When can abundance surveys be analysed with “design-based” methods?

David L. Miller[123] and Mark V. Bravington[4]

7th March 2017

[1]Integrated Statistics, Woods Hole, MA
[2]Centre for Research into Ecological and Environmental Modelling & School of Mathematics and Statistics, University of St Andrews, St Andrews, Scotland
[3]Woods Hole Oceanographic Institution, Woods Hole, MA
[4]Commonwealth Scientific and Industrial Research Organization, Hobart, TAS

1 Introduction

The IWC Scientific Committee (SC) often has to consider abundance estimates derived from line transect surveys which have been analysed using “design-based estimators”, but where for various reasons it is not clear whether the resulting estimates of abundance and uncertainty are trustworthy for, say, RMP purposes. The SC has therefore in recent years (up to 2016) considered revising its formal Guidelines[1] to take advantage of methodological developments, in particular the increased flexibility offered by “model-based” abundance estimates, constructed statistically around smooth estimates of animal density across space. As background for that revision, document IWC/65b/RMP11 (Hedley & Bravington 2014) describes the randomization assumptions *required* by design-based *principles*; it also introduces some of the practical issues around model-based estimates, which are more flexible but also more complicated to implement. In section 11 of that document, the authors note that it may sometimes be possible to derive acceptable estimates of abundance and uncertainty using a design-based calculation— which in *this* document we call an HT (“Horvitz-Thompson-like”) estimate— even when the underlying design-based assumptions are not strictly met, provided that (among other things) realised coverage is sufficiently even. This has been common practice at IWC and elsewhere, but generally on an *ad hoc* basis with no clear criteria for “how bad is too bad?”. For such cases, IWC/65b/RMP11 recommends instead that HT acceptability needs to be verified on a case-by-case basis, using diagnostics derived from model-based analysis.

This document follows up on that suggestion. We briefly review the main differences between HT and model-based estimates (section 2), explain where problems can occur with the former (3), and through simulation demonstrate those problems and how they might be checked for using model-based criteria (sections 4, 5). The idea is to consider a range of scenarios about underlying density gradients, then fit different spatial models (including a “null model” that is equivalent to an HT estimate) to data simulated from the actual survey tracks and each density scenario, then check the consistency of point estimates and variance estimates across the different models. Software implementing these criteria/checks is available as an accompanying R package.

[1] The Requirements and Guidelines for Conducting Surveys and Analysing Data within the Revised Management Management Scheme: 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.

In order to keep things tractable, we have made certain assumptions, beyond the usual requirements for distance sampling (see e.g., Buckland, Anderson, Burnham, Borchers et al., 2001; Buckland, Anderson, Burnham, Laake et al., 2004; Buckland, Rexstad et al., 2015)):

- Transect lines are the ones actually visited ("realised effort"), not whatever may have been planned in the original design.
- We assume there is *some* environmental covariate related to sighting probability that has been measured along (within) transect lines, not just when sighting is made (e.g., Beaufort sea state recorded as "good" or "worse" categories). We refer to this covariate simply as "weather". For evaluating a survey, it should not matter too much exactly how the weather covariate is defined. Without a weather covariate, though, we do not consider it possible to adequately evaluate HT acceptability *post hoc* (see section 3).
- We assume that detectability, broadly speaking has already been estimated in a preliminary step. So $g(0)$ and availability etc have already been estimated . What is required for evaluation, is the integrated detection probability out to the truncation distance; the details of functional form do not matter.
- We have ignored school/group size; if a survey is not going to lead to reliable HT analysis even if all groups are size 1 with equal detectability, then variations in group size (and detectability) are hardly going to improve matters. Group size error is an even harder issue. See section 9 of IWC/65b/RMP11 for some discussion.

With these assumptions in mind, we now go on to describe the HT and spatial approaches to estimating abundance.

2 Methods

We do not spend time here describing fitting the detection function to the distance data and refer readers to Buckland, Anderson, Burnham, Borchers et al. (2001), Buckland, Newman et al. (2004) and Buckland, Rexstad et al. (2015) for information on model formulation and selection. Assuming a fitted detection function (which we will denote $\hat{g}$— a function of distance (y), observed detection-related covariates such as weather ($\mathbf{z}$), and estimated parameters ($\boldsymbol{\theta}$)), we can estimate the average (over distance) probability of detection for an observation (indexed by i, conditional on observed covariate values), $\hat{p}_i$, by integrating $\hat{g}(y; \mathbf{z_i}, \boldsymbol{\theta})$ across perpendicular distances y out to the truncation distance.

2.1 Horvitz-Thompson-like estimators

We first describe the "Horvitz-Thompson-like" estimator (e.g., Borchers and Burnham, 2004). 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 (number of groups if animals occur in clusters), indexed by i, s_i is the size of the i^{th} group (or always 1 if animals are solitary) 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 sum of 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, which we then rescale to the study region. The HT estimator can only estimate animal density at a constant level within the area A; the rescaling assumes that the density in the observed area is representative of the larger area. The unbiasedness of equation 1 relies on the inclusion probability of the observations being known, we estimate these (via estimating detectability and implicitly assuming equal inclusion probability; see IWC/65b/RMP11 section 1) so equation 1 is only unbiased when these assumptions hold. These assumptions may be approximately valid in some situations, but it is hard to check directly.

We can perform pre or post hoc stratification, slicing-up the study area into smaller regions (*strata*) and estimating abundance for each of these. These may be geographically defined ("near vs. far from shore", "east/west of some feature", 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 in a given stratum), 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. The assumptions and shortcomings mentioned in the previous paragraph still apply, but on a smaller scale, so may be more defensible.

Variance is estimated for $\hat{N}_{HT}$ by noting that there are (at least) two sources of stochasticity in the equation: *(i)* from the variance in the model for $\hat{p}_i$ (i.e., the detection function) and *(ii)* by noting that the observed number of observations, n, is random. If animals are observed in groups, then there is also variation from the group size to consider. 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 on the scale of $\hat{p}_i$ (Borchers, Buckland and Zucchini, 2002, Appendix C). Variance in n is usually calculated as variance in n/L: the *encounter rate* variance. There are a number of formulations 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 (though we use an encounter rate estimator of $\hat{N}/L$ as in Innes et al., 2002). Note though that this estimator only considers variance between the sample units: the transects, so if transects are long and there is significant variation in density along the transect, we may obtain variance estimates that are far too small. If animals occur in groups, we calculate the variance in group size empirically.

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 modelling* (DSM; Hedley and Buckland, 2004; Miller, Burt et al., 2013). As with HT estimators, 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 (Hastie and Tibshirani, 1990; Ruppert, Wand and Carroll, 2003; Wood, 2006a) to build smooth, spatially explicit terms describing the distribution of the species and their response to other biological/physical variables (though here we only consider models that include smooths of location). Rather than dealing with whole transects (which are generally long and in which animal density and environmental covariate values can change appreciably), we cut the transects into smaller pieces, which we call *segments*. The mean response of the model can be written as:

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

where j indexes the segments, which have area A_j and all observations in that segment have a probability of detection of $\hat{p}_j$. The response, n_j (count per segment), is distributed according to some count distribution (usually Tweedie or negative binomial) for which exp is the inverse link function[2]. The model intercept is β_0. The s_k are usually splines (Boor, 1978): smooth functions of one or more covariates (denoted z_{jk}), though could be more exotic things like random effects, tensor products, smooth-factor interactions and so forth (Wood, 2006a; Wood, Pya and Säfken, 2016). 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 (IWC/65b/RMP11 also discusses this in section 7.2).

A typical DSM may take a form such as:

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

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 such as weather, observer shift/team or ship can be included in the detection function, but sex of the animal or observer ID (if there were multiple observers on deck at once) cannot be[3].

3 Why do spatial modelling?

A criticism of the DSM approach is that it is more complex, as we explicitly model the spatial and possibly environmental covariate effects. More mode checking is required and there is a wide literature of often conflicting advice on how to construct models. However, explicit modelling is the only way to deal with the heterogeneity in spatial distribution of the study species. In this section we give some some arguments about why HT estimation may not be a good idea from a theoretical perspective.

3.1 Appeals to pooling robustness

An appeal to "pooling robustness" (Buckland, Anderson, Burnham, Laake et al., 2004, Section 11.12) does not get around heterogeneity in the spatial distribution of the population. Before explaining why, we first define and explain pooling robustness in a distance sampling context. From Burnham, Anderson and Laake (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 r 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 if pooling robustness holds, the estimates from a stratified analysis would be the same as those for an unstratified analysis. Buckland, Anderson, Burnham, Laake et al. (2004, Section

[2] Though any exponential (and beyond) 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.

[3] These covariates can be included in a more general formulation of DSMs, though we don't consider them here for clarity of presentation. See Miller, Burt et al. (2013) for further details.

11.12) state: “if only an overall abundance estimate is required, standard methods without covariates [in the detection function] are satisfactory under rather mild conditions, provided heterogeneity in detectability is not too extreme.” This is a statement about HT estimation in the presence of (mild) detection heterogeneity and does not address spatial variation in density within strata. If we knew the optimal R strata to chop our survey into, according to the density gradient we might be onto something, but this is impossible in practice.

