Shortcomings of Horvitz-Thompson-based estimates of abundance for large-scale cetacean abundance estimation

David L. Miller and Mark V. Bravington

25th July 2016

Integrated Statistics, Woods Hole, MA & Centre for Research into Ecological and Environmental Modelling, St Andrews, Scotland

Commonwealth Scientific and Industrial Research Organization, Hobart, TAS

1 Introduction

A great deal of time and money is spent on large-scale cetacean surveys. Though sophisticated statistical methods have been developed to deal with the complications of complex spatial data, investigators can opt to perform an overly simplistic analysis based on randomisation principles which are not appropriate for the data in question. An example of a potential issue of this type is the use of "vanilla" Horvitz-Thompson estimators of abundance from line transect surveys when the underlying distribution of the study species varies in space. In this paper we show by simulation when such analyses are inappropriate in simple situations and reasoning that if an estimator shows poor properties in a simple situation, then when the distribution, availability or detectability is more complex in nature then the estimator will perform even more poorly.

Here we are interested in two methodologies for estimating abundance from line transect distance sampling surveys, one is a design-based estimate, the other is a model-based estimate. We assume in both cases that the usual assumptions regarding distance sampling surveys have been met (see e.g., Buckland et al 2001, 2004, 2015). We also assume that:

- Distances are recorded for each observation (along with the size of each group of animals, without error)
- The vessel's location along the transect is recorded during the survey (for example using an automated waypoint function on a GPS unit), i.e., the transect lines are recorded as they were visited (the *realised design*), rather than as they were designed.

- A covariate that gives some indication of sighting conditions was also recorded (we simply refer to this as "weather" but it could be Beaufort Sea State or some other omnibus measure of visual conditions).

We now describe the two ways in which one could analyse this data, as used below.

2 Methods

We do not spend time here describing fitting the detection function to the distance data and refer readers to Buckland et al (2001, 2004, 2015) for information on model formulation and selection. Assuming a fitted detection function (which we will denote g, a function of distance, observed detection-related covariates and estimated parameters), we can calculate the average (over distance) probability of detection for an individual (conditional on observed covariate values), $\hat{p}_i$, by integrating out distance.

2.1 Horvitz-Thompson-like estimators

We first describe the "Horvitz-Thompson-like" estimator (henceforth HT) first described in (**blah, CITE**). In its simplest version, the HT estimator is

$$\hat{N}_{HT} = \frac{A}{a} \sum_{i=1}^{n} \frac{s_i}{\hat{p}_i} \tag{1}$$

where n is the number of observations, index by i, s_i is the size of the i^{th} group and $\hat{p}_i$ is the detectability estimated for the i^{th} group, which will be a function of the covariates for that observation. The total area surveyed (*covered area*) is a, which is the product of the line lengths and their corresponding strip widths (if the strips were all the same width then $a = 2wL$, where w is the strip half-width as in Buckland et al 2001, and L is the sum of all the lines' lengths). Finally, A is the area that we wish to estimate abundance for (sometimes referred to as the *study area*). Intuitively, we take the group sizes, correct them for detectability and sum to get an estimate of abundance in the covered area, we then rescale this to the study region. The HT estimator assumes that animal density is constant within the study area. This assumption may be justifiable in some situations, but seems very unlikely in a dynamic environment such as an ocean.

In order to circumvent this shortcoming, we can perform pre or post hoc stratification, slicing-up the study area into smaller subsets and estimating abundance for each of these. These may be geographically defined ("near vs. far from shore", "east/west of some longitude", etc), based on conditions at sea ("dense vs. non-dense ice") or based on oceanographic features ("deep or shallow water"). Mathematically, this consists of changing the limits of the sum (summing over the animals which occur at a given depth or in a particular area etc), then changing a and A accordingly to reflect the effort in the given stratum

and the area of that stratum, respectively; abundance estimates can then be summed to obtain total abundance or given per-stratum form. In some sense these estimates reflect a crude, "blocky" spatial model, which try to address the deeper drivers of distribution in a given species. Stratification can be performed using animal/group-specific covariates (e.g., abundance of males/females, juveniles/adults etc), though we do not address this here.

