounters $Y_i$ in snippet $i$ located around a position $x_i$ in some coordinate system, is described by a log-link function with a linear (or at least additive) predictor, like in a GLM:

$$\log \mathbb{E}[Y_i] = \log \hat{\mu}_i + s(x_i; \beta, \omega) + \ldots$$

  where $\hat{\mu}_i$ is an offset for the estimated detection probability or effective strip width in that snippet, and $s()$ is a function that can compute the log-density surface of animals at *any* location $x$ in the Region (not just on the tracks). Specifically, $s()$ is a "smoother" (section 7.2) that uses -parameters $\beta$ to control the actual shape of the density surface[1], and "surface variance" parameter(s) $\omega$ to describe how much the surface is likely to vary from place to place. The dots are there in case extra terms are needed (e.g. section 8.4.3).

- the observed number of encounters $y_i$ is assumed to follow some statistical distribution with the above expected value and perhaps extra parameters to describe overdispersion or correlation between neighbouring $y_i$;

- some algorithm is used to estimate the parameters $\omega$ and $\beta$ (and any others related to the dots).

[1] Some smoothers (mostly old ones) are set up with an implicit rather than explicit parametrization, and there is no numerical "$\beta$". We do not consider them here.

- the estimated $\beta$ are used to predict animal density across a fine grid of points, with the log-offset fixed at zero (as if detection was certain). These density estimates are scaled by area and added up to form the abundance estimate
- the uncertainty in this abundance estimate is computed based on the estimated uncertainty in the $\beta$ parameters and the estimated detection probabilities (and perhaps the variance-related parameters too).

In fact, it is *almost* possible to set up a conventional stratified abundance estimate in this framework. The crucial feature of a spatial model for abundance estimation, though, is that *animal density is assumed to be spatially correlated*. In fact, the key is the parameter $\omega$, which governs how strong the correlation is as a function of distance (in a way that is specific to the type of smoother used). In a stratified estimate, density is assumed to be constant within strata, but there is no assumed relationship between density in neighbouring strata (and there is no need to use a parameter to describe "correlation", hence no $\omega$). The correlation is what makes it possible to predict (with uncertainty) the density across the whole Region, just using encounter rates on tracklines that cover only a small proportion of the total area.

Smoothers have been developed to suit more than one statistical paradigm, but to us the most straightforward is a Bayesian perspective in which the density surface is viewed as a single realized "value" from a prior distribution of possible density surfaces (Wahba, 1990). The prior distribution is controlled by the type of smoother chosen, and the value of the "surface variance" $\omega$; in complete contrast to design-based analyses, the data (regardless of why it was collected in a certain way) updates the prior, and inferences are made from the posterior distribution of the density surface. For a broad class of smoothers, this leads naturally to a "penalized likelihood framework" (Wood, 2006) in which the shape-controlling parameters $\beta$ are treated as random effects drawn from a prior with hyperparameter $\omega$. The penalized-likelihood framework is how GAMs are represented in the R software `mgcv`, which is the shrine of computationally-sound practically-tested smoothers[2]. In this document, we use the term "basis-and-penalty smoother"; they are so computationally convenient, and the class is so broad, that we have not considered other options.

This Bayesian approach to smoothing leaves considerable computational freedom in how estimation and inferences are carried out: MCMC is possible, but not necessary. For example, the software INLA (Rue and Martino, 2007; Martino and Rue, 2009) can do high-accuracy fully Bayesian inference much faster than MCMC, for a subset of basis-and-penalty smoothers[3]. Alternatively, fits based on REML principles (using Laplace approximation) can give approximate but quite accurate answers and often very quickly; this is the default approach used for GAMs in `mgcv`. REML-based inference can also be implemented in a more flexible but still largely automated way in ADMB (Skaug and Fournier, 2006). It is important to realize that, even though MCMC fits appear superficially quite different from the GAMs fitted by `mgcv`, the underlying ideas are quite similar.

Our own spatial abundance estimation efforts have (since about 2004) always used basis-and-penalty smoothers, either directly inside `mgcv` or by starting with basis-and-penalty matrices generated by `mgcv`, then adding extra statistical nuances that are not part of GAMs or `mgcv` and doing all estimation and prediction inside an ADMB-like environment. The software `dsm` (Distance Spatial Models; Miller et al., 2013), which is designed for spatial abundance estimation from line transect and distance-sampling data, uses `mgcv` as its computational engine. Our comments in this document are of course somewhat determined by this background: in particular (but not only) the development of spatial models for the IDCR/SOWER CP II/III Antarctic minke whale

[2]There are, or used to be, other types of GAM where smoothers were represented in different ways, but they have fallen out of favour because of the theoretical and computational advantages of the penalized-likelihood framework.

[3]It is not clear whether INLA can (yet) handle the additional distance-sampling aspects required for spatial abundance estimation.

surveys, a well-collected dataset which nevertheless exhibits almost all the difficult features that make spatial modelling sometimes both a necessity and a challenge. However, the general issues requiring attention should be applicable to all smoother-based approaches to spatial abundance estimation.

# 7 Choices in spatial abundance estimation

Spatial modelling forces the analyst to make several additional choices, beyond those required in any Distance-Sampling analysis. From the PoV of the Committee reviewing a paper that presents a spatial abundance estimate, [if we the authors were reviewing...] we would expect to see some commentary addressing at least the points below, explaining how the choice was made in each case. In this document, we try to explain why the choice might matter and, in several places, we have also made suggestions about likely good or bad choices, based on our own experience. These suggestions should not be treated as definitive; our experience is decidedly finite, and there might be good reasons why another analyst might make a different choice in some cases. Our main aim is to suggest which issues the analyst needs to think about and report on so that Committee can assess whether a presented estimate is satisfactory, not to dictate what every choice must be.

## 7.1 Coordinate system

Spatial models assume that places closer together are more correlated, and so the definition of "closeness" can matter. Usually (with exceptions, some noted below), smoothers measure closeness using Euclidean distances $(x_1 - x_2)^2 + (y_1 - y_2)^2$ in some 2D coordinate system, which should therefore be chosen so that its Euclidean distances are approximately the same as real physical distances "as the whale swims" (though see the discussion of boundaries in section 7.2.1). Clearly, latitude and longitude are generally not suitable unless transformed, because the real distance travelled in moving 1 degree West is not the same as in moving in 1 degree North except near the equator. One option is to use a spherical coordinate transform that moves the "equator" to pass through the centre of the Region, but we have usually employed the simpler trick of using $(\text{long}, \text{lat} \times \cos(\text{long}))$.

The earth is not flat (Cohen, 1994), so some distortion of distance and/or area is inevitable when mapping to *most* coordinate systems suitable for smoothers, but it is not important to have extreme fidelity in preserving distances. Accuracy matters more for short distances, because density will be effectively uncorrelated when "the distance" is large enough, regardless of the co-ordinate system. If the model has to cover a huge area (e.g. the entire Pacific Ocean), then it *might* be worth considering a spherical-shell co-ordinate system with less distortion, such as the `Spherical.Spline` smoother.

Not all smoothers use their coordinate system in the way just described. In particular, it may not make sense to assume *isotropy,* i.e. that the rate of change of density is the same whether moving North-South or East-West. When a strong environmental gradient is expected, for example with distance offshore or seabed depth, then it may be more sensible to parametrize in terms of the 2D system such as (distance-along-coast, distance-offshore) and to deliberately use a *tensor product smoother* (section 11.2) which is not isotropic. In certain cases, this "along and away from" approach can be more effective statistically than using general-purpose 2D coordinates with an isotropic smoother, in terms of requiring fewer "estimated degrees of freedom" to fit the data. The final abundance estimate still needs to be formed by predicting over an evenly-spaced grid in true spatial coordinates (which are mapped into "along and away" coordinates to make the predictions). Having said that, more conventional isotropic smoothers that have been adapted to deal with boundaries (section 7.2.1) should still give reasonable inferences in those settings (just with larger CVs), and in any case if the boundary

is strongly curved then it may not be possible to construct an "along and away from" coordinate system without greatly distorting physical distances.

It would be conceptually possible to take this line of thought further, by assuming that density is a smooth function of chosen environmental covariates and/or spatial covariates. That lies rather beyond the scope of this document, but we note that such approaches are likely more reliant on having chosen "a good model" than a standard spatial smoother is (whether isotropic or non-isotropic).