So far we have only considered the case where detectability is certain on the line ($g(0) = 1$). If we start to consider issues around uncertain detectability on the line (Borchers, Zucchini and Fewster, 1998; Borchers, Buckland, Goedhart et al., 1998; Borchers, Laake et al., 2005; Burt et al., 2014) this situation only gets more complicated. Buckland, Anderson, Burnham, Laake et al. (2004) note that if the detection function includes covariates “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 in a cetacean survey), pooling robustness does not apply.

3.2 Weather covariates

Following on from pooling robustness, we now think more about recording a sighting conditions covariate, such as weather. It is not enough to “see” that tracklines physically cover the entire study area in a seemingly even way; if most of the survey effort in the West are in good sighting conditions whereas most in the East are in poor conditions, then the actual effort can be quite imbalanced. (The tracklines on the West have higher sighting probabilities, so they are in effect “wider” and cover a higher proportion of nearby areas, but the width difference is not visible on the scale of a plot.) The average probability of detection of a whale (across all whales in the region) is not the same as the average detection probability averaged along the trackline.

If density does vary within strata 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. For example, say we saw many animals in the West of the study area and fewer in the East. Let us also say that there were good sighting conditions in the West and poor sighting conditions in the East. If we ignore the weather covariate, we have know way of knowing if there truly were many animals in the West or whether it was just the effect of having good sighting conditions.

This is the reason for our stipulation that HT-acceptability (i.e., robustness to animal density variations) can only be assessed when a weather covariate is available. For a full analysis of data, some care is usually given to the choice of covariate (e.g., number of levels, if the covariate is a factor; use of several covariates together), but for broad assessment of uniformity, we suspect that the choice does not matter too much. For example, splitting Beaufort into two levels such that the “good” level contains 1/3—2/3 of total sightings, and the “poor” level contains the rest, should be sufficient. Using a simple classification of that type makes it easy to estimate a separate detection probability by “weather level”, simply by splitting the sightings according to weather before fitting a separate detection function to each.

3.3 Designs with unknown coverage probability

Implicit in the HT formula (1) is the assumption that all locations in the study area are (at least approximately) equally likely to be sampled (e.g., IWC/65b/RMP11 section 1; Borchers and Burnham, 2004). With a truly randomised (or systematic) design, this is certainly the case, but it is often the case that designs are “constructed” rather than realised from a randomisation process (such as that provided in Distance for Windows (Thomas et al., 2010)). This can be for a variety of logistical reasons, but the lack of randomisation invalidates the design-based justification for HT estimation of abundance.

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

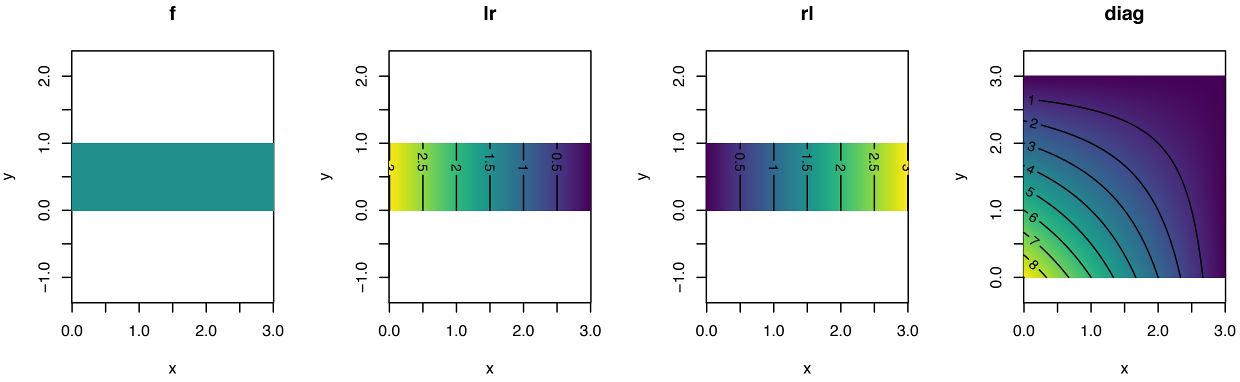

Figure 1: The proposed density surfaces. From left to right: flat, left-right gradient, right-left gradient, diagonal gradient. Shorthand names are given as the title for each plot.

grid, ensuring that unreasonable extrapolations are not made. 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 perform poorly for these simple density surfaces (where gradients are clearly defined) then it is 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.
- "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.
- "diag": increasing gradient from top right to bottom left

These are shown in Figure 1.

4.2 Detectability

Detectability in the survey was set at two fixed detectabilities "good" and "bad" by varying the parameters of a hazard-rate detection function. We also simulated a two-level weather covariate that changed from left to right across the study area, on the left side the weather was "good" (using the "good" detection function) and on the right side the weather was "bad" (using the "bad" detection function). The transition between the

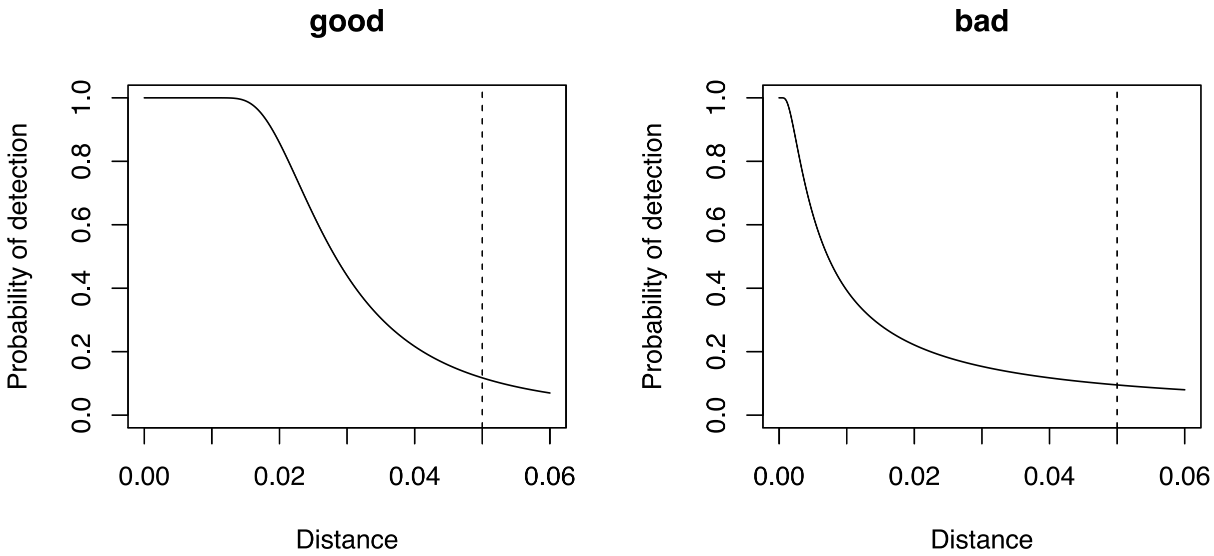

Figure 2: Detection functions used in the simulations. Dashed line indicates the truncation distance used. For the "good" detection function, the scale parameter was set at 0.025, shape parameter at 3. For the "bad" detection function, the scale parameter was set at 0.005, and the shape at 1. In both cases truncation was at 0.05.

detection functions was controlled by a logistic function. Due to its change along the x axis, the covariate is confounded with the "lr" and "rl" density gradients – this is eminently possible in survey data as discussion in section 3.2. Plots of the detection functions used to simulate the data are shown in Figure 2 along with their parameters.

For clarity and simplicity we do not consider the case where detectability is uncertain at zero distance ($g(0) \neq 1$) here — we assume that if animals occur directly in front of the observer they will be seen. We also do not consider any availability issues (that cetaceans are often underwater and cannot be seen). Finally, we also assume that observations are of single animals (that group size is one). These are simplifying assumptions, but a modelling strategy that fails in this simple situation is unlikely to perform well once any of these assumptions are relaxed.

4.3 Designs

We experimented with four designs. A good design with good realised coverage across the whole study area ("manyzigzags"; to confirm that what we consider to be a good design gives us the results we expect from our metrics). Two "iffy" designs: one with a large gap between each contiguous section of realised effort ("zzl") and one with two very different effort distributions for two parts of the survey ("twozigzags"). A bad design where the effort is concentrated along two sides of the area ("corner"). The designs are shown in Figure 3. In each case the box surrounding the design indicates the area used for prediction in the simulations, dashed lines indicate divisions between strata used for stratified HT estimates.