Variance is estimated for $\hat{N}_{HT}$ by noting that there are two sources of randomness in the equation: *(i)* from the variance in the model for $\hat{p}_i$ and *(ii)* from the randomness in the number of observed groups, n. The variance component from $\hat{p}_i$ can be calculated by using the variance from the detection function estimation procedure and using a sandwich estimator to express that parameter variance in terms of $\hat{p}_i$ (Borchers, Buckland and Zucchhini, 2001 Appendix C). Variance in n is usually calculated as variance in n/L, that is the encounter rate variance. There are a number of options for this estimator, depending on the possible design used. Fewster *et al* (2009) proposes a series of estimators and evaluates them. Here we use the Fewster *et al* (2009) R2 estimator (**CHECK this, IIRC this is the default in mrds**).

2.2 Density surface models

We now describe one spatially explicit approach to modelling distance sampling line transect survey data, which we refer to as *density surface models* (DSMs; **Hedley, Miller etc**). These are similar to the HT estimators above, as they assume that the detection function has already been adequately fitted to the distance data and estimates of probability of detection are available for the spatial model to use. The spatial part of the model uses the generalized additive modelling framework (**wood, carrol etc**) to build smooth, spatially explicit terms describing the distribution of the species and their response to other biological/physical variables (though in this paper we only consider models that include smooths of location). Rather than dealing with whole transects (which are generally long and can include large changes in animal density along their length, as well as covariate values), we cut the transects into smaller pieces, which we call *segments*. The mean response of the model can be written as:

$$\mathbb{E}(n_j) = A_j \hat{p}_j \exp(\beta_0 + \sum_k s_k(z_{kj})),$$

where j indexes the segments, each of which have area A_j and all observations in that segment have a probability of detection of $\hat{p}_j$. The response, n_j, is distributed according to some count distribution for which exp is the inverse link function[1]. The model intercept is β_0. The s_k are usually splines (**some papers**): smooth functions of one or more covariates (denoted z_{jk}), though could be more exotic things like random effects, tensor products, smooth-factor

[1]Though any exponential family response can be used in the GAM framework with any appropriate link function, we just talk about count distributions and log link functions for clarity of notation.

interactions and so forth (**cite Simon doubly general paper**). The exact form of each s_k depends on the nature of the effect we wish to model, for smooths of location there are quite a few options, some of which are enumerated and described below.

A typical DSM may take a form such as:

$$\mathbb{E}(n_j) = A_j \hat{p}_j \exp(\beta_0 + s_{x,y}(x_j, y_j) + s_{\text{Depth}}(\text{Depth}_j)),$$

that is: the expected count in segment j is a function of its location and the water depth at that location, this is then exponentiated onto the response scale, and multiplied by the area of the segment and probability of detection in that segment.

Note that here the probability of detection has a subscript j not i as in the HT estimator. This is because in this formulation of the DSM we only consider detectability as varying at the scale of the segments, not the observed individuals/groups. This means that covariates that effect detectability such as weather, observer shift or ship can be used, but sex of the animal or observer ID (if there were multiple observers on deck at once) cannot be[2].

3 Motivation

A criticism of the DSM approach is that it is more complex, as we explicitly model the spatial and environmental covariate effects, but this explicit modelling is the only way to deal with the heterogeneity in spatial distribution of the study species. We note that an appeal to "pooling robustness" (Buckland 2004, section 11.12) does not get around this issue. Before explaining why, we first define and explain pooling robustness in a distance sampling context. From Burnham et al. (1980):

$$n\hat{f}(0) = \sum_{r=1}^{R} n_r \hat{f}_r(0)$$

where there are R strata chosen to minimise heterogeneity, n is the total number of observations, n_r is the number of observations in stratum j and $\hat{f}(0)$ and $\hat{f}_r(0)$ are the probability density functions of the observed distances, evaluated at zero distance, for the whole sample and by stratum, respectively. Equivalently we can write:

$$\frac{n}{\hat{p}} = \sum_{r=1}^{R} \frac{n_r}{\hat{p}_r}$$

Intuitively, we say that pooling robustness holds, the estimates from a stratified analysis would be the same as those for an unstratified analysis. Buckland

[2]These covariates can be included in a more general formulation of DSMs, though we don't consider them here for clarity of presentation.

et al (2004, section 11.12) state: "if only an overall abundance estimate is required, standard methods without covariates are satisfactory under rather mild conditions, provided heterogeneity in detectability is not too extreme." This is a statement about Horvitz-Thompson estimation in the presence of detection heterogeneity and does not say anything about the case where density varies within strata — in this case the effect of detectability and distribution are confounded, unless data on observation conditions (e.g., a weather covariate) and spatial distribution (e.g., location of transects) is recorded and modelled. A spatial model that includes data on the location of the observations and the sighting conditions will be able to tease apart these effects and attribute appropriate uncertainty.