## 7.2 Type of smoother

There are many different types of smoother[4] but, provided there is ample survey effort throughout the prediction region, then it should not matter which type is used: the estimated abundance should be similar, and so should its CV. More formally, with smoothers the "model" is not very restrictive and it is often reasonable to expect that "model uncertainty" (i.e. different estimates due to different models, which is hard quantify formally and also entails a lot of computational effort to assess) should be small relative to the "intrinsic uncertainty" in results from any single model (which is easy to assess); this would not be the case, for example, if someone was trying to estimate abundance by fitting a linear relationship between density and water temperature. A few comments on particular smoothers are given in the Appendix.

However, "ample survey effort throughout the prediction region" is far from universal, and we are aware of two common situations where it is unsafe to just assume that smoother choice is immaterial:

- (less common) near irregular boundaries. For example, if the region of interest includes a peninsula, then a badly-chosen smoother will smear the densities from both sides. If there are imbalances in both effort *and* density across the peninsula, the smearing will give a biassed estimate overall. (EG pic)
- (more common) extrapolation. Note that extrapolation is not a well-defined concept in 2D ; *all* spatial models in 2D involve some extrapolation (. EG pic a V shape in a square.)

In both cases, some types of smoother work better than others. Note that the obvious test of "try several and see if the estimates are robust?" is not very sensible here; a well-chosen smoother can give quite a different answer to a badly-chosen smoother, and the mere fact that the answers are different is no reason to discard *both* of them. Instead, it is necessary to pay attention to theoretical reasons (e.g. proven optimality in specific situations) and practical experience for choosing between smoothers. It turns out that solutions to the two issues are related.

### 7.2.1 Boundary behaviour (finite-area smoothers)

A variety of solutions have been proposed from the late 1970s onwards, but most have had theoretical weaknesses that have eventually been exposed as practical problems. As of about 2012, the most successful seemed to be "soap-film smoothing" (Wood et al., 2008), which is available in `mgcv`. It does have a respectable theoretical underpinning, and we have generally found it to behave well, although problems have reported in some examples. Certainly, some care is needed to set up Soap, there are limits to the amount of coastal detail that it

[4] Not all spatial models generate predictions that are smooth, nor even continuous: for example, spatial blocks treated as independent random effects, and certain types of wavelet. Nevertheless, such models may be able to give perfectly reasonable inferences about *aggregated* abundance across a region, and there might be good reasons to use them in some situations. The conceptual and practical issues are largely the same whether or not the predictions are smooth/continusous, and we refer to all "spatial-random-effect models" as "smoothers" here.

can handle before its predictions become unreasonable, and it lacks a clear theoretical justification when part of the "boundary" is an artificial limit in open sea. A visual check for artefacts in the fitted surface is advisable, as always when fitting smoothers.

**B**etter performance than Soap has been reported from more *ad hoc* smoothers, e.g. Scott-Hayward et al. (2013)); however, if (as in that paper) the smoother lies outside the basis-and-penalty framework, then it could be difficult to propagate detection function uncertainty (see section 8.4.3). More recently, Miller and Wood (2014) have proposed a two-stage scheme, with apparently good properties, that can be embedded in a basis-and-penalty framework. The history of boundary-respecting smoothers shows that some apparently promising smoothers lead to rather strange properties in their fitted surfaces, e.g. in the direction of gradients near the boundary. If boundary results are important (e.g. if a high proportion of abundance is estimated to be close to the boundary) and smoothers are used that have not been investigated practically and theoretically in the literature, then we recommend a sensitivity check against Soap.

### 7.2.2 Extrapolation

It is important to realize that some degree of extrapolation is unavoidable when fitting 2D density surfaces to line transect data. There are always subregions outside the "survey perimeter" where animal density must be predicted, e.g. near some corners and/or along the edges of the Region. It is also worth remembering that survey effort in poor sighting conditions can be very unlikely to .

This does not usually cause serious problems for spatial abundance estimation unless these subregions are large or very far from informative data (or unless a ill-advised choice of smoother is made). Because smoothers "know" that density can vary spatially, they tend to give high CVs for predictions in subregions that are far away from meaningful data— which is exactly what is desired. (In contrast, for example, a model that assumes constant mean density would be much more dangerous for extrapolation; it would predict a misleadingly low CV because it believes that the mean will not change.) However, point estimates are important too, and some types of smoother are prone to extrapolating to extremely large point estimates— a clearly undesirable behaviour— while others are less so. The reason lies in what parts of the "signal" (underlying trend) a smoother is designed to penalize or compress" (i.e. to smooth away).

Thin-Plate Splines were among the earliest 2D smoothers in general use, and are still something of a default choice. They have a basis-and-penalty interpretation, where the penalty is proportional to the integrated squared second derivative of the fitted surface. Linear trends have zero second derivative, so are not penalized by TPS and are "free parameters" that the spatial model will adjust to fit the data as it likes. The linear trend component of a fitted TPS will, of course, be increasing over half of the $(x,y)$ plane, and it will continue increasing linearly however far it is extrapolated. Since the smoother is actually describing *log* density, this means there will be *exponential* rates of increase in predicted density in half the plane. It is easy to see how extremely high abundances can be extrapolated. This leads to overall positive bias as well as occasional downright stupid answers, because the increase and decrease do not balance after exponentiation: extrapolation in the direction of decreasing trend can never drop below zero.

The problem is the that prior distribution assumed by TPS is biologically unreasonable; in nature, density trends do not stay exponential over large spatial ranges. The solution is to use smoothers that penalize linear trends (as well as higher-order trends, which are penalized by all common smoothers including TPS). Even though the fitted log-density surface will probably have a linear component over the range of the data, the linear component will

One example is, coincidentally, the Soap-film smoother just mentioned (see also section 11.2), which can be used with fictitious boundaries as well as real physical ones; it is probably this "smoother taming" behaviour, as much as its boundary-respecting properties, that have led to its occasional use for spatial models of animal density. Unless physical boundaries are involved, though, a better solution on theoretical grounds (and thus probably practical ones) is to choose a particular type of Duchon spline (section 11.2). Duchon splines are a recent addition to the `mgcv` stable, but we have heard encouraging reports on their stability (D.L. Miller, *pers. comm.*).

Extrapolation (a rather ill-defined term in this setting) is thus unavoidable with spatial models, but not necessarily problematic. If smoothers are working properly, they should automatically give high CVs when make large extrapolations (high compared to, say, the CV of a prediction over just the central part of the surveyed region). And if the smoother is chosen appropriately (avoid TPS!), then ludicrous extrapolations should not occur, and should at least be obvious from examination of contour or image plots if they do occur.

Larger extrapolations do place more dependence on the "underlying model" (i.e. the type of smoother), as shown by the TPS vs. Duchon spline comparison. We have no particular suggestions for assessing "how much extrapolation is too much", but it is fairly clear that the case of least concern is when estimated densities are falling with increasing distance around the "perimeter" of the surveyed areas, so that extrapolations form a small proportion of the total estimate. If extrapolation contributes a high percentage of the total, it may not be wise to trust the estimate.

### 7.3 Implications for design

A spatial analysis can be used as a "rescue option" when a design-based analysis is inappopriate, but it can also be planned deliberately from the start. If so, one might choose to design the survey differently. For example, since strange behaviour at the edges remains a potential concern with spatial models, even when good smoothers are chosen, it might be more statistically efficient (i.e. lower final CV per dollar spent) to spread out the effort unevenly, putting more effort close to the outer edge. Also, strata are not a concern, which may improve logistic flexibility. And, importantly, it is OK to have gaps in coverage with spatial models, so it may be reasonable to design tracklines with built-in expectation of randomly incomplete coverage due to bad weather; this might permit wider overall coverage that is in effect less dense, with better CV overall. "Optimal" design for model-based analysis does not seem to exist yet as a statistical field, but it should; meanwhile, case-specific investigation may be useful in developing efficient designs.

Peel et al. (2013) used this model-based approach to design multi-species fishery surveys (effectively with point samples, not line transects), using pre-existing CPUE data to determine spatial and temporal variability, and allocating effort to keep overall CVs low. It is not necessary to simulate in order to predict the CV, so different candidate designs could be checked quickly.

## 8 Variance estimation

Smoothers include one or more variance-related parameters that can be estimated from data: at least one "smoothing parameter" (a.k.a. "wiggliness penalty") will dictate the scale of large-scale spatial correlation, and there may be one or more governing "local clustering" and/or "overdispersion". Once those parameters have estimated, there is a conceptually straightforward way to compute the uncertainty of an abundance esti-

mate, and there are not many choices to make[5]. However, some choices are needed in order to estimate the parameters.

## 8.1 Snippet size

Sighting surveys are basically continuous in time and sometimes space, but for analysis and computation it is convenient to break them into smaller units (here called "snippets", or "segments" in `dsm`). The general idea is that "conditions" (both effort-related, and animal-density-related) should be roughly constant within each snippet.