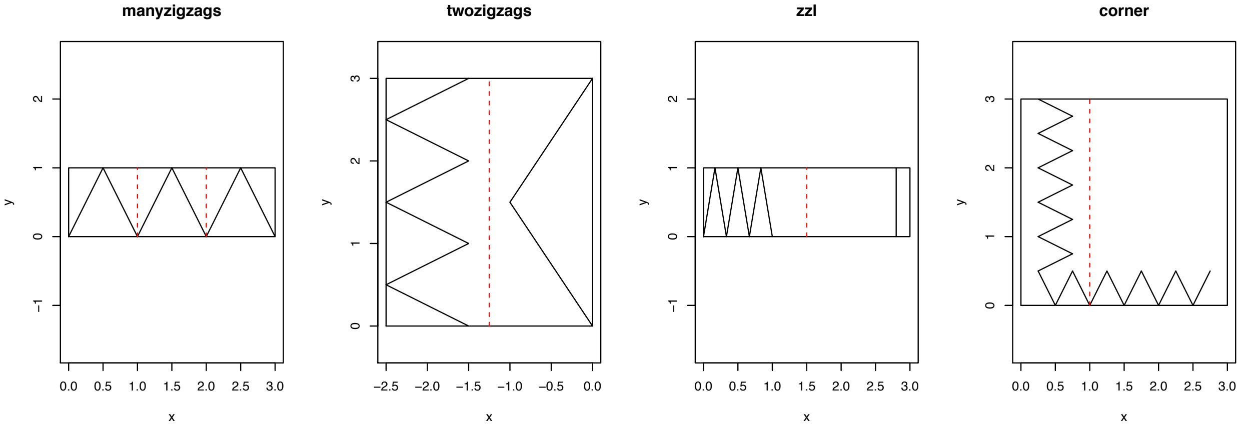

Figure 3: Realised effort for the designs used in the simulations. From left to right: good design with even coverage; an "iffy" design where effort high in one side and low in the other; an "iffy" design where effort is more sporadically allocated; the bad design, with most of the effort in the left and bottom of the survey area. 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 HT estimator (see "Models").

Design 1: zig-zag with good coverage ("manyzigzags")

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

Design 2: zig-zag with straight line ("zzl")

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

Design 3: two different zig-zags ("twozigzags")

To show how designs often consist of different coverages in different areas of the survey region, a design with more effort (many zig-zags) is placed next to that with much less coverage (a single zig-zag). This is shown in the third panel of 3.

Design 4: corner ("corner")

This design concentrates along two sides of the study area, this design mimics the commonly used technique where near-coast transects are used to extrapolate well beyond the covered area. 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 coordinate system by 45° using the rotation matrix:

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

The models tested were (starting with their shorthand code used for plots later):

- Isotropic smooths
  - TP: thin plate spline, `bs="tp"` (Wood, 2003)
  - TPSH: thin plate spline with shrinkage, `bs="ts"` (Marra and Wood, 2011)
  - Duchon: Duchon spline, `bs="ds", m=c(1, 0.5)` (Miller and Wood, 2014)
- Tensor product smooths (smooths listed below were used in both directions)
  - TPTE: thin plate spline, `bs="tp"`
  - TPTER: thin plate spline with rotated covariates, `bs="tp"`
- Non-spatially explicit models
  - HT: Horvitz-Thompson (assuming one stratum, using 1)
  - HTstrat: stratified Horvitz-Thompson using strata as shown in Figure 3
  - HTcovar: Horvitz-Thompson with covariates included (where applicable)
  - HTstratcovar: stratified Horvitz-Thompson with covariates using strata as shown in Figure 3 (where applicable)

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. For simulations where the weather covariate was simulated, all spatial models use a detection function with the covariate included, we include estimates for HT-based methods both including and not including the weather covariate. The spatial models used are not *all* the ones that might be considered in practice for a full-on *bona fide* spatial analysis (e.g., we did not consider soap-film smooths, because there are more application-specific insights to set up appropriately), but we expect that they should encompass a wide range of ways to respond to data.

4.5 Software

All simulations were generated using `DSsim` (R package version 1.0.6), with a wrapper scripts used to generate data that could be easily analysed, these are collected in the R package `ltdesigntester` (available at `http://github.com/dill/ltdesigntester`); a vignette is provided with the package to illustrate its use. 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.

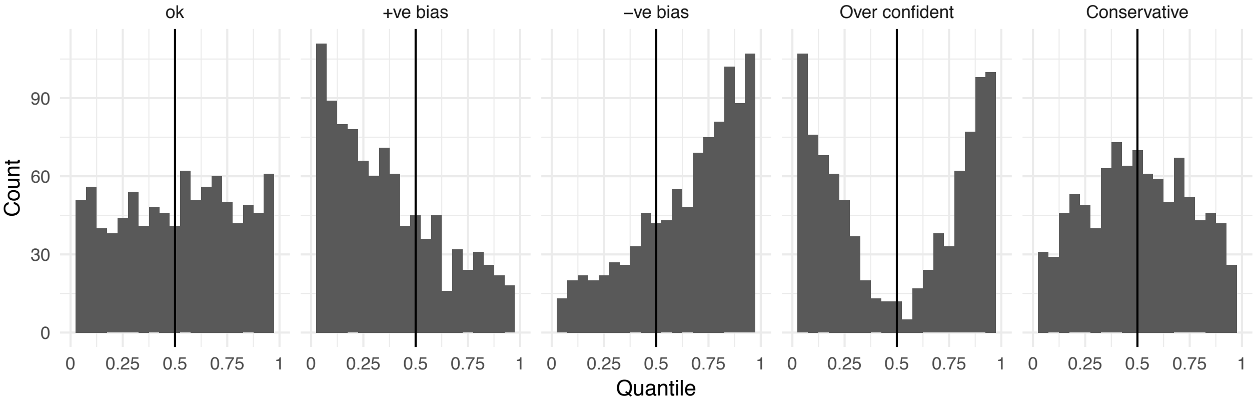

Figure 4: Illustration of the "self-confidence" measure. From left to right: "ok" denotes shows a plot with flat quantile distribution (no problems), "+ve bias" shows the spike at 1 indicating positive bias, "-ve bias" shows a large spike at zero indicating negative bias, "overconfident" shows a spike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

4.6.1 Bias

We can simply calculate the bias ($\hat{N} - N_{\text{truth}}$) and plot boxplots of these values. Bias may only be interesting in certain situations, for example when we are very uncertain about abundance, the bias may not be our main concern.

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 "self-confidence" statistic gives some idea of both bias and variance. If the model is working as it should, then the self-confidence quantile should be uniformly distributed across simulations. 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. Skew to both ends indicates overconfidence. 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 in practice we probably have not 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

We experimented with possible combinations of simulation scenarios based on the models above. In each scenario, 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 500 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 HT estimators and the spatial models in particular situations. Differences between the spatial models can be attributed to the model formulation process, for example in the cases where the gradient was simply right-left, left-right or flat, fitting a bivariate spatial model will likely not perform particularly well as it is too flexible. Additionally, none of the models underwent model checking in the usual way (e.g., Miller, Burt et al., 2013; Winiarski et al., 2014), so the spatial models represent the "dumbest possible" spatial model, without any thought to checking or calibration.

We summarise the results in the remaining subsections of this section. Bias and "self-confidence" plots for all simulations are provided in Appendix 2, results (in RData format) are include in the project GitHub repository.

5.1 General comments

Before looking at where particular models succeeded or failed, we first look at general trends in the results. First we note that for any design, flat densities are easy to estimate for HT-type methods. We note that the rotated coordinate spatial models (TPTER) sometimes perform better than their non-rotated variants; we believe this is down to the non-rotated models having another direction of variability (in the y direction) that is effectively unused (and which we would therefore like to be estimated as a zero effect), but that may pick up on minor variations due to sampling variability. In the rotated models both coordinates are used, so estimating an effect of exactly zero for the y component is not an issue.

5.2 Designs with good coverage work well

As expected, when coverage is even (Appendix 2, plots labelled "manyzigzags"), the methods based on HT estimators perform well in terms of bias in estimating $\hat{N}$. Only when the detectability becomes particularly bad does the "self-confidence" diagnostic start to look bad for the HT estimators (at which point it also begins to look worse for the spatial models).

5.3 Uneven coverage is bad

Moving to an uneven design like "zzl" or "twozigzag" makes unbiased estimation of abundance much harder and we see all models perform worse. Though in particular, we see that the unstratified HT estimate (HT) perform much worse and the stratified HT also perform poorly — this is particularly revealing as in our simulations, the "correct" stratification was provided to the estimator, a luxury that we do not have in real life. The "self-confidence" diagnostic plots also show a shift in all models, though again the HT-based methods perform worse than the spatial models on the whole. If performance is poor with the correct stratification, then we can assume that things will be much worse when stratifications are decided *a priori* based on logistical constraints rather than information on true distribution. This is likely amplified when density is not a simple gradient.

Things get substantially worse across the board if the gradient runs against a poor design. This gets even worse when weather confounds the problem. For example, Figure 17 right panel ("zzl" design with right-left density, weather covariate) shows poor performance for all models. In this case the combination of confounding and low effort where there are lots of animals means that nothing can be done — it is probably best to give up on such data.

5.4 Covariates make things complicated

Ignoring the even coverage design, once the weather covariate is included in the data generation process, things look much worse for the models which do not include it. Both bias and "self-confidence" show that stratification and including the covariate do improve the HT estimates, though again we are assuming the correct stratification scheme.

The most complicated simulation setup (corner design with diagonal density gradient and weather covariate) highlights a more usual situation than the others — we are often confronted with uneven coverage (non)-designs, densities that are complex and weather is almost always an issue. Given the performance in the simpler situations, this scenario was bound to be more taxing for the HT-based estimation methods (Figure

8). The “self-confidence” measure is rather bad for all HT-based methods, but the performance of the spatial models (Figure 18) looks like the “conservative” example shown in Figure 4 .