So far we have only considered the case where detectability is certain on the line ($g(0) = 1$), this situation only gets more complicated once we start thinking about double observer methods. Buckland et al (2004) say: "If $g(0, z)$ is a function of z, then the model robustness criterion fails, and we must model the heterogeneity to avoid bias", so if we do expect that the probability of detection at zero distance is influenced by covariates (which it almost surely will be), pooling robustness does not apply.

Pooling robustness is implicitly conditioned on having a "reasonable" design, so appeals to it should only be made in the case where the realised design has (approximately) even coverage.

Given the above, there is a temptation then to fit a "dumb" spatial model and hope for the best. Properly configuring a spatial model is a time-consuming process requiring some "expert" judgement. As well as formulating, fitting and selecting between models, the investigator also needs to select an appropriate prediction grid, ensuring that unreasonable extrapolations are not made (**maybe cite Mannocci and Conn papers here?**). There is no "quick fix" to obtain a good spatial model, care must be taken in the construction and checking of the model if reasonable inferences are to be drawn.

4 Simulation setup

With the above in mind, we set about constructing some simple simulations of plausible survey data. We attempted to keep the underlying densities as simple as possible and the realised designs as fairly realistic.

4.1 Density surfaces

We used a series of simple density surfaces to test for differences between the proposed models. Although animal distribution is much more complicated than the patterns shown below, if models fail for these simple density surfaces (where gradients are clearly defined) then it's likely that there will be more severe issues when more complex surfaces are used. In the simulations presented here the following surfaces were investigated:

- "f": flat density, uniform distribution across the region.

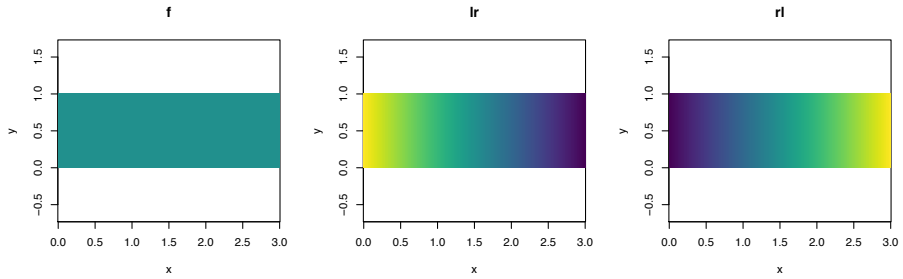

Figure 1: The proposed density surfaces. From left to right: flat, left-right gradient, right-left gradient.

Detection function	Scale	Shape	Truncation
1	0.025	3	0.05
2	0.005	1	0.05

Table 1: Parameters for the detection functions used in the simulations. Plots of the detection functions are shown in Figure 2.

- "lr": left to right gradient, high on the left, decreasing as we go right.
- "rl": right to left gradient, high on the right, decreasing as we go left.

These are shown in Figure 1.

4.2 Detectability

Detectability in the survey was set at two fixed detectabilities "high" and "low" by varying the parameters of a hazard rate detection function.

Plots of the detection functions used to simulate the data are shown in Figure 2 and a table of the parameters of the detection functions is given in Table 1.

4.3 Designs

We experimented with three designs: one bad design with a large gap between each contiguous section of realised effort, one "iffy" design where there are gaps between "chunks" of effort and one good design with good realised coverage across the whole study area (to confirm that what we consider to be a good design gives us the results we expect from our metrics). The designs are shown in Figure 3.