It is certainly useful to break large/long chunks of effort, such as transects that last for hours, into shorter pieces. There are several reasons, perhaps the most important is

- an important benefit of spatial models is that *within-transect* variability can help to assess the uncertainty in the final estimate of a spatial model, whereas with most design-based analyses only the between-transect variability is useful, and there may not be enough transects to reliably estimate the latter. Smaller snippets give more opportunity to estimate variability
- detection-related covariates such as weather might change within long snippets, which is computationally awkward especially when computing diagnostics;

However, with smaller and smaller snippets, diminishing returns eventually set in, and it is possible to make the snippets too short:

- when there is a substantial probability that a sighting event could begin in one snippet but end (e.g. pass abeam of the vessel) in another one;
- when the sheer "number of numbers"— almost all of which are zero— becomes computationally burdensome
- when the snippets are too small to allow estimation of clustering/overdispersion

The lattermost point is subtle, but can cause negative bias in CVs. Most types of spatial smoother[6] can only deal with fairly large-scale spatial effects, where a substantial change in density can only happen over a distance that corresponds to many hours of survey effort. However, in real surveys it is quite common to come across smaller obvious hotspots, where numerous schools/animals are seen over a much shorter stretch of time. The precision of the final abundance estimate really depends on how many of these hotspots are encountered, not on how many schools/whales are encountered overall. For example, if only three hotspots are found in a survey, then a hypothetical repeat survey might well have encountered just one hotspot, or six; it doesn't really matter whether each hotspot comprised two sightings or 10. If the snippets are too small, less than the size of a hotspot, then each one is likely to contain just one or zero sightings, and a typical spatial model is not equipped to "notice" that there tend to be runs of consecutive snippets with non-zero sightings. In some sense, "the residuals will be autocorrelated" (though this is hard to diagnose in practice), and the overall CV will be underestimated. If longer snippets are used, then autocorrelation is reduced and will be manifested instead as overdispersion, in

[5] In some contrast to design-based estimation, where several different variance estimators have been proposed based on different underlying principles

[6] At least, amongst the smoothers we know of that have been used in cetacean abundance estimation; the statement isn't universally true, e.g. for kriging-based methods.

the sense that the variance of encounters across snippets will exceed what would be expected from a Poisson distribution. This can be allowed for statistically by changing the statistical "family" of the GAM to something other than Poisson. For example, the `dsm` package allows Tweedie, Negative Binomial, and Quasipoisson responses. We don't have direct experience of Negative Binomials in this context, but we have found Tweedie preferable to Quasipoisson (which is a special case of Tweedie) because of the extra modelling flexibility.

Provided that the snippets are neither too big nor too small, and that overdispersion is allowed for, then the overall CV of an abundance estimate should not be very sensitive to snippet size, because the estimated overdispersion parameter will largely compensate. However, overdispersion can only be realistically estimated when a reasonable number of snippets, say 10, have (after fitting some kind of spatial model) a reasonably high expected number of encounters, say above 1. If there are very few encounters overall, this might entail making the snippets unreasonably large. In such cases, the overall CV should be large anyway, so that missing any extra variability due to overdispersion would not matter.

In a fully Bayesian setting, it is possible to devise "smoothers" that handle small-scale clustering and large-scale spatial variation in exactly the same framework, though we have no experience with such approaches. In REML-like settings, an elegant and computationally efficient solution is to model the encounters in neighbouring snippets with a local fine-scale clustering process, whose parameters can also be estimated. For example, the Norwegian North Atlantic minke whale surveys originally used a Neymann-Scott clustering process, now changed to the simpler Markov-Modulated Poisson Process (Skaug, 2006). The SPLINTR model for SOWER minke whale sightings also uses an MMPP. There are computational benefits in separating small-scale from large-scale spatial variability this way, because the small-scale variability can be ignored when estimating aggregate abundance, and can be allowed for without adding huge numbers of parameters. The disadvantage is that special software needs to be written (e.g. in ADMB), at least until the proposed incorporation of MMPP into `dsm` . The advantage for the analyst is that results should no longer be at all sensitive to the choice of snippet size.

In principle, a perfectly-chosen small-scale clustering model could make it possible to use encounter rates from Closing-mode as well as Passing-mode sectors of the same survey. One possible problem with Closing-mode encounter rates is that, if animals or schools are clustered and the time-to-close (which is off-effort for encounter rate purposes) is comparable with time between encounters in high-density zones, then overall search effort will tend to be systematically reduced in of higher-density zones, leading to negative bias. A fine-scale clustering model might be able to allow for this effect. We tried doing so for SOWER Antarctic minke whale data using the MMPP process for fine-scale clustering, but still found that overall abundance estimates were substantially reduced when including Close-mode encounter rates. It may not be realistic to "model away" bias in Closing mode encounters, and careful attention should be paid to justifying (or bounding the likely magnitude of any bias from) the use of Closing-mode encounter rates for abundance estimation.

To summarize: a bad choice of snippet size can bias the CV. Smaller is good, but too small can lead to underestimated CVs. The practical advice might be:

- ideally, to use a fine-scale "local clustering" model with estimable parameters, as well as a large-scale smoother;
- if not, then the possibility of local clustering should be allowed for by using an overdispersed "family" in the GAM (e.g. via a Tweedie distribution), and do sensitivity checks to show that the abundance estimate and CV is not sensitive to snippet size when snippets are in a size range sufficient to estimate overdispersion.

- if it is not possible to make the snippets big enough to give a chance of detecting overdispersion, then the CV is probably high enough anyway that the possibility of local clustering is irrelevant.

## 8.2 Smoothing parameter estimation

General-purpose smoothers that could be used for density-surface estimation can perhaps be dated back to **Hast:Tibs:1990** and were available in the software Splus during the early 1990s. At that time it was necessary to "estimate" the smoothing parameter either informally by, at best, somewhat *ad hoc* cross-validation methods. There is no longer any excuse for that approach, since with modern GAM software it is computationally straightforward to automatically estimate smoothing parameters (and other variance-related parameters) for a wide class of smoothers (basis-and-penalty) by appeal to well-founded statistical principles. In particular, the widely-used `mgcv` package in R implements several criteria for smoothing-and-variance-parameter estimation including GCV, UBRE, and REML (ref Google). All the available criteria should work well (i.e. should give similar, stable, low-bias estimates of the smoothing parameters, across hypothetical replicate datasets) if there is ample data— but of course there might not be. Since there is no way to know what is "right answer" for any particular dataset, and since the diagnostics are not obvious, it is preferable to pick a criterion in advance on theoretical/empirical grounds, rather than to make a decision based on what happens to pop out of analysing any particular dataset. Simon Wood, who is the developer of `mgcv` and who has extensive experience with GAMs, nowadays (2014) leans toward REML as the most reliable (see Wood, 2011, or current documentation for `mgcv?mgcv::gam.selection`). Our own approach has also been to use REML, not just for computational but also for philosophical reasons: it embeds the smoother into a fully probabilistic (empirical Bayes) formulation that matches the way other parameters are handled, and which leads to straightforward inferences, at least conceptually.

Fully Bayesian approaches to spatial modelling have similar notions of smoothing parameters and inference to `mgcv`'s REML approach (at least for smoothers with a basis-and-penalty equivalent, as in Wahba, 1990). They differ in that, instead of aiming for point estimates of smoothing/variance parameters, they place hyperpriors on the *a priori* distribution of smoothing parameter(s), and end up with a posterior distribution of possible values for the smoothing parameter, rather than a single point estimate (see next section). Choosing a hyperprior is vaguely similar to choosing an estimation criterion for GAMs (in that it should not matter if there is plenty of data, but there may be some choices that aren't as good when data is lacking), and there is an extensive literature on choosing hyperpriors for random effect models.

### 8.2.1 Uncertainty in the estimated smoothing parameter

A statistically coherent way to think about inferences from spatial models is: first estimate the variance-related parameters (including the surface variance), then make "second-stage" inferences about the actual density surface conditional on a particular value of the variances. The question arises of whether it is acceptable to take an Empirical Bayes approach of just using the point estimate of the variances for those second-stage inferences, or whether it is important to also allow for uncertainty in the estimated variances themselves (referred to as "hyperparameter uncertainty" in statistical literature). Empirical Bayes is known to give some negative bias to overall CVs, so the second version is theoretically preferable. However, it may be computationally awkward, depending on what software is used.