6 Discussion

Simulations presented here highlight potential issues when HT estimators are used in the case where animal distribution is not constant within a given stratum. Violation of the assumption of flat density within a stratum causes issues when coverage is uneven. Though the designs and densities presented here are relatively simple, the problems that arise only become amplified with added complexity. We do not believe that the most complex scenario (corner design with diagonal density gradient and weather covariate) is in any way pathological — on the contrary, we think it represents a fairly mild version of what is often seen in the field. A more realistic version of this scenario would involve incomplete transects (holes in the transects) and a more complex density surface.

In general we see that spatial models, even when applied in a very naïve way, provide a coherent way to think about modelling phenomena that are, by their very nature, spatial. Although formulating, fitting, checking and validating spatial models is more complex than simply using HT-type methods to obtain density, the added time and resources clearly leads to more reliable results. It is worth noting that here for the majority of the density surfaces tested (aside from diagonal case) the spatial models were all overparameterised, in that the “correct” model would be a single smooth of x for the left-right and right-left gradients, not a bivariate smooth of x and y. Nevertheless, the spatial models performed well. If the true density is flat, estimating that density surface (equivalently estimating almost all of the coefficients of the bivariate smooths to be zero), as would be assumed by HT, when sampling of the density surface is not even, is very difficult (not to mention an unrealistic situation). In these cases the spatial model performed almost as well as the HT estimator.

The `ltdesigntester` package developed for this paper is quite general and can be used for any design or density surface. We considered only a two-level “weather” covariate. We encourage investigators to input their current designs and use the supplied simple density surfaces to test how well abundance can be estimated. The software could potentially be adapted and extended in numerous ways, perhaps linked to planned developments in the `dsm` package such as allowance for spatial variation in group size.

We envisage three possible outcomes from testing HT-acceptability, starting with the approaches/software described here, but not necessarily finishing there:

1. An HT analysis should be OK;
2. HT would be risky, but spatial modelling might deliver an acceptable point estimate and variance estimate;
3. The data are so imbalanced that spatial modelling is unlikely to deliver a reliable estimate.

Clearly, the analyses and software proposed here can only differentiate directly between 1. and 2./3. If a spatial model is to be used instead of HT, then the appropriate model — which will use the real locations of sightings, not just the total number — may well be outside the range of simple models in our tests (e.g., if there are edge effects requiring the use of, e.g., soap film smoothers). It may of course also be necessary to enlarge the model to deal with group size etc. Selecting a preferred spatial model, and reporting appropriate diagnostics for it, is beyond the scope of this particular document, but see (IWC/65b/RMP11; Chandler, 2005; Wood, 2006b; Augustin, Musio et al., 2009; Augustin, Sauleau and Wood, 2012; Augustin, Trenkel et al., 2013; Miller, Burt et al., 2013; Winiarski et al., 2014).

It is worth pointing out that we have deliberately avoided using the actual location of sightings. Of course, the sighting-locations themselves could be used to directly fit a spatial model, so why not use that as the point-of-comparison with HT? Our motivation is twofold: first, actually picking a good spatial model for a real dataset is not something that can be automated; second, especially if there are not many sightings, the real density surface is very uncertain, and it is “putting the cart before the horse” to assume that a spatial model fitted to limited data will actually pick the truth accurately. Instead, our approach is a robustness test to assess the survey tracks against a range of density surface scenarios, and as such should be safer. The possible downside is that, if animal density really is quite uniform, then an uneven sampling design might fail our HT-acceptability checks even though the actual estimate would be OK. But this is hardly a disaster; all that would be needed to rescue the data, is to proceed to fitting a spatial model “for real”. For many datasets, this is certainly not an overwhelming task with modern software such as `dsm`.

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:

- `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 - Full simulation results

This appendix gives all plots of results from the simulations that were run. Plots are aggregated by underlying density used to generate the data. Note that boxplots are clipped at their limits, so some extreme outliers may not be shown.

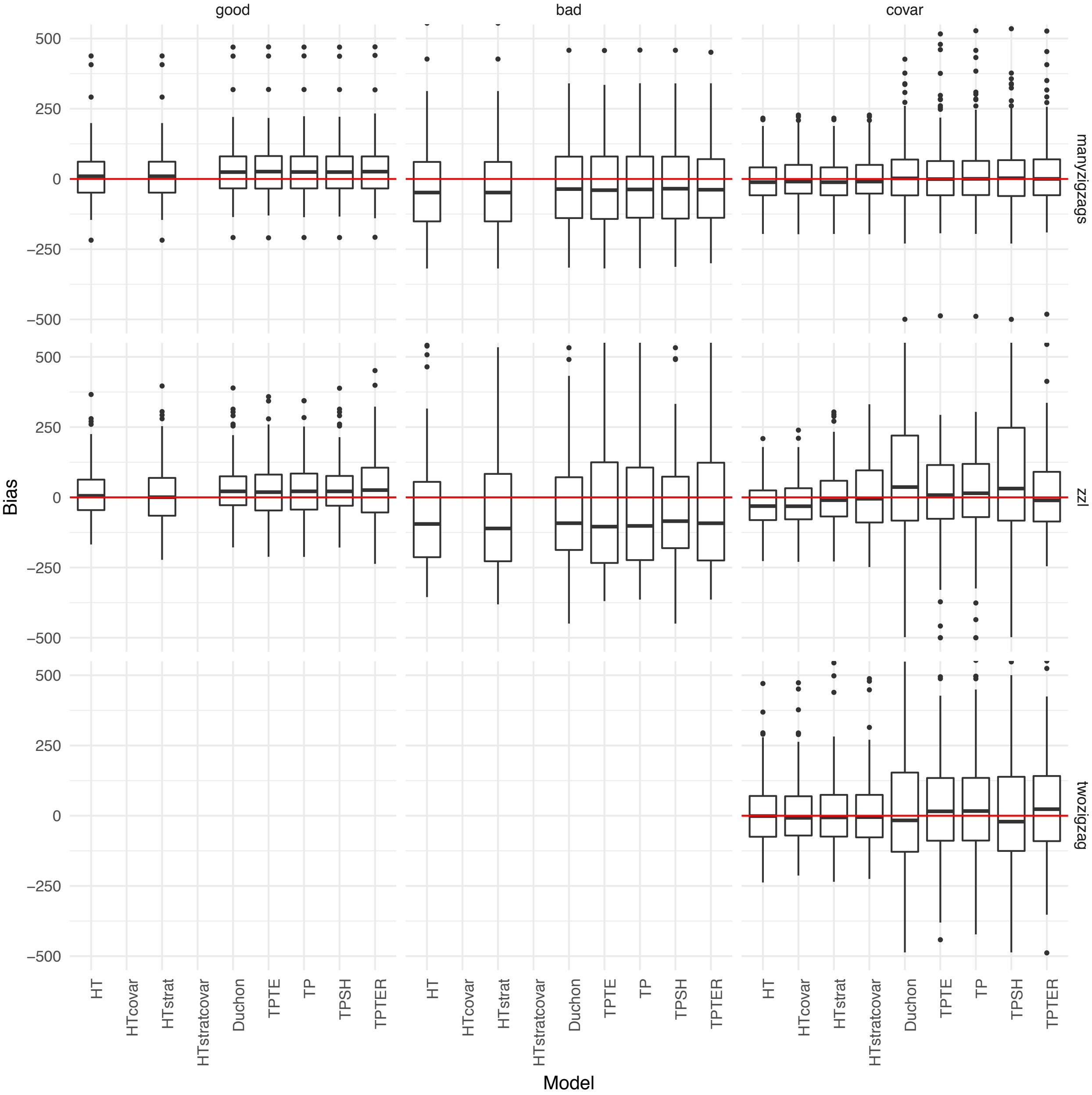

Figure 5: Bias in abundance for each of the models per simulation scenario for the flat density. Columns give the detection function type (see Section4.2) and rows give the design used (see Section 4.3). With a flat density all models perform relatively well, even when the detectability is low (though there is some negative bias).