4.3.1 Design 1: zig-zag with straight line

This is supposed to mimic the situation in which a zig-zag design went well on the left side of the study area, but not realised in the middle of the survey

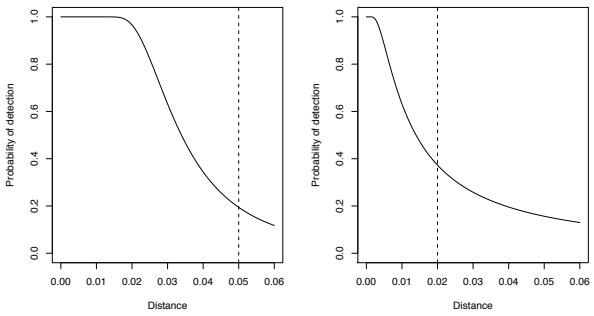

Figure 2: Detection functions used in the simulations. Dashed line indicates the truncation distance used. Left to right have the parameters given top to bottom in Table 1.

(perhaps due to bad weather), then to the left we have a lonely transect (perhaps weather picked-up). Shown in the left panel of Figure 3.

4.3.2 Design 2: IWC

This design illustrates the case where conditions were more amenable than in the first example, but coverage is still quite patchy with some gaps in coverage. Shown in the middle panel of Figure 3.

4.3.3 Design 3: zig-zag with good coverage

The final design has good coverage over the whole study area and would be the ideal realised design. Shown in the right panel of Figure 3.

4.4 Models

Both spatially explicit models and HT methods were used to estimate abundance for each simulation. These are enumerated below. Since we only include spatial terms in our simulations and we believe that in general our spatial effects can be estimated by bivariate smooths (even if in this example the underlying densities are better suited to univariate smooths, we never know this *a priori*). The spatial models are separated into two classes: those which have the isotropy property (that a unit change in one direction is considered to be equivalent to a unit change in an orthogonal direction, sometimes referred to as *rotational invariance*) and those which do not; these are constructed by a *tensor product* of univariate splines. We also test that the setup of the smoother isn't unduly advantageous to a particular model by rotating the coordinates system by $45^{\underline{o}}$

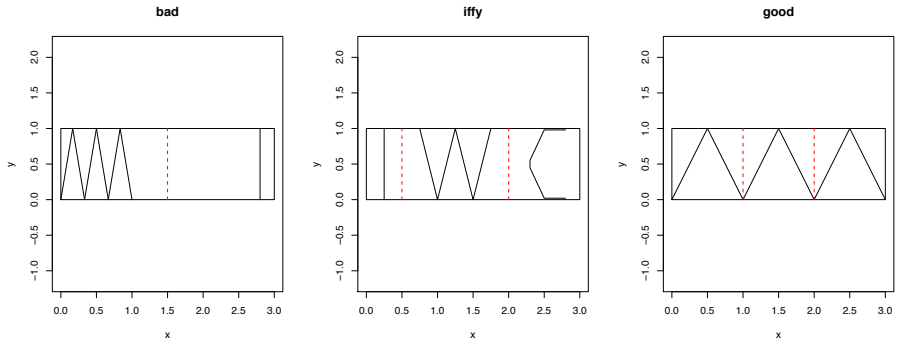

Figure 3: Realised effort for the designs used in the simulations. From left to right: the bad design, with most of the effort on the left of the design; the "iffy" design where effort is more sporadically allocated; the good design with even coverage. In each case the black box around the designs indicates the limits of the study area. Red dashed lines indicate the boundaries of the strata used with the stratified Horvitz-Thompson estimator (see "Models").

using the rotation matrix:

$$R = \begin{pmatrix} cos(\pi/4) & -sin(\pi/4) \\ sin(\pi/4) & cos(\pi/4) \end{pmatrix}$$

- Isotropic smooths
  - Thin plate spline (`bs="tp"`; Wood, 2003)
  - Thin plate spline with shrinkage (`bs="ts"`; Marra and Wood, 2011)
  - Duchon spline (`bs="ds", m=c(1, 0.5)`; Miller and Wood, 2014)
  - Thin plate spline with rotated covariates (`bs="tp"`)
- Tensor product smooths (smooths listed below were used in both directions)
  - Thin plate spline (`bs="tp"`)
  - Thin plate spline with rotated covariates (`bs="tp"`)
- Non-spatially explicit models
  - Horvitz-Thompson (assuming one stratum, using 1)
  - stratified Horvitz-Thompson using strata as shown in Figure 3.

Note that for the detection function part of each model we fit a model of the same form as the generating model, we do not consider model uncertainty or selection for the detection function.