Fully Bayes approaches (e.g. MCMC and INLA) should handle hyperparameter uncertainty automatically (but see the possible exception noted below). There should also be no difficulty with ADMB's random effects op-

tion, where a Bayes Empirical Bayes approach is automatically used for an approximate delta-method-style adjustment to the overall CV. With `mgcv::gam()`, though, it would be necessary to make multiple sets of inference for different values of the variance parameters, which could be time-consuming. Wood (2006) recommended one version of such an approach, based on a parametric bootstrap of say 100 replicates; however, the results underlying that advice contained an error, and we have been told that the author has subsequently found hyperparameter uncertainty to be less important for inference than the book states. Our suggestion is that (with one exception noted next) most of the uncertainty in an abundance estimate is captured in the conditional inference based on a point estimate of variance; allowance for hyperparameter uncertainty is desirable, but lack of if it is not automatically a reason to reject an abundance estimate (or its CV).

One exception could be if the estimated surface variance is zero, i.e. infinite smoothing parameter, *and* substantial extrapolation is required. If e.g. a Duchon (1, 0.5) spline is being used, this would mean that the fitted surface is completely flat. Nevertheless, provided that no substantial extrapolation is needed, then fitted models with zero surface variance might still give perfectly reasonable CVs on abundance[7] without needing allowance for hyperparameter uncertainty. Assuming that the model allows for overdispersion or local-clustering, and that the overdispersion estimate is non-zero, then the fitted model is simply "moving" the spatial variability into small-scale rather than large-scale processes. If there are simply too few sightings to yield non-zero estimates of either surface variance or overdispersion/local clustering, then the estimated CV will likely be very high anyway, and it does not matter whether extra components of variability are ingored. If there are plenty of sightings, well-chosen snippet sizes, zero estimated surface variability, and zero estimated overdispersion/local-clustering, then we recommend checking whether the analysis has accidentally used golf tee data instead of real cetacean sightings.

If estimated surface variance is zero and substantial extrapolation is required, then ignoring hyperparameter uncertainty is likely to substantially underestimate the true uncertainty; even if animal density is genuinely quite constant in a small (surveyed) region, it does not follow that it will still be flat in a larger (prediction/extrapolation) region. (In contrast, when the estimated surface variance is non-zero, extrapolation is less likely to give a misleadingly low CV, because the model "knows" that abundance can vary.) The problem is *how* to allow for hyperparameter uncertainty when the point estimate is zero. The distributional theory underlying Empirical Bayes and Bayes Empirical Bayes breaks down because the marginal likelihood (over the variance parameter) is not flat at the maximum. While full-Bayes methods do not rely on such approximations, their CVs may be sensitive to the choice of hyperprior (whereas if there is enough data to estimate a non-zero surface variance, then presumably data is dominating the hyperprior).

To summarize:

- when surface variance is estimated non-zero, hyperparameter uncertainty is probably not the main contributor to overall uncertainty; it is desirable but not essential to allow for hyperparameter uncertainty, approximately or otherwise.
- if surface variance is estimated at zero, and minimal extrapolation is involved, then it is not obvious how to robustly allow for hyperparameter uncertainty— but the CV may nevertheless be reasonable even without such allowance.
- if surface variance is estimated at zero and appreciable extrapolation is desired, estimates and CVs from a spatial model may be untrustworthy.

[7] Of course, the true underlying animal density may truly be quite flat, so it is not biologically implausible.

## 8.3 Analysing several surveys simultaneously

If there are several surveys that use the same protocols and platforms for the same species, in separate regions in the same year and/or the same region in separate years, then there are clear statistical gains to be made by analysing all the DS data together— provided that the DS analyses include all covariates that substantially affect detection probability and might also vary substantially between years/regions, to avoid (one version of) "unmodelled heterogeneity", and that suitably disaggregated diagnostics are checked. The gains apply not just via a bias-variance trade-off that improves the Root-Mean-Square-Error of the resulting abundance *point* estimates, but also to making the estimated CVs more stable. Having stable CV estimates is particularly important for management procedures such as RMP, which can be somewhat sensitive to noise in the estimated CVs.

The same idea, that is statistically desirable to analyse several datasets within the same model when it makes sense to do so, applies to surface variance and overdispersion parameters. These depend on biology, not on the observation process, so there is reason to expect consistency even if there are changes in detection probability and coverage e.g. from changing survey platforms, protocols, or observers. Note that this is *not* the same as assuming that density surfaces have the same shape from year to year; the assumption is just that there is consistency in the scales of clustering and rates of change of log density with location. This certainly seems like a reasonable assumption for surveys of the same place in different years, though it may not be so true across different regions. If one survey is conducted near a complicated coastline while another is mid-oceanic, then the true surface variances might be quite different (and so should be estimated separately), and it may even be best to use different types of smoother.

If the right software is used, it is not difficult nowadays to simultaneously fit several spatial models that must have the same variance parameters; with `mgcv::gam()`, for example, one need only add "by" and "id" arguments to the specification of the smooth term. We strongly advocate doing so, because the resulting CV estimates are liable to be more stable, and less prone to hyperparameter uncertainty.

## 8.4 Variance propagation

Classical Distance Sampling, with equal coverage probabilities and no covariates affecting detection except perpendicular distance, breaks up an abundance estimate into three parts involving the encounter rate, the estimated mean school size, and the effective strip width. In fairly standard notation (e.g. Buckland et al., 2001a), this becomes

$$\hat{N} = \frac{n}{L} \times \hat{\bar{s}} \times \frac{A}{2\hat{\mu}}$$

The three terms are statistically independent, so the CV of the final estimate can be approximated by

$$\mathrm{CV}\left[\hat{N}\right] = \mathrm{CV}\left[n\right] + \mathrm{CV}\left[\hat{\bar{s}}\right] + \mathrm{CV}\left[\hat{\mu}\right]$$

The CVs of the three components are estimated by some combination of model-based (at least for $\hat{\mu}$) and design-based (at least for $n$) arguments. Unfortunately, this separate-estimation approach does not work for spatial models (nor for more complicated design-based settings where covariates are used) because detection probabilities vary spatially and so are inextricably linked with the estimation of "$n/L$".

The formulation $\log \mathbb{E}[y_i] = \log \hat{w}_i + s(x_i)$ that underlies spatial modelling, whether fitted by GAM software or any other approach such as MCMC, needs to account for uncertainty in the offset, i.e. in the estimated detection probability $\hat{w}_i$ (something which also needs to be accounted for in more sophisticated covariate-based design-based settings that use Horvitz-Thompson-Like estimators). This is not done automatically by GAM software. There are three options in current use:

- Fully Bayesian analysis that *simultaneously* fits the detection functions and the spatial model
- Bootstrap: resample all the data, and for each resampled dataset first fit the detection function, then fit the spatial model conditional on the (re-)estimated detection probabilities, then form the point estimate of abundance
- First fit the detection probability, then fit the spatial model as a GAM conditional on the point estimates of the detection probabilities but with augmented extra .

The simultaneous-Bayesian approach is implemented in the R package `DSpat` ([8]). It is hard to argue against philosophically, but in practice the computational machinery is unavoidably cumbersome, and we are skeptical about whether it is really feasible to conduct the detailed exploration of models and diagnostics that is often necessary to get satisfactory detection probability estimates from Distance-Sampling[9].

### 8.4.1 Nonparametric bootstraps

Nonparametric bootstrapping can be a reasonable way to assess the overall CV in a design-based Distance-sampling setting. Even then, though, there is an important choice which has practical consequences and which is not easy to make: what should the resampling unit be? The obvious answer is to say "the transect" but this may not be appropriate[10], and it is not necessarily to find a good solution:

- If there are few resampling units, (the bootstrap estimate of variance is unstable (i.e. would be highly variable across repeated surveys)
- Small resampling units may not be statistically independent, leading to negatively biassed CVs. For example, this would be true if there could be mass movements of animals over distances comparable to two sampling units during the course of the survey, so that neighbouring units would have correlated encounter rates (i.e. "residual" encounter rates assuming a constant density across the region).

The same question of resampling unit applies when trying to bootstrap a spatial model, and the computational burden is also likely to be much heavier especially if variance parameters are re-estimated each time. More seriously, though, there are two additional theoretical problems with important practical consequences:

- Support
- Failure to capture "random effect variance"

[8] `DSpat` currently has limitations for density surface modelling, concerning detection function shape, certain detection on the track-line, no handling of school size

[9] Especially across the broader class of "Distance-Sampling" models that include MRDS, MCDS, and cue-based methods.

[10] In addition to the reasons listed, there is the usual difficulty with design-based variance estimates, of how to separate "systematic trend" from "variability", and choice-of-resampling unit is important for that too. There are sophisticated bootstrap schemes which try to do this, but the fundamental issues mentioned here still apply.