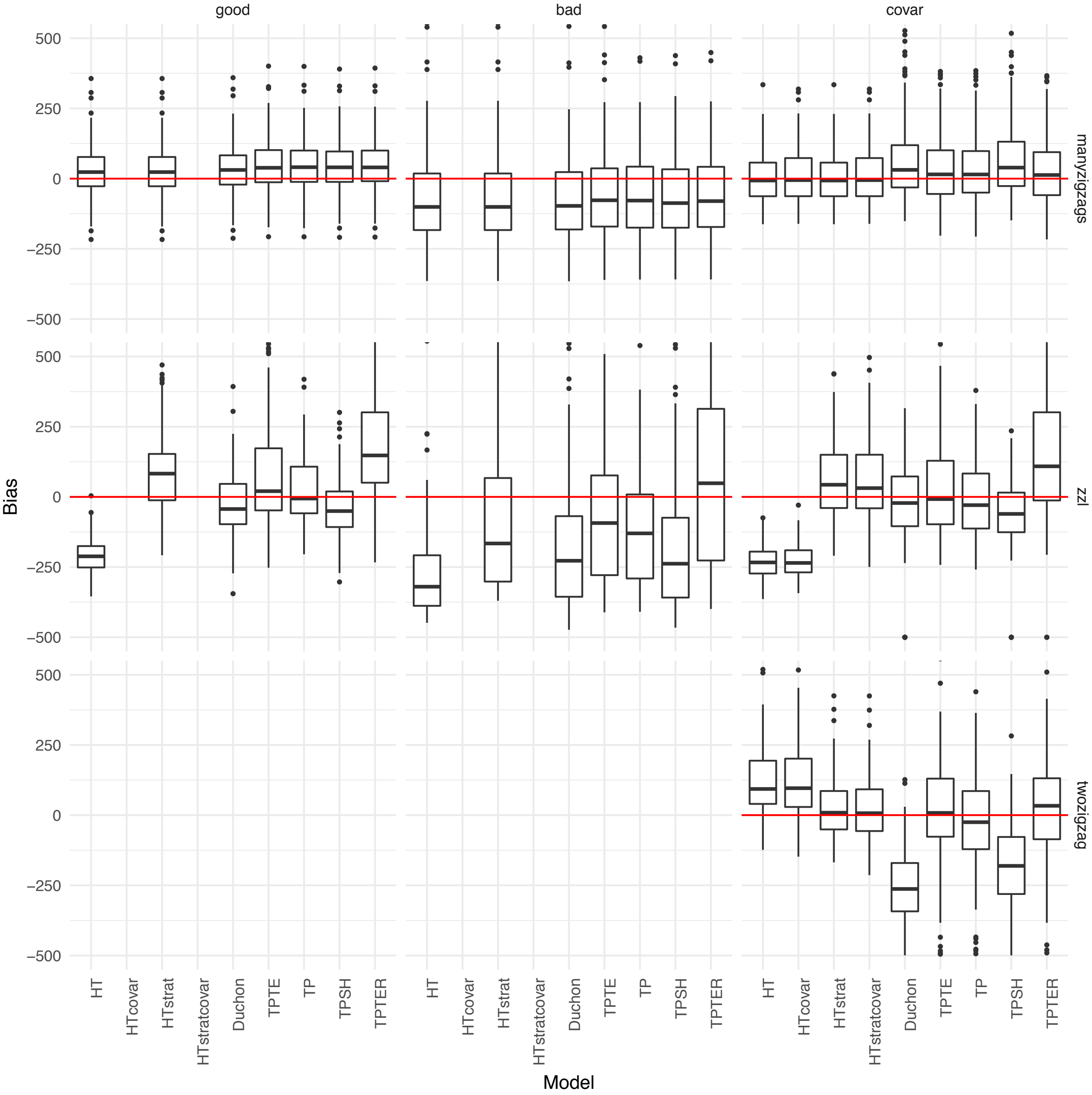

Figure 6: Bias in abundance for each of the models per simulation scenario for the left-right gradient density. Columns give the detection function type (see Section4.2) and rows give the design used (see Section 4.3). In this case for the "zzl" design, most of the effort is concentrated on the higher densities, the low effort side is where there are fewer animals, the HT-based methods perform poorly here, as do some of the spatial models. With such a small amount of effort, it would be hard to detect that the gradient was to blame for low numbers of sightings. For the covariate analysis, high detectability coincides with high density and high effort, so the good performance of the stratified HT and spatial models is likely down to a feature of the confounding between density and covariate value (hence the poorer performance of the non-rotationally invariant TPTER model).

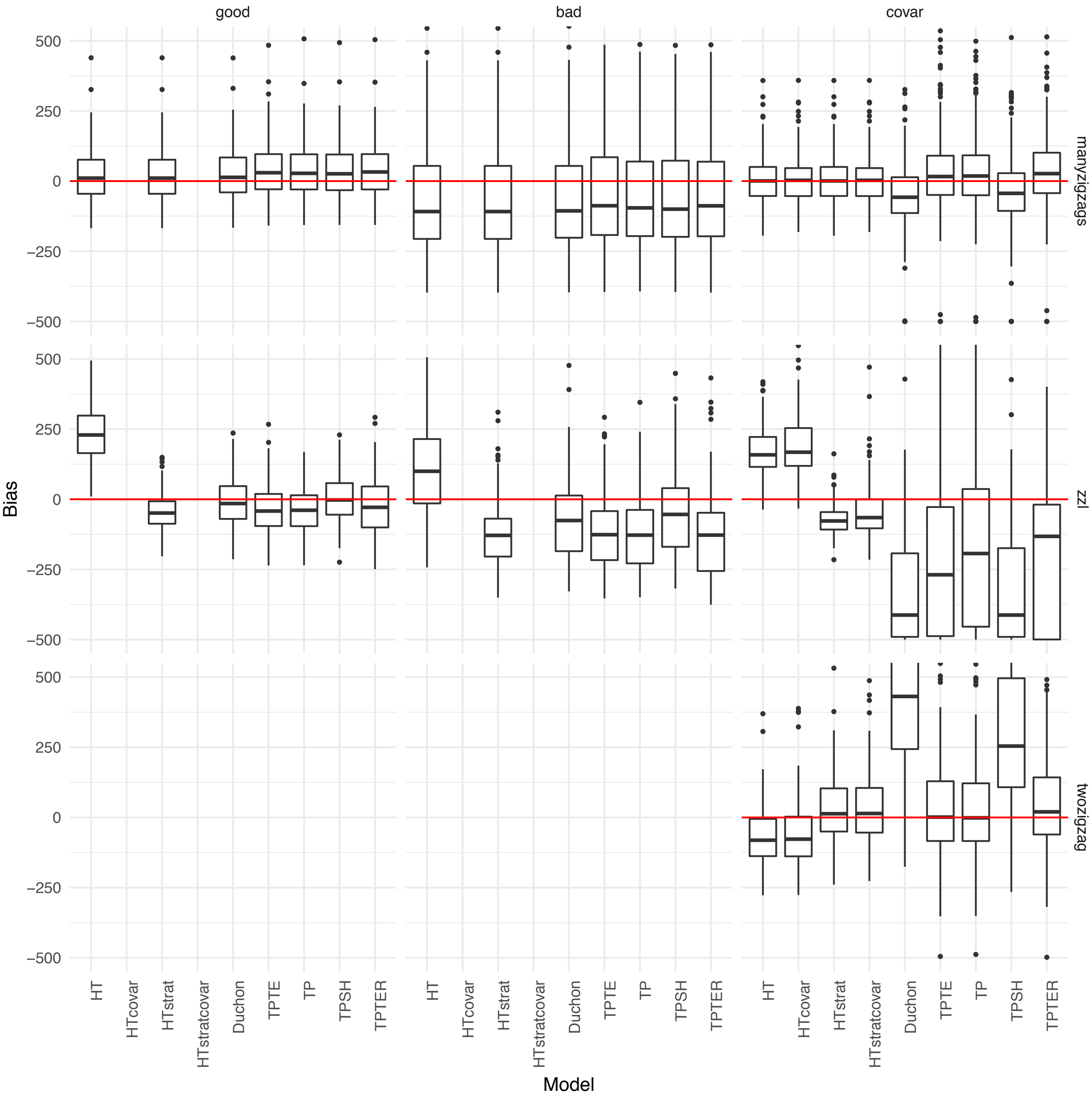

Figure 7: Bias in abundance for each of the models per simulation scenario for the right-left gradient density. Columns give the detection function type (see Section4.2) and rows give the design used (see Section 4.3). Here the "zzl" design allocates more effort to the low density areas, this leads to the non-stratified HT-based estimators to overestimate abundance (assuming high abundance everywhere); stratification improves this somewhat (assuming the correct stratification) but the spatial models are able to pick up on the distribution gradient. For the covariate analysis, high detectability coincides with low density and high effort, leading to bias in the abundance estimates; spatial models have a much greater spread of estimates, obtaining the true abundance at least some of the time.