4.5 Software

All simulations were generated using `DSsim` (R package version 1.0.4), with a wrapper scripts used to generate data that could be easily analysed. Detection functions were fitted in `Distance` (R package version 0.9.6) and spatial models were fitted using `dsm` (R package version 2.2.12). Code for the simulations and this paper is available at `http://github.com/dill/spatlaugh`.

4.6 Metrics

In order to assess the performance of the abundance estimates, we use two graphical methods.

4.6.1 Bias

We can simply calculate the bias ($N_{\text{truth}} - \hat{N}$) and plot boxplots of these values, however this does not get to the uncertainty, which we're more interested in.

4.6.2 Where does the truth lie in the distribution of the model?

If we know the true abundance in our simulation (N_{truth}), then we can derive a useful diagnostic measure by asking at what quantile does N_{truth} lie in the distribution implied by the model (i.e., find $\mathbb{P}[N_{\text{truth}} \leq \hat{N}]$). Here we assume log-normally distributed $\hat{N}$, so use the usual formulae to find the resulting quantiles. This summary statistic gives some idea of both bias and variance.

Obtaining the quantile for each simulation, if the distribution of the statistic is skewed to either end then we can infer under or over estimation of abundance for a particular estimator. A flat distribution shows good performance, whereas a "dome" in the middle indicates a conservative estimate in the sense that confidence intervals are slightly too wide. This more conservative behaviour seems desirable, since we probably have no accounted for all of the sources of uncertainty in our model. An example of plots of this statistic is given in Figure 4.

5 Results

For each of the designs illustrated above, we applied each of the three densities along with each of a low, high and changing detectability (both left-right and right-left) leaving us with $3 \times 3 \times 4$ simulation scenarios. In each of these scenarios we tested all of the modelling options listed above and recorded the metrics from the previous section.

The density surfaces described in Section 4.1 only describe the *relative* density of the population in question. We (arbitrarily) fixed the total population size to be 200 individuals for each simulation.

Note that the aim here is not to show which spatial model is best out of those presented here, it is to show that there are large differences between the

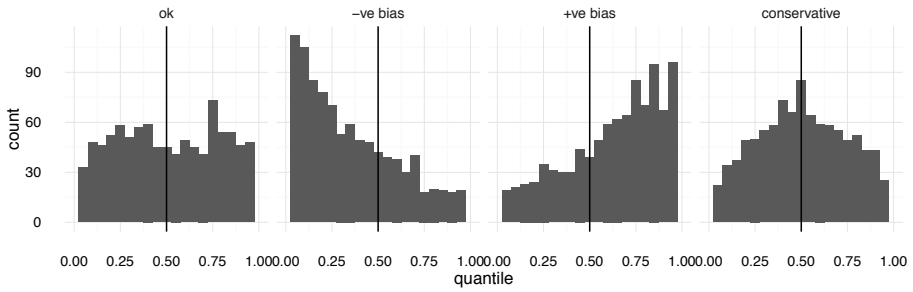

Figure 4: Illustration of the "self-confidence" measure. From left to right: "ok" denotes shows a plot with flat quantile distribution (no problems), "-ve bias" shows a large spike at zero indicating negative bias, "+ve bias" shows the spike at 1 indicating positive bias, "conservative" shows behaviour where the confidence intervals are slightly too wide, which we might prefer (see main text).

HT estimators and the spatial models. There are large differences between the spatial models which can be attributed to the model formulation process. None of the models underwent model checking in the usual way (see e.g. Miller et al 2013, more papers), so the spatial models represent the "dumbest possible" spatial model, without any thought to checking or calibration.

Figure 5 shows the bias in abundance estimates.

6 Discussion

(of things that are still negotiable to include post-Bled :)

- note that we only need consider these simple gradients not multiples, as we can consider each as a stratified version of a "multiple ripple" setup

Appendices

Appendix 1 - Data format for `Distance` and `dsm`

In this appendix we describe the data format required for the two packages used above. The text below is adapted from their respective manuals.