**8.4.1.1 Change of support** In a design-based abundance estimate, it doesn't matter where within the region each transect is located (at least for point estimation purposes), so all bootstrap replicates have about the same "information content". However, spatial models are sensitive to the location of sampling units. If there is only one transect in the corner of the region, then density estimates in that corner will be highly sensitive to how many times the transect appears in a resampled dataset— especially if it appears zero times. It is common to find ludicrous abundances amongst the bootstrap estimates, corresponding to resamples with strange spatial coverage that no sane analyst would dare to use for abundance estimation. It is not obvious that those resamples contain any relevant information for assessing uncertainty conditional on a realized survey design that looks very different. This "change of support" phenomenon is well-known even for situations as simple as 1D linear regression. More sophisticated bootstraps can alleviate it. In particular, in non-random-effect settings (see below) we have had better results from a "Bayesian bootstrap" (Rubin, 1981) whereby each unit in a resample receives a *non-integer* weight drawn from an Exp(1) distribution (whereas a conventional bootstrap effectively assigns weights drawn from a Poisson(1) distribution). With a Bayesian bootstrap, resamples always have the same support as the original dataset, and ludicrous estimates are eliminated. However, Bayesian bootstraps do require that model-fitting can incorporate weights (which is possible in `mgcv::gam()`, for example).

**8.4.1.2 Uncaptured variance of random effects** Bootstraps and random effect models come from very different statistical paradigms, and there is a fundamental sense in which they do not mix. Bootstraps describe the frequentist variability in *point estimates* of some quantity, say density within some sub-region; random-effects models (being Bayesian) should describe the entire posterior uncertainty about that quantity. The following example is perhaps extreme, but at least should be clear.

Consider a "spatial model" where the whole region is divided into 3 blocks; density is assumed to be constant within each block, but to vary across blocks as independent random effects with the "surface variance". Now suppose there is substantial survey effort and lots of sightings in the 1st and 2nd blocks, but not the 3rd. In reality, the analyst will have a good idea of abundance in 2/3 of the region, but very little idea about abundance in the remaining 1/3, so that the overall CV should be substantial. This is the inference that would be reached by analysing the data with a GAM (this is a special case of a GAM with an unusual "smoother"), ignoring uncertainty in detection probabilities. However, with a nonparametric (or parametric) bootstrap, every resample will continue to have zero effort in the 3rd block, so the point estimate of the random effect in that block will remain at zero (relative to the mean of the others) and the variability of the total abundance estimate across bootstraps will be very low.

The example is extreme, but the same phenomenon applies in a harder-to-spot way to any spatial model even when data is merely sparse in some areas rather than completely absent. Although there are some suggestions in the statistical literature for how to deal with this problem, to the best of our knowledge they are specific to "blocked" situations like the example above, but not like real spatial models where there are long-range correlations. We are not aware of any general solution to this fundamental problem, and it is not obvious how to guess the extent of bias for any particular real data set.

### 8.4.2 Parametric bootstraps

Provided that the statistical framework is capable of modelling both large-scale density variations (via the "spatial model") and fine-scale clustering (in this case, preferably via e.g. an MMPP model, or perhaps via an overdispersion parameter), then a parametric bootstrap that simulates spatial variability along the trackline,

and the probabilistic process detection, should automatically solve the choice-of-resampling units and change-of-support problems. However, parametric bootstraps are still fundamentally incompatible with random-effect models because of the uncaptured-variance problem (section 8.4.1.2). Note that this is not a problem for the Norwegian minke whale abundance estimates, which use parametric bootstraps and a fine-scale clustering process. The reason is that the stratum densities are *fixed effects* not random effects, so they are not shrunk to the mean in the way that everything is when a smoother is used.

The `dsm` package includes a "parametric bootstrap option", but it does not work in the way described because there is no explicit fine-scale clustering process. Instead, clustering is handled by a moving-block (nonparametric) bootstrap. The manual warns that ludicrously large bootstrap abundance estimates— "rogue replicates"— are prone to occurring (as with the change-of-support problem for a fully nonparametric bootstrap, but presumably for a different reason), and these "outliers" have to be removed by elaborate but *ad hoc* algorithms. Also, at least as of version 2.2.3, the `dsm` parametric bootstrap cannot properly handle covariates in the detection function. We do not recommend it.

### 8.4.3 Approximate analytical variance propagation

Provided that the spatial model is embedded in a basis-and-penalty framework (regardless of whether `mgcv::gam()` or ADMB or a fully Bayesian MCMC algorithm is used to fit it), then it is possible to approximately propagate the uncertainty arising from estimated detection probabilities without any changes to the basis-and-penalty computational framework, by adding random effects to represent the detection parameter uncertainty. (This is probably only practical if hyperparameter uncertainty is being ignored.) The first stage is to estimate all the detection function parameters, and to compute the estimated detection probability for each snippet. The second stage is to fit the spatial model as a GAM, with the log-offset still computed from the point estimates of the detection probabilities, but with additional columns in the design matrix of random effects. For the $i^{\text{th}}$ row (snippet), the additional columns contain the derivative of the log-detection-probability for that snippet with respect to the entire vector of detection-probability parameters. The corresponding random effects are deviations of the DS parameters away from their first-stage point estimates, and there is an associated penalty matrix (with *known* rather than estimated smoothing parameter) which is the inverse variance matrix from the first-stage estimate. Thereafter, estimation and inference can be done as usual for any GAM.

This is really an extension of the delta-method. We first developed it for the SPLINTR spatial model of Antarctic minke whales, and have found it to be simple and effective in several applications since (see **Williams2011blue** for one example, albeit with a typo; a paper explaining several variants of the approach is in preparation). It is easy to automate, and has now been incorporated into the `dsm` package (v 2.2.3; Miller et al., 2013).

### 8.4.4 Using short-distance encounter rates to help estimate detection probabilities

A pitfall with two-stage fitting is that the estimated detection probabilities can end up inconsistent with the empirical ratio of encounter rates across covariates. For example, there may be clearly "too many" sightings in Good weather conditions compared to Bad, even after allowing for the (higher) estimated detection probabilities in Good weather; this is an important diagnostic to check (section 10). Although it is tempting to use all this information when estimating the detection probabilities, that is not safe practice because there can be systematic variations in weather across large spatial scales which become aliased with changes in animal density; also, it would mean "using the data twice" under the approach of section 8.4.3. However, a compromise is possible. Spatial models (at least the basis-and-penalty type found in GAMs) deal only with large-scale variation;

small-scale clustering is treated as noise. By comparing encounter rates either side of each change in sighting covariates, over short enough distances that the large-scale density surface will not change much, it is possible to substantially improve estimates of detection probability using only information that is otherwise discarded by the spatial model. We first developed this approach for SPLINTR, after finding unpleasant observed/expected diagnostics without it (section 10). It greatly improved the plausibility of the detection probability estimates without worsening the conventional goodness-of-fit diagnostics for Distance sampling. It was computationally simple, but is not currently part of standard software; it requires adding an extra term to the detection-function likelihood, and also requires that all detection-probability parameters be fitted simultaneously.

**Summary** Spatial abundance estimates are subject to uncertainty from estimating detection probabilities, as well as from fitting the spatial model itself with detection probability as an offset. It is important to include both sources of uncertainty. A theoretically ideal way is to use a fully Bayesian framework that fits both parts of the model simultaneously; however, we suspect this may be too cumbersome for many applications. Two-stage fitting is simpler, and allows the usual thorough investigation of Distance-Sampling models. Estimated detection probabilities under different covariates should be validated against empirical encounter rates (section 10) and if there is some suggestion of inconsistency, then a comparison of short-range encounter rates may fix the problem.

Although bootstraps have often been used, they are *not* reliable for spatial models. Even if the change-of-support problem is addressed by using a Bayesian bootstrap or a parametric bootstrap, there remains a fundamental negative bias in bootstrap CVs for spatial models.

Instead, we recommend the use of analytical approximations such as that now found in `dsm`.

# 9 School size

When a substantial proportion of animals occur in groups/pods/schools, several complications arise for Distance-sampling, arising partly from the sometimes difficult task of estimating school size accurately and the implications for protocols, and partly from the extra statistical tasks even when school size is accurately estimated. Even when the analyst may just be aiming for a simple estimate of the form $\hat{N} = (n/L) \times \hat{\bar{s}} \times \hat{\mu}$, it is not straightforward to get unbiassed and mututally coherent estimates of "mean school size" and "effective strip width for schools".