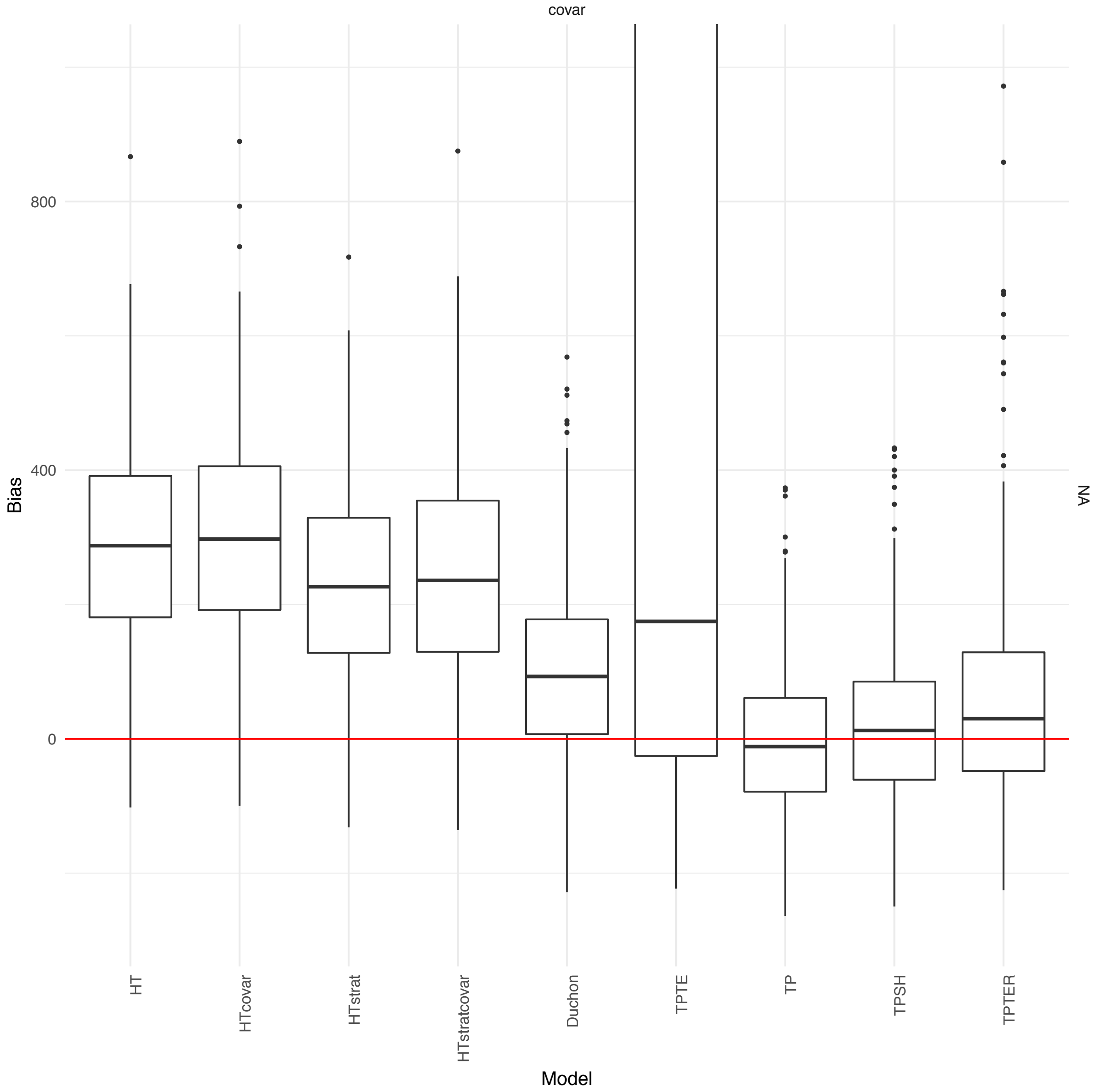

Figure 8: Bias in abundance for each of the models per simulation scenario for the diagonal gradient density. In this case the stratification of the corner design is somewhat arbitrary and there is a large extrapolation to the top right of the survey area (see Figure 3) where the flat density assumption of the HT-based methods fail. Spatial models perform well, aside from the tensor product of thin plate splines (TPTE).

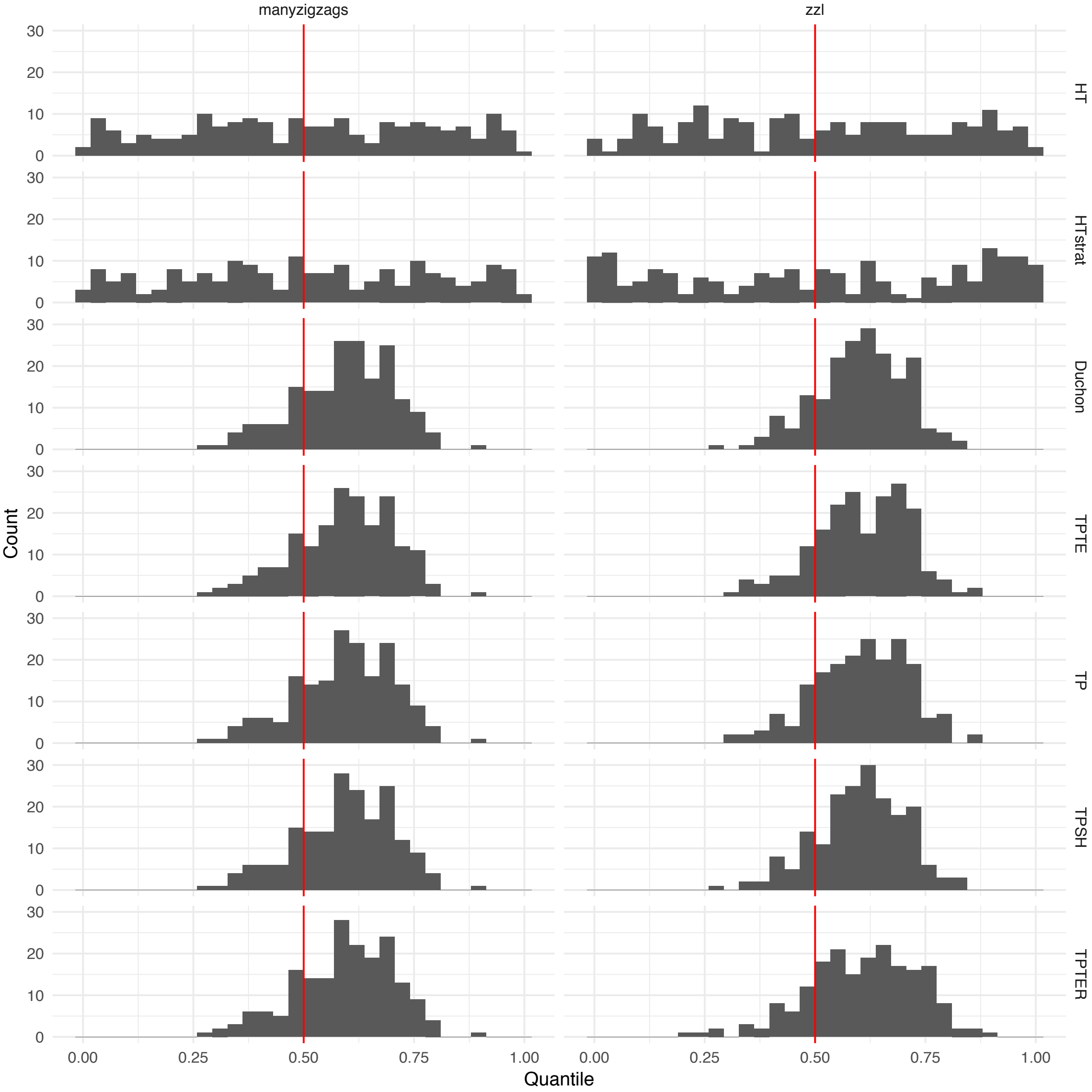

Figure 9: Histograms of the "self-confidence" measure for each model for the flat density surface with a "good" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). Here we see generally good behaviour in all models, though the spatial models are slightly positively biased, which is likely down to model over parametrisation (smooths should be estimated with zero effects, which is hard).