`Distance`

A single `data.frame` should be provided to `Distance` to fit a detection function, or to estimate abundance using the HT estimator. To simply fit a detection function we require the following columns in our data:

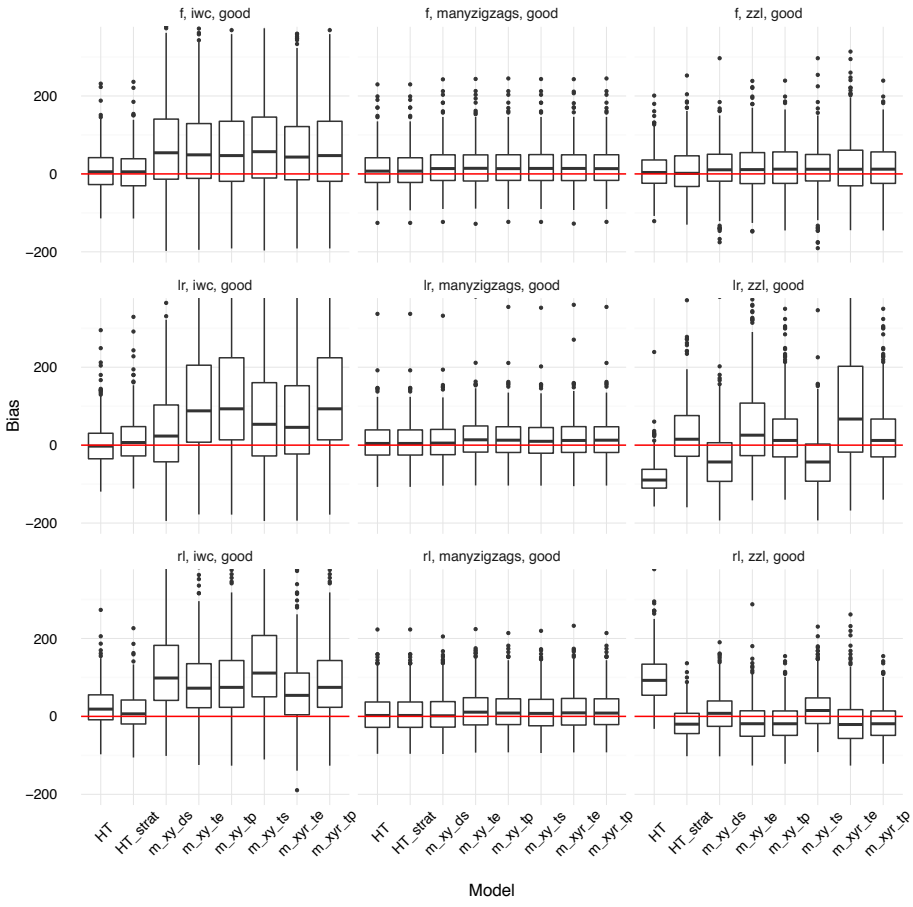

Figure 5: Bias in abundance for each of the models per simulation scenario.