We first note that the spatial density surface can be set up to describe either animal density (e.g. Cooke's IM, where the response variable is number of animals seen per snippet) or school density (e.g. SPLINTR, where the response variable is number of schools seen). When the response describes school density, the overall abundance estimate is slightly more complicated, because of the need to first multiply predicted mean school size *at each point in the prediction grid* by the corresponding predicted school density before adding up the predictions. In our view, the school-density version is mildly preferable on statistical grounds, since the "error model" for the response is liable to be less skewed, and less modelling of "overdispersion" may be needed to get the overall spatial model working well. Usually, though, we would expect similar abundance estimates from either approach. One exception might be if a very small number of very large schools are seen. That is something of a nightmare scenario for animal abundance estimation, and the arguably good news is that it presumably will lead to a high CV under any good statistical model, so that the details of differences may not matter. (The bad news is that finding a few huge schools in one year begs the question: might there have been

similarly huge schools in other years that were not seen due only to chance? This would have implications for the reliability of CV estimates in those other years.)

Detection probabilities can be estimated in the standard way, usually with school size as a covariate. The next step is to compute the overall detection probability for each snippet, which needs to be averaged over school sizes (regardless of whether the density surface describes animals or schools, which merely affects how the average is weighted). To do the averaging, it is necessary to "know" not just the mean, but the entire frequency distribution of school size, either via a parametric distribution (e.g. Negative Binomial, as in the OK model) or a grouped nonparametric distributions (e.g. polytomous/multi-category logistic regression, as in the SPLINTR). Nonparametric distributions are more flexible, but a parametric distribution may be adequate; however, if a parametric form is assumed then the predicted and observed school size distributions should be compared, as a diagnostic check.

If density varies spatially, then it is plausible that mean school size might vary too. Unless there is strong statistical evidence *against* spatial variation in mean school size, we recommend that some kind of spatial model be used to allow variations in mean school size; once the local mean is available, the entire frequency distribution can then be determined according to the parametric/nonparametric distributional assumption discussed in the previous paragraph. For the "spatial model" of mean school size, the observed school size at each sighting is used as the response, with probability proportional to the true frequency distribuation at that location and school size multiplied by the estimated detection probability by that school size (as an "offset"). At a minimum, a "stratified model" (constant within block, variable between block) should be tried, as in OK and Cooke's IM; if the blocks are chosen in a biologically arbitrary way, say to coincide with predefined strata, then a sensitivity check to block size should be used. However, it is also possible to use a second smooth spatial model to describe mean school size, as in SPLINTR.

Variance propagation can be done as described in section 8.4.3. The extension to 3 stages of modelling (first detection probabilities, then spatial modelling of mean school size, then spatial density of schools/whales) does not cause any conceptual problem, although the computation can be fiddly and the complete formulae for the school size case are not yet published.

## 9.1 School size measured with error

School size error is a tough problem in Distance-sampling, whether spatial models are being used or not. Well-designed data collection protocols are required to resolve it satisfactorily, e.g. to avoid confounding between the perpendicular detection function (which presumably is affected by true school size) and the probability of school size error.

When school size is measured with error, all the above remarks about spatial modelling and school size continue to apply, with an additional complication. It seems inevitable that detection function parameters must now be estimated *simultaneously* with the frequency distribution of school size[11]; because the true school size will not be known for many sightings, and the posterior distribution of each true school size depends not just on the observed size but also on the "local prior", i.e. on the frequency distribtion of school size in the vicinity. This simultaneous estimation is statistically possible given adequate data, but is outside the realm of off-the-shelf statistical models and requires specially-written software in e.g. ADMB: for SOWER Antarctic minke whale sightings, SPLINTR tackled it using a smooth spatial model for mean school size, and OK tackled it using

[11] Unless *observed* school size does not affect empirical detection functions.

block-stratified mean school sizes. Propagating the variance from this step through to the final abundance is not much worse than for the no-school-size error case.

# 10 Diagnostics

All the diagnostics associated with fitting detection functions are still applicable when the detection probabilities are to be used in a spatial model.

With respect to the spatial model itself, there is an extensive literature on diagnostics for GLMs, and GAMs can be viewed as a natural extension. For the statistical and computational aspects of basis-and-penalty smoothers (using enough knots, etc.), we cannot improve on the help page for `?gam.check` in the `mgcv` package (regardless of whether the model is fitted with `mgcv::gam()` or with other software). The comment in section 8.1 is apposite: if snippets are very small, so that very few expected encounters are above 1 say, then there is not much chance of detecting problems using unaggregated observations at the snippet level (in the same way that residual-based diagnostics for binary data do not work unless the residuals are first aggregated in a question-specific way). The most obvious diagnostic is, of course, an image or contour plot of the density surface, to check there are no hotspots of ridiculously high predicted density.

We do not know of any formal checks for "excessive extrapolation", but the presence of hotspots (even if not excessive) near the edge may be a warning signal (though not necessarily a red light). In such cases it would be useful to compute how much of the total abundance estimate comes from the areas that look "most extrapolated", and perhaps to compare the overall CV with the CV a prediction that avoids those areas; there should be a much higher

If the spatial model includes overdispersion via a Tweedie or Negative Binomial distribution (which it should, unless a local-clustering process such as MMPP is included), then it is usual to set the "extra" parameter by hand rather than to estimate it explicitly. Hence, it is important to check that the variance-mean relationship is satisfied, via the usual sqrt(abs(resid)) plot.

For cetacean abundance estimation, we have found one additional type of diagnostic to be very useful: comparing observed and expected numbers of sightings at Wvarious categories of disaggregation (the overall totals should match closely, by construction). As per section 8.4.4, one example is by covariates related to sighting conditions (whether these are included as covariates in the detection function or not). Another is to look for anomalies in the density surface: are there too many/few sightings close to a coastline, for example? When applying SPLINTR to the 2004 SOWER survey of the Ross Sea, we found a big discrepancy between observed and expected sightings in one part of the region; on inspection, we realized that the discrepancy corresponded to a major shift in the ice edge part way through the survey, which led us to a much better fit using two separate spatial surfaces[12] instead of one.

When a substantial proportion of animals are found in school, then checks of observed versus expected frequencies of school sizes should be made, stratified by sighting conditions and/or location if sample sizes permit. Contour/image plots of mean school size are also useful.

With respect to simulation-based diagnostics: these can be extremely useful in checking the performance of estimators that have clear theoretical deficiencies but might be practically adequate for some datasets. Simulation is particularly useful for some aspects of abundance estimation, such as estimating detection probabilities.

[12] With shared smoothing parameters

Its utility may be less spatial models, because there may not be many many serious theoretical deficiencies left. GAMs have been around for 20 years (Breslow and Clayton, 1993), with improvements in both inferential theory and computational practice to the point where they are a routine tool of statistical analysis; we do not think it is necessary nowadays to check every spatial model *by simulation* to see if the smoother is working (though it is necessary to check diagnostics). In certain cases— heavy extrapolation comes to mind— specific simulations could be useful, but care is needed to choose the prior distribution of density surfaces in a "reasonable" way that does not make it too hard or too easy for the estimator (since each type of smoother implicitly has its own type of prior). Simulation could be very helpful in cases where the spatial model fits well but its implicit assumptions are violated in a "trans-statistical" way, e.g. when there is migration during the survey.

# 11 Conclusions and suggestions for text

In some cases, it is clear that a design-based analysis (including the associated CV estimates) is perfectly reasonable (although a spatial model could still be useful to stabilize CVs across years). In other cases, a spatial model is clearly the only reasonable approach, when it is clear that coverage is quite unbalanced. Sometimes, though, a survey may fail to qualify as "suitable for design-based analysis" under the criteria in Part 1. Nevertheless, if the spatial coverage is quite even, it might be reasonable to expect that *calculation* using design-based formulae will give a similar point estimate to a spatial model. And— perhaps less likely— if there is not much apparent trend and no major hotspots in encounter rate, then it might even be reasonable to expect that a design-based calculation will give a similar CV to a spatial approach. Under those conditions, the Committee *might* be tempted to just accept a "design-based estimate" rather than asking for a more complicated model-based estimate, even though the latter is more appropriate, because the answers would probably be similar. However, our own sympathy for that position is considerably tempered by the following:

- for statisticians using modern software, many spatial models are simply GAMs and are really are not hard to fit[13];
- thanks to improved estimation algorithms and better-founded smoothers, spatial abundance estimates are better behaved than in the early days;
- even when coverage is balanced enough for bias[14] not to be of concern, spatial models provide a simple and coherent basis for variance estimation that is robust to systematic trends in density and is not sensitive to small numbers of transects;
- in the context of repeated surveys over time (or space), spatial models are liable to give less variable CVs because the "biological" variance parameters can reasonably be shared (i.e. made equal) across years (or areas).