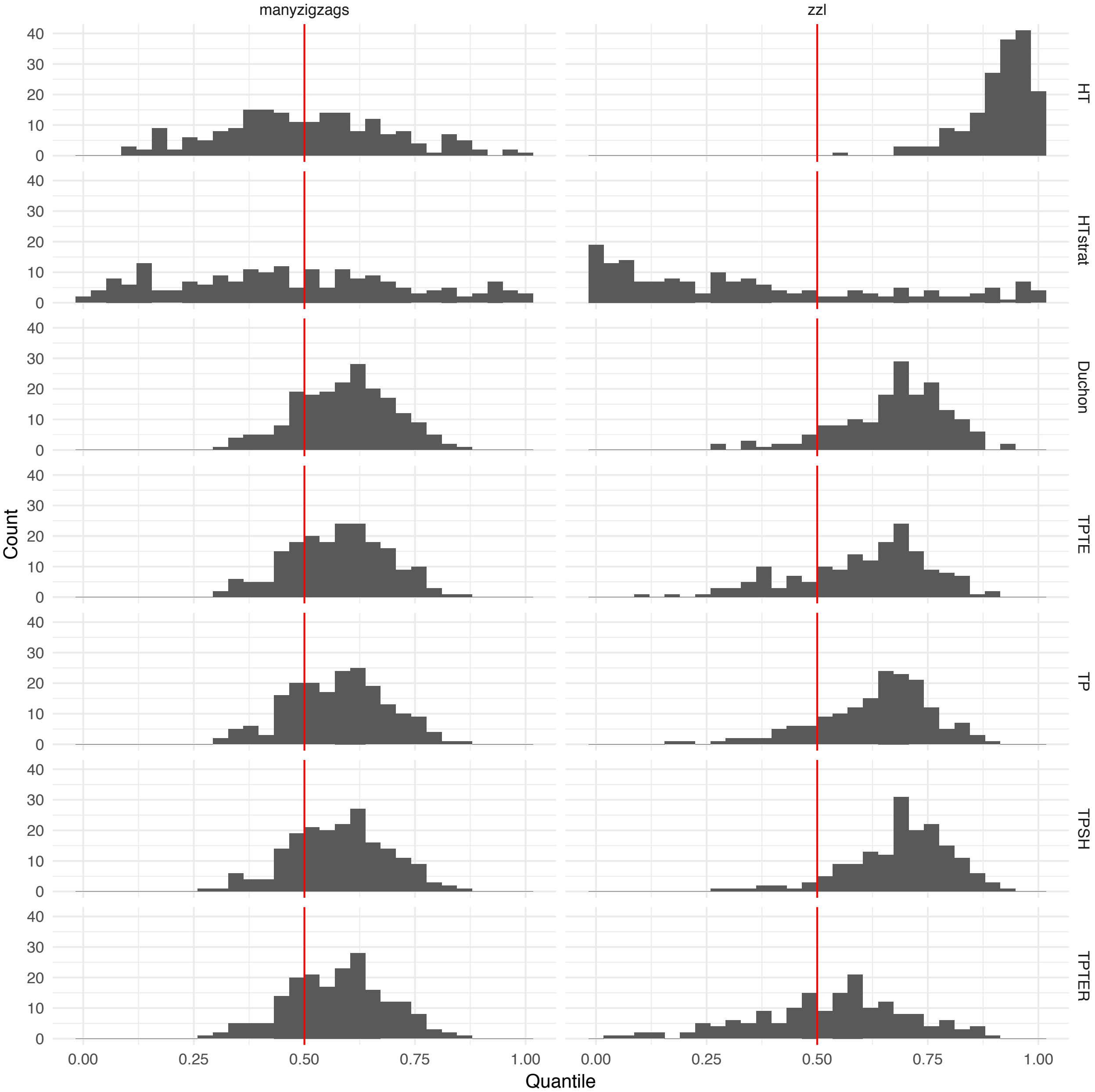

Figure 10: Histograms of the "self-confidence" measure for each model for the left-right density surface with a "good" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). In this case for the "zzl" design, most of the effort is concentrated on the higher densities, the low effort side is where there are fewer animals, as seen in Figure 6, stratification improves things somewhat, though using the rotated covariates makes the spatial model's job easier (again there is an issue with model specification, as we should really only build a model with a smooth of x in this case).

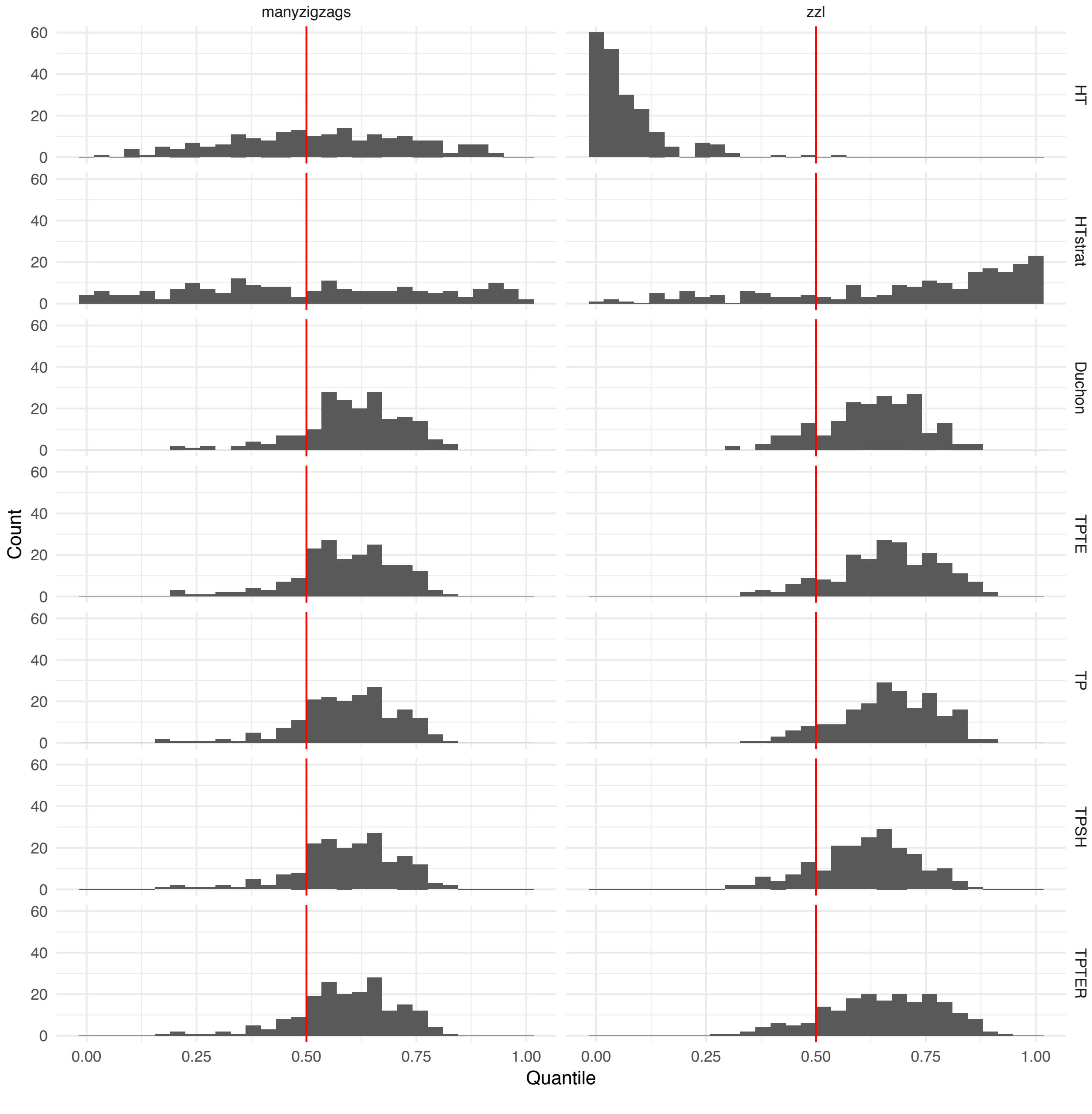

Figure 11: Histograms of the "self-confidence" measure for each model for the right-left density surface with a "good" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). Here the "zzl" design allocates more effort to the low density areas, as in Figure 7 non-stratified HT estimation is positively biased, stratification improves this but the spatial models are better able to estimate the gradient.

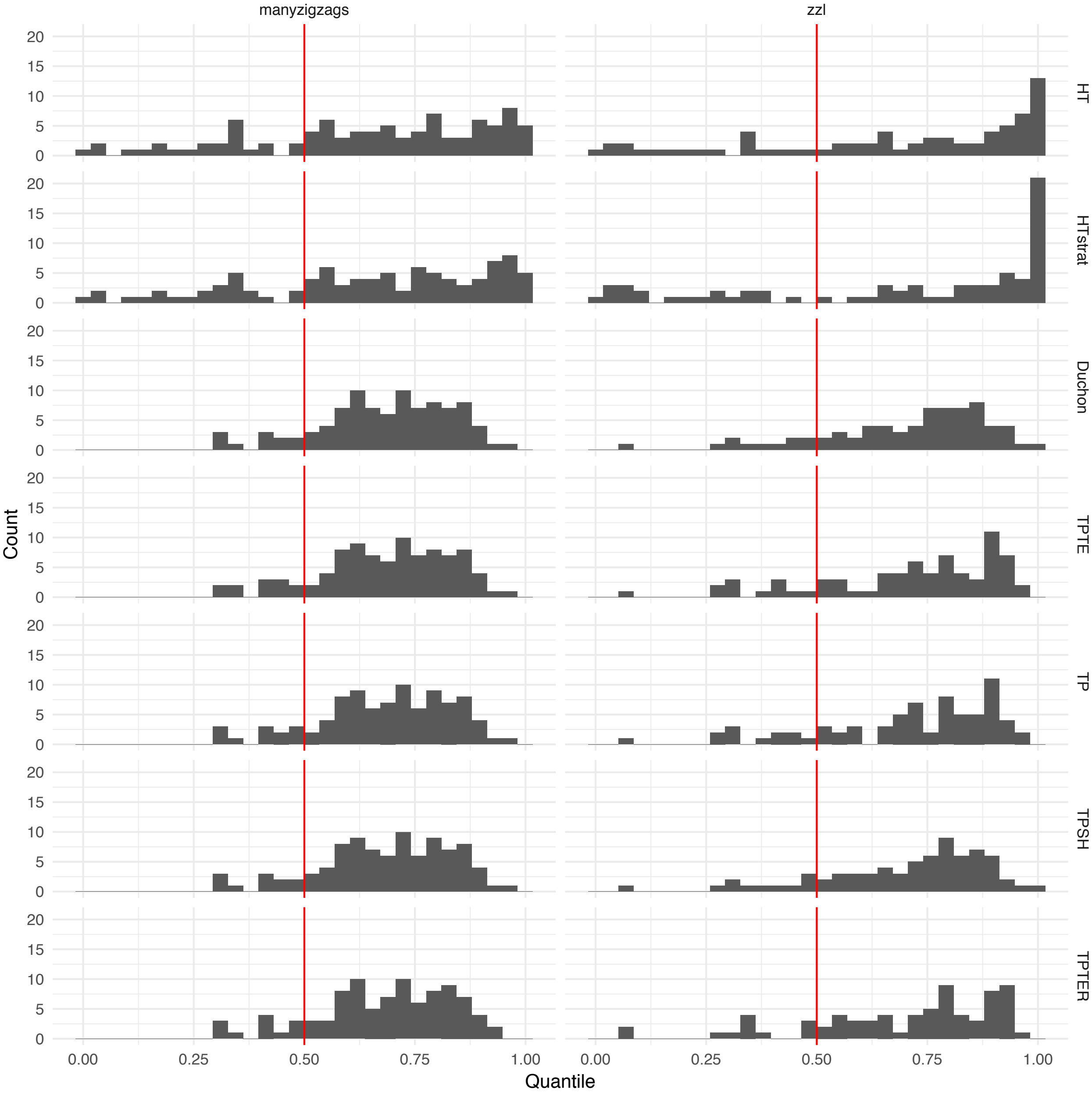

Figure 12: Histograms of the "self-confidence" measure for each model for the flat density surface with a "bad" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). Here we see generally good behaviour in all models (some negative bias across the board, perhaps more severe for the HT-based methods than Figure 5 would have us believe), though the spatial models are slightly positively biased (though more spread out than with the good detectability), which is likely down to model overparamatrisation (smooths should be estimated with zero effects, which is hard).