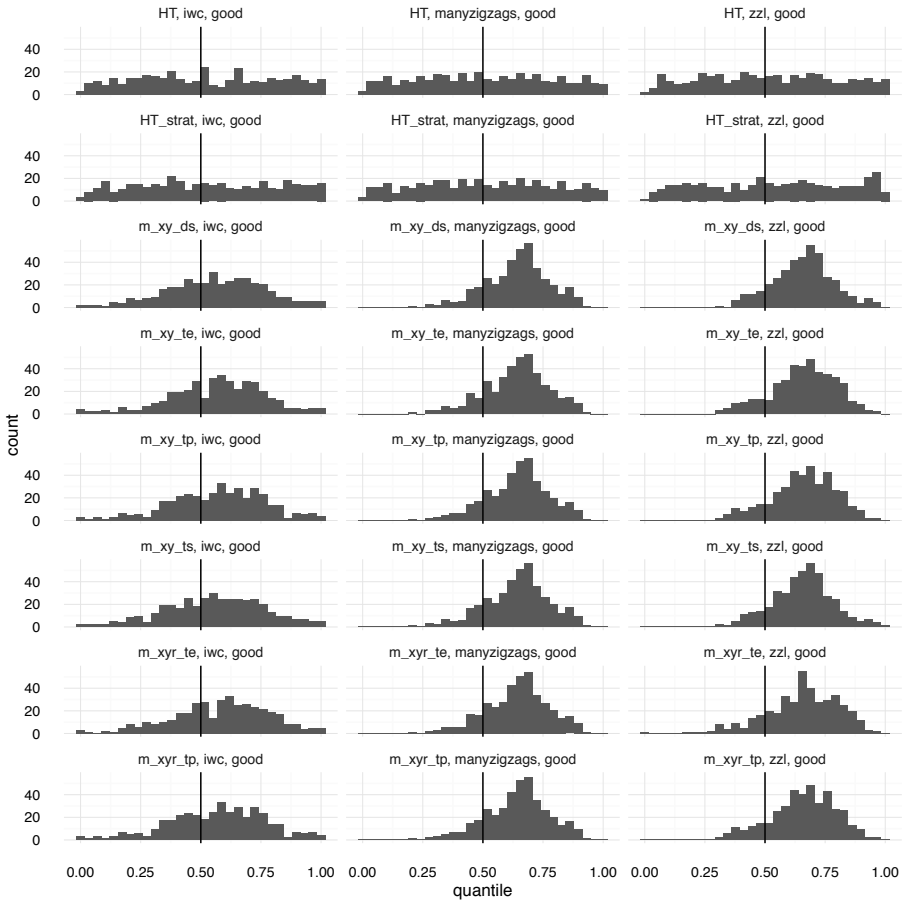

Figure 6: Plots of the "self-confidence" measure for the flat density surface.

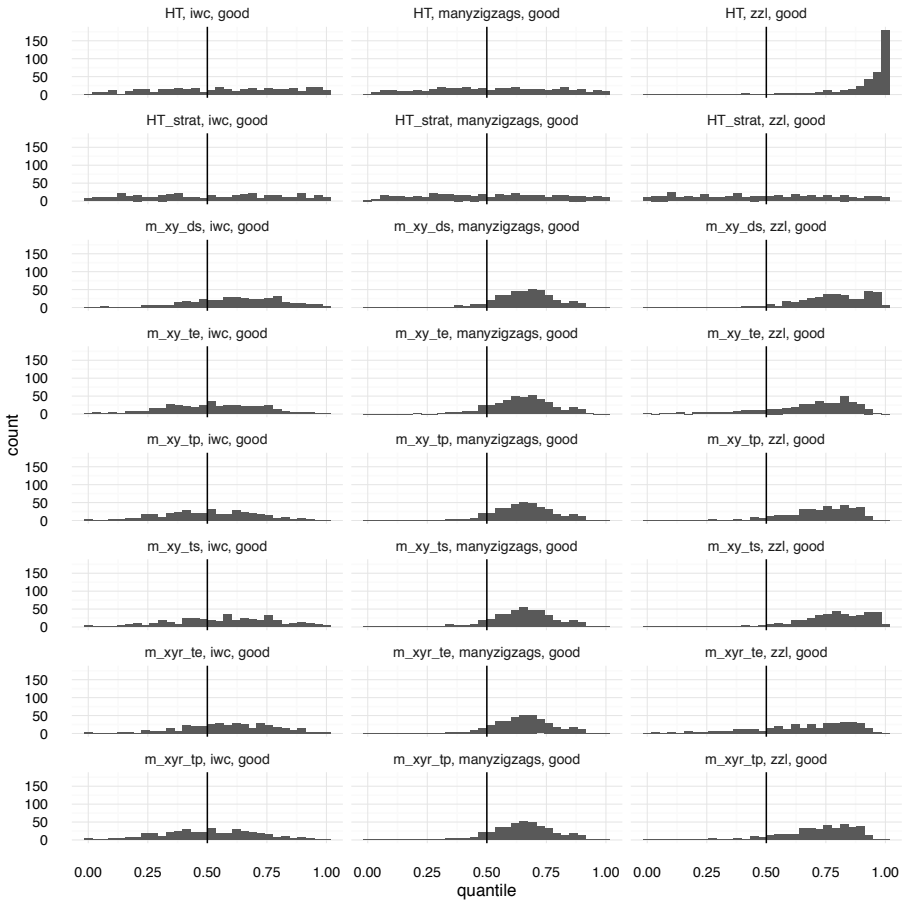

Figure 7: Plots of the "self-confidence" measure for the left-right density surface.