In such a case, we consider that the onus should be on the analyst to demonstrate that a design-based *estimate* is acceptable (for that particular case), based either on comparison with a spatial analysis (perhaps not polished to the same degree as if it was the analyst's preferred approach, but still subject to the QC in this document), or on a simulation that shows low bias for the design-based estimate and CV using the *actual* design but with the *real* density surface of whales varying across simulations in a plausible way. This begs the question: what is

[13] Harder cases do exist, e.g. when school size gets complicated; but such cases are perhaps even more likely to benefit from a spatial model.

[14] IE bias of observed mean encounter rate as an estimate of encounter rate across the whole Region

plausible here? One obvious response is: using the posterior distribution obtained from fitting a spatial model. (Conceivably, the prior might be used instead, with the prior variances fixed at estimated values.) Either way, we cannot see a general-purpose way to avoid fitting a spatial model.

## 11.1 Proposals for amended text

Our draft amendments are shown below, for two sections of the Guidelines: 3.3 and a new section 6.3. Amendments are in italic (either for entire paragraphs or parts thereof), with comments in [square brackets]. The Committee may also want to give some attention to 6.1 (variance estimation for design-based analyses), for which some of the material in this report may be useful. We also note that the comments about migration in current section 3.3 are rather terse, and are not reflected in current section 6 (Analysis) but perhaps should be.

### 11.1.1 3.3 Cruise tracks

The first stage is to define the area that is to be surveyed and to which the resultant estimates will apply. In many cases, the most efficient way to survey an area is to stratify it. The shape and size of strata will be determined by physical factors such as the surrounding land masses and limitations on the endurance of the survey platforms. If prior knowledge of the distribution and relative abundance of whales is available, this should be used when delimiting strata. If qualitative or quantitative information on the relative abundance of animals is available, more effort should be devoted to strata of known high abundance e.g. see Fig. 1.

*When designing cruise tracks, the type of analysis intended to be conducted (design-based or model-based) should be known in advance. This is essential if a design-based analysis will form the basis of an estimate, as the randomization and replication of the transects within the survey area then forms a critical component of the analysis. Inference from a model-based analysis does not rely on a randomization scheme (it relies on the appropriateness and goodness-of-fit of the model), but cruise track placement is still important. The same general points about tracklines apply as for design-based analyses, but the emphasis is no longer on randomness; for model-based analyses, tracklines should cover the range of the spatial covariates and be quite 'evenly' spread across the survey area. Gaps in realized transect coverage (eg. due to poor weather) are more acceptable than in design-based studies, and should allow the survey to cover the wider range of spatial covariates overall that is required for a reliable fit.*

*When a design-based analysis is proposed, surveys should be designed so that the coverage probability in each stratum is uniform or can otherwise be determined. Where inference from an analysis relies on (nearly) uniform coverage probability, then the randomization scheme that forms part of the design should be clearly described for the specific survey in question, taking into account the shape of the area to be surveyed and any other salient features. It is useful, and in some cases necessary, to include a map of estimated coverage probability, to provide evidence that the survey design meets sampling assumptions. Program Distance (versions 5.0 and later; Thomas et al., 2010) provides coverage probability maps for some designs, but other trackline designs are of course acceptable if coverage probability can be shown to be uniform. Where a design-based analysis explicitly uses unequal coverage probabilities, then a description of how they were calculated must be included. If coverage probability is not uniform, then even if coverage probability is estimated, the design should minimise the amount of survey area with very low relative coverage probability. Design-based estimates of abundance obtained from new surveys whose design does not meet this criterion will not be accepted for use in the CLA. If a design-based analysis has been used to estimate abundance, but the criterion of equal or estimable coverage probability has not been demonstrated to have been met, then further analytical investigations will be*

*necessary before the Scientific Committee may accept them. Either (i) sensitivity analyses of the results should be conducted against a spatial model, or (ii) some simulation testing be conducted to demonstrate that the design-based estimates and CVs are likely to be unbiased, given the realized design and a plausible range of spatial distributions (and it might entail fitting a spatial model to establish that plausible range). If either (i) or (ii) suggest that the design-based analysis may be unreliable, then the Scientific Committee may recommend that a model-based estimate be employed instead (see section [new] 6.3).*

The aim of a survey should be to maximise coverage of the area within the resources available. However, this should not be done at the expense of the experimental and calibration work necessary to ensure proper analysis of the data (e.g. estimation of g(0), distance estimation; see Section 4). When determining the length of cruise tracks, due consideration should be given to time that may be expected to be lost as a result of bad weather.

When considering the placement of cruise tracks, care should be taken that they do not follow physical features that may be correlated with whale abundance. For example, cruise tracks should not run parallel to the coastline or to depth contours in the vicinity of shelf breaks but should rather run across such features. Although the 'classical' approach is to place a random grid of evenly spaced parallel lines within each stratum, this can be inefficient if line separation is large, particularly for shipboard surveys. A commonly accepted alternative is to use regular zig-zag lines with a randomly chosen start point. Fig. 1 provides examples of different types of cruise track design. *[Deleted reference to Distance here since it now occurs earlier on.]* If more than one stratum is to be surveyed by different platforms operating at the same time, consideration should be given to the timing of the surveys in order to minimise the difference in time between surveying the area on one side of a stratum boundary and the adjacent area on the other side of the boundary.

If there is a known or suspected migration of whales through the survey area, care should be taken when designing cruise tracks and survey direction to ensure that the data collected are representative. For example, it is clearly inappropriate to survey following the direction of the migration, particularly with a slow moving platform. *[We suggest this paragraph could do with an update, but it is outside the scope of our report.]*

## 11.2 6.3 Spatial models for abundance estimation

*Spatial models can be used as a rescue option for analysing that surveys that fail to meet design-based criteria, or that had to deviate from plans for logistic reasons. However, they can also be used deliberately on any survey dataset (provided of course the general criteria in the Guidelines are met). They offer several potential advantages: no bias when dealing with unbalanced coverage, no need for coverage probability estimates, no assumption of "pooling robustness"*[15]*, no need to assume independence between weather/school size/school density, no confounding of systematic trend with noise when computing CVs, greater stability in CVs from year to year (if variance-related parameters, at least, are shared across years). The statistical tools for fitting spatial models, based on GAMs and other implementations of spatial random field models, have become much more flexible and reliable in recent years, and it is possible to marry essentially any method of estimating detection-probability with any of a broad class of spatial models. The recent experience of the Committee in simulation-testing of spatial methods has been encouraging [here we mean SPLINTR, where the spatial aspects seemed to work well]*

*Spatial models are powerful, but they can give poor results if used inappropriately. In each application, there are a number of analysis choices that need to made, based on both theoretical considerations and empirical aspects of each dataset, as in (ref** long text: coordinate system, type of smoother, extent of extrapolation,*

[15] *Pooling robustness is a valid assumption in some but not all Distance-sampling applications*

*snippet size and fine-scale clustering, smoothing parameter estimation, treatment of school size var, computation of overall variance). It is not possible to give in advance a definite recommendation about each choice that would cover all situations addressed by the Committee. Therefore, it is critical that any model-based estimate be accompanied by a clear explanation of why each particular choice was made. The reason might be theoretical or empirical, or even arbitrary; in some cases, a choice will be somewhat arbitrary and will be made simply for reasons of computational or other convenience. That need not be a problem provided that there is either an explanation of why it is unlikely to matter for inference, or a sensitivity check.*

*A variety of diagnostics for spatial models are also required, beyond those normally required for Distance sampling (ref** long text). Of greatest importance are contour maps of predicted densities (and associated summaries, e.g. of proportion of total abundance in dubious-looking predicted hotspots) to look for extrapolation problems, and comparisons of observed/expected sightings over different covariates and subregions. Both the accompanying explanations, and the diagnostics, are essential prerequisites for scientific review by the Committee.*

*Multi-year estimates will have shared parameters, and thus will be correlated although, if it is only variance-related parameters that are shared (as opposed to detection function parameters), the correlation is likely to be weak. Correlated multi-year abundance estimates are already allowed for in CLA (Punt,* pers.comm.*).*

## Appendix: notes on various smoothers

**Soap** Unlike most isotropic smoothers, Soap has *two* variance parameters (for the internal and boundary smooths) to estimate, and requires that the prediction and data region are enclosed by a closed boundary. The theory of Soap assumes that the entire boundary is real and physical, but the computation runs even with an artificial boundary, although the theoretical justification is then unclear. Some care is needed with boundary knot placement especially when some parts are real and some artificial— it seems generally advisable to make the boundary itself fairly smooth, and to place the boundary knots without cramming too many of them into short wiggly sections. To get the most powerful smoother taming, the boundary-smooth should be set to `xt=list(bndSpec=list(bs='cp',m=c(3,1)))`. A zero variance estimate for the boundary smooth is not unreasonable, but a zero estimated variance for the internal smooth would warrant attention in the sense of section 8.2.1.

**Thin-Plate Splines** we would not recommend for animal density applications, as per section 7.2.2. Technically, the nullspace includes linear trends which, after exponentiation to the response scale, can easily lead to wild extrapolation.

**Duchon splines** are now available in `mgcv` and can be used as a plug-in replacement for TPS (which are in fact a special case); see `?mgcv::smooth.construct.ds.smooth.spec` for explanation, or Duchon (1977) for the brave. For good smoother taming behaviour, set `m=c(1,0.5)`.

**Separate 1D smoothers** e.g. an `mgcv` formula such as `s(lat)+s(long)`: we would not usually recommend this as a starting point because it does not allow for the possibility of interaction; a tensor product spline can assess interaction automatically. If interaction can be ruled out as unlikely (because of the nature of the coordinate system chosen and the animals' biology) or seems not to be important in the tensor product, then separate smoothers can be acceptable.

**Tensor product splines** e.g. an `mgcv` formula such as `te( lat, long, <type.of.smoother>)` (or the statistically equivalent but more cumbersome `ti(lat,...)+ti(long,...)+ti(lat,long,...)`, which can be useful in checking for the importance of interactions): these are very useful for non-isotropic smoothers, but require more surface variance parameters to be estimated. Be careful of the extrapolation behaviour; we are not sure how far it is possible to "tame" the underlying 1D components of a 2D tensor product spline.

**Space time smoothers** can be implemented easily using tensor products. They are an attractive possibility for handling large-scale animal movements *within* the Region during the course of a survey (which presumably covers parts of the Region more than once), but they cannot cope automatically with constant migration through the region. However, there is limited experience with their performance.

**Kriging** is in our opinion is not *per se* a good tool for spatial abundance estimation of animals. It is usually presented as a stand-alone procedure, rather than a smooth component of a GAM. It does not adapt well to low-expected-value count data in the presence of overall trend, e.g. cetacean sighting data The "estimation" of the correlation parameters for kriging is also notoriously subjective. Kriging does have one theoretical advantage, in that it can flexibly vary the spatial correlations over different scales; most other smoothers have just one surface variance parameter, so that small-scale correlation is fixed relative to large-scale correlation. In our experience, cetacean sightings are usually too sparse to warrant such niceties (though see the comments on clustering section 8.1). If such flexibility is required, it would be preferable to stay with the GAM-like framework but use a more modern smoother formulation, such as the Gauss-Markov random fields used by INLA.

# References


Barabesi L and Fattorini L (2013). "Random versus stratified location of transects or points in distance sampling: theoretical results and practical considerations". *Environmental and Ecological Statistics* 20, pp. 215–236.

Beekmans B, Forcada J, Murphy E, De Baar H, Bathmann U, and Fleming A (2010). "Generalised additive models to investigate environmental drivers of Antarctic minke whale (Balaenoptera bonaerensis) spatial density in austral summer". *J. Cetacean Res. Manage.* 11, pp. 115–129.

Bravington M and Hedley S (2012). *Abundance estimates of Antarctic minke whales from the IWC IDCR/SOWER surveys, 1986-2002.* Scientific Committee paper SC/64/IA13. International Whaling Commission.

Breslow NE and Clayton DG (1993). "Approximate inference in generalized linear mixed models". *Journal of the American Statistical Association* 88.421, pp. 9–25.

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, and Thomas L (2001a). *Introduction to Distance Sampling*. Oxford University Press, Oxford. 448pp.

Buckland S, Anderson D, Burnham K, Borchers D, and Thomas L (2001b). *Introduction to distance sampling*. Oxford University Press.

Cohen J (1994). "The earth is round ($p < 0.05$)". *American psychologist* 49.12, p. 997.

Duchon J (1977). "Splines minimizing rotation-invariant semi-norms in Sobolev spaces". English. In: *Constructive Theory of Functions of Several Variables*. Ed. by W Schempp and K Zeller. Vol. 571. Lecture Notes in Mathematics. Springer Berlin Heidelberg, pp. 85–100.

Fewster R (2011). "Variance estimation for systematic designs in spatial surveys". *Biometrics* 67, pp. 1518–1531.

Fewster R, Buckland S, Burnham K, Borchers D, Jupp P, Laake J, and Thomas L (2009). "Estimating the encounter rate variance in Distance Sampling". *Biometrics* 65, pp. 225–236.

Hakamada T and Matsuoka K (2012). *Preliminary analysis of abundance estimate for sei whale in the North Pacific based on sighting data obtained during IWC-POWER surveys in 2010-2012.* Scientific Committee paper SC/65b/IA04. International Whaling Commission.

IWC (2012). "Requirements and Guidelines for Conducting Surveys and Analysing Data within the Revised Management Scheme". *J. Cetacean Res. Manage. (Suppl.)* 13, pp. 509–517.

– (2014). *Report of the Planning Meeting for the 2014 IWC-POWER Cruise*. Scientific Committee paper SC/65b/Rep01. International Whaling Commission.

Johnson DS, Laake JL, and Ver Hoef JM (2010). "A Model-Based Approach for Making Ecological Inference from Distance Sampling Data". *Biometrics* 66.1, pp. 310–318.

Martino S and Rue H (2009). "Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the INLA program". *Department of Mathematical Sciences, NTNU, Norway*.

Miller DL, Burt ML, Rexstad EA, and Thomas L (2013). "Spatial models for distance sampling data: recent developments and future directions". *Methods in Ecology and Evolution* 4, pp. 1001–1010.

Miller D and Wood S (2014). "Finite area smoothing with generalized distance splines". *Environmental and Ecological Statistics* online.

Peel D, Bravington M, Kelly N, Wood SN, and Knuckey I (2013). "A Model-Based Approach to Designing a Fishery-Independent Survey". *Journal of Agricultural, Biological, and Environmental Statistics* 18.1, pp. 1–21.

Rubin DB (1981). "The bayesian bootstrap". *The annals of statistics*, pp. 130–134.

Rue H and Martino S (2007). "Approximate Bayesian inference for hierarchical Gaussian Markov random field models". *Journal of statistical planning and inference* 137.10, pp. 3177–3192.

Scott-Hayward LAS, MacKenzie ML, Donovan CR, Walker C, and Ashe E (2013). "Complex region spatial smoother (CReSS)". *Journal of Computational and Graphical Statistics* just-accepted.

Skaug HJ (2006). "Markov modulated Poisson processes for clustered line transect data". *Environmental and Ecological Statistics* 13, pp. 199–211.

Skaug HJ and Fournier D (2006). "Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models". *Computational Statistics and Data Analysis* 51, pp. 699–709.

Strindberg S and Buckland S (2004). "Zigzag survey designs in line transect sampling". *Journal of Agricultural, Biological and environmental Statistics* 9, pp. 443–461.

Thomas L, Buckland S, Rexstad E, Laake J, Strindberg S, Hedley S, Bishop J, Marques T, and Burnham K (2010). "Distance software: design and analysis of distance sampling surveys for estimating population size". *Journal of Applied Ecology* 47, pp. 5–14.

Thomas L, Laake J, Rexstad E, Strindberg S, Marques F, Buckland S, Borchers D, Anderson D, Burnham K, Burt M, Hedley S, Pollard J, Bishop J, and Marques T (2009). *Distance 6.0. Release 6.1*. Tech. rep. Research Unit for Wildlife Population Assessment, University of St. Andrews, UK.

Thomas L, Williams R, and Sandilands D (2007). "Designing line transect surveys for complex survey regions". *J. Cetacean Res. Manage* 9, pp. 1–13.

Thompson S (2002). *Sampling*. Wiley Series in Probability and Statistics. Wiley.

Wahba G (1990). *Spline models for observational data*. Vol. 59. Siam.

Wood SN (2006). *Generalized Additive Models: An Introduction in R*. Boca Raton: Chapman and Hall/CRC.

Wood SN, Bravington MV, and Hedley SL (2008). "Soap film smoothing". *J. R. Statist. Soc. B* 70, pp. 931–955.

Wood SN (2011). “Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.1, pp. 3–36.