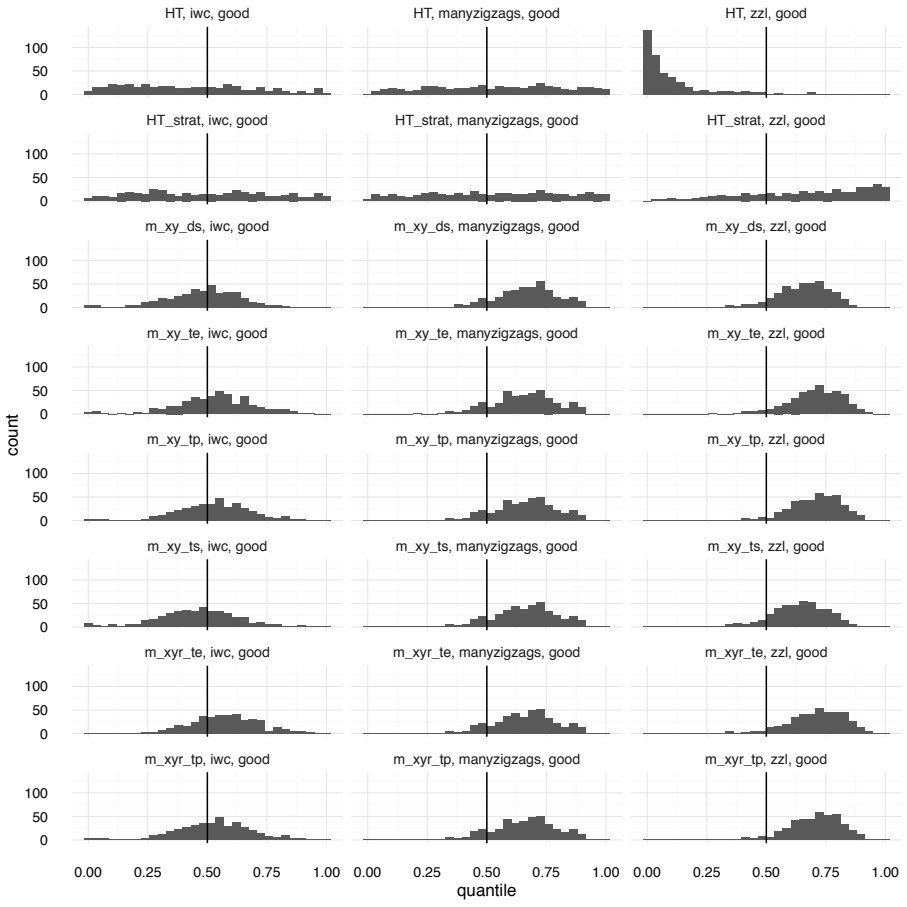

Figure 8: Plots of the "self-confidence" measure for the right-left density surface.

- `distance` observed perpendicular distance to observation from the line
- `object` an unique identifier for the observation

If one wishes to estimate abundance, the following columns are also required:

- `Sample.Label` Identifier for the sample (transect)
- `Effort` effort for this transect (transect length)
- `Region.Label` label for a given stratum
- `Area` area of the strata

Each row corresponds to one observation. In some cases a given transect or even stratum may contain zero observations. In this case the transect(s) are still included, along with their effort, but their corresponding `object` and `distance` fields are set to "not available" (in R "`NA`").

dsm

Two `data.frame`s must be provided to `dsm`. They are referred to as `observation.data` and `segment.data`. The `segment.data` table has the sample identifiers which define the segments, the corresponding effort (line length) expended and the environmental covariates that will be used to model abundance/density. `observation.data` provides a link table between the observations used in the detection function and the samples (segments), so that we can aggregate the observations to the segments (i.e. `observation.data` is a "look-up table" between the observations and the segments).

`observation.data`

The observation `data.frame` must have (at least) the following columns:

- `object` unique object identifier
- `Sample.Label` the identifier for the segment that the observation occurred in
- `size` the size of each observed group (e.g 1 if all animals occurred individually)
- `distance` distance to observation

One can often also use `observation.data` to fit a detection function (so additional columns for detection function covariates are allowed in this table).

`segment.data`

The segment `data.frame` must have (at least) the following columns:

- `Effort` the effort (in terms of length of the segment)
- `Sample.Label` identifier for the segment (unique!)
- ??? environmental covariates, for example: location (projected latitude and longitude), and other relevant covariates (sea surface temperature, bathymetry etc).

Appendix 2 - Mathematical equivalence between HT and DSM

This should go somewhere, why not here?