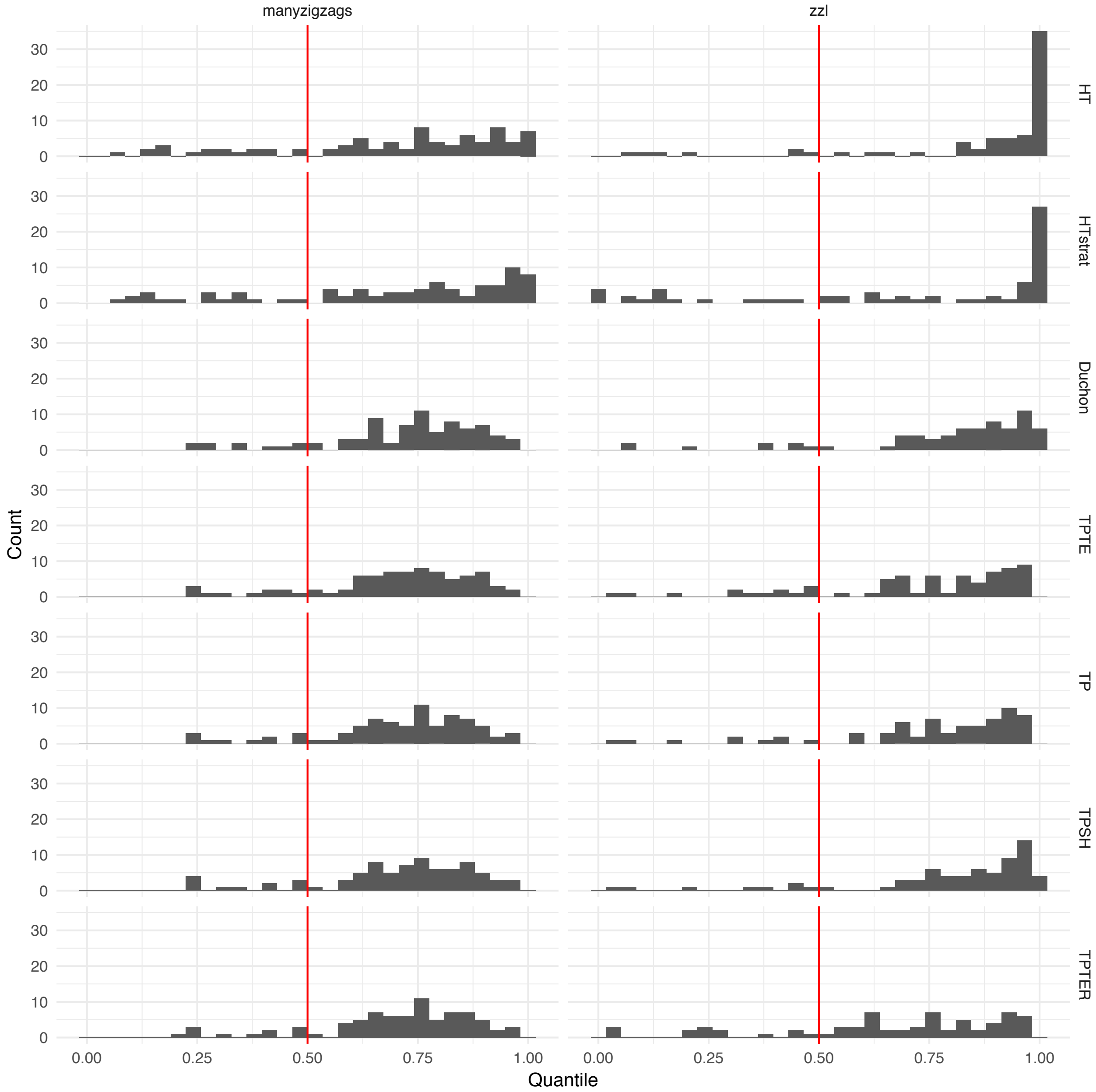

Figure 13: Histograms of the "self-confidence" measure for each model for the left-right density surface with a "bad" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). As in Figure 5 we see negative bias in the results, though again this is more extreme for the HT-based methods for the "zzl" design.

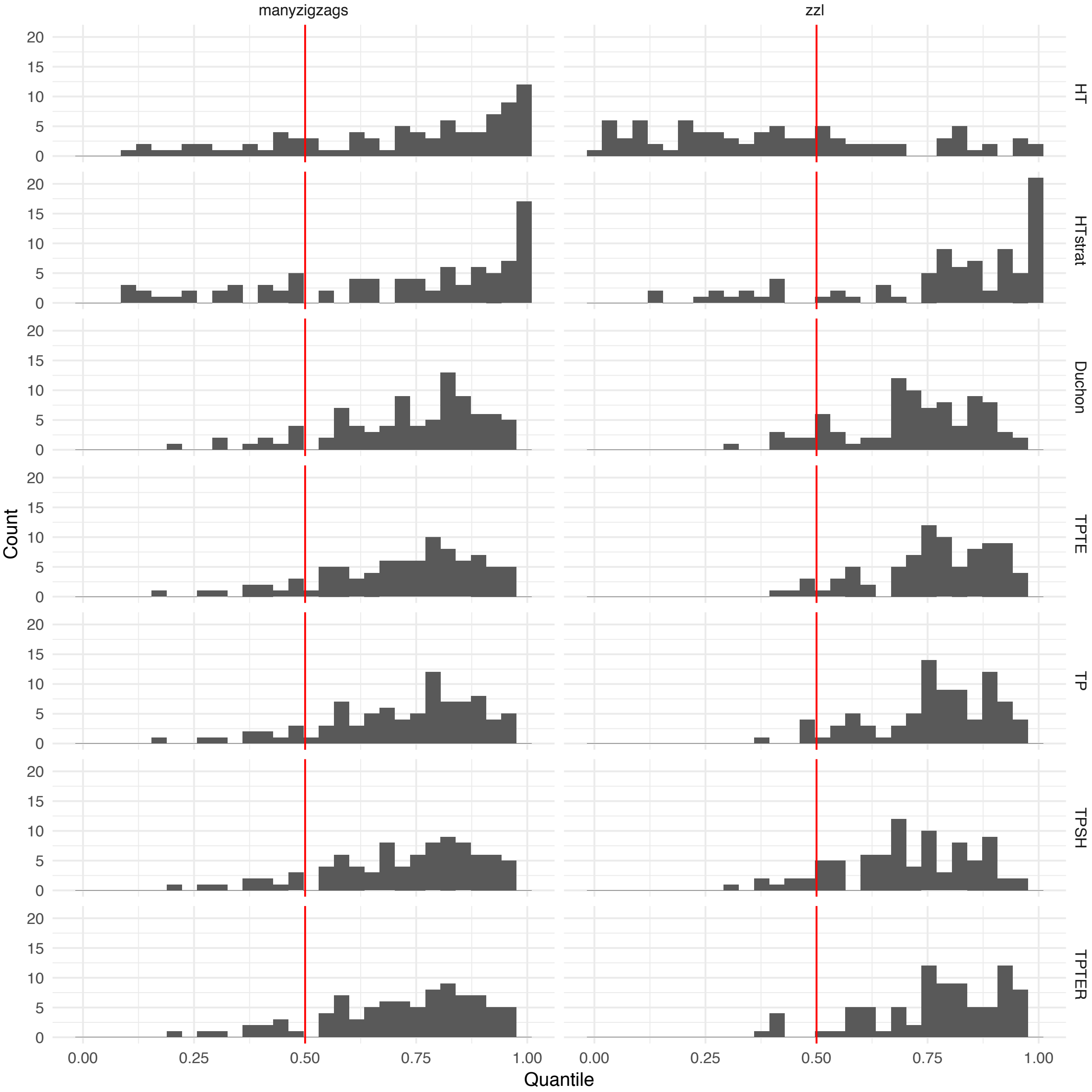

Figure 14: Histograms of the "self-confidence" measure for each model for the right-left density surface with a "bad" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). As in Figure 5 we see negative bias in the results, though again this is more extreme for the HT-based stratified method for the "zzl" design, for the unstratified HT estimate we see positive bias for the "zzl" design as the model thinks that there is low detectability everywhere and hence underestimates detectability, overestimating abundance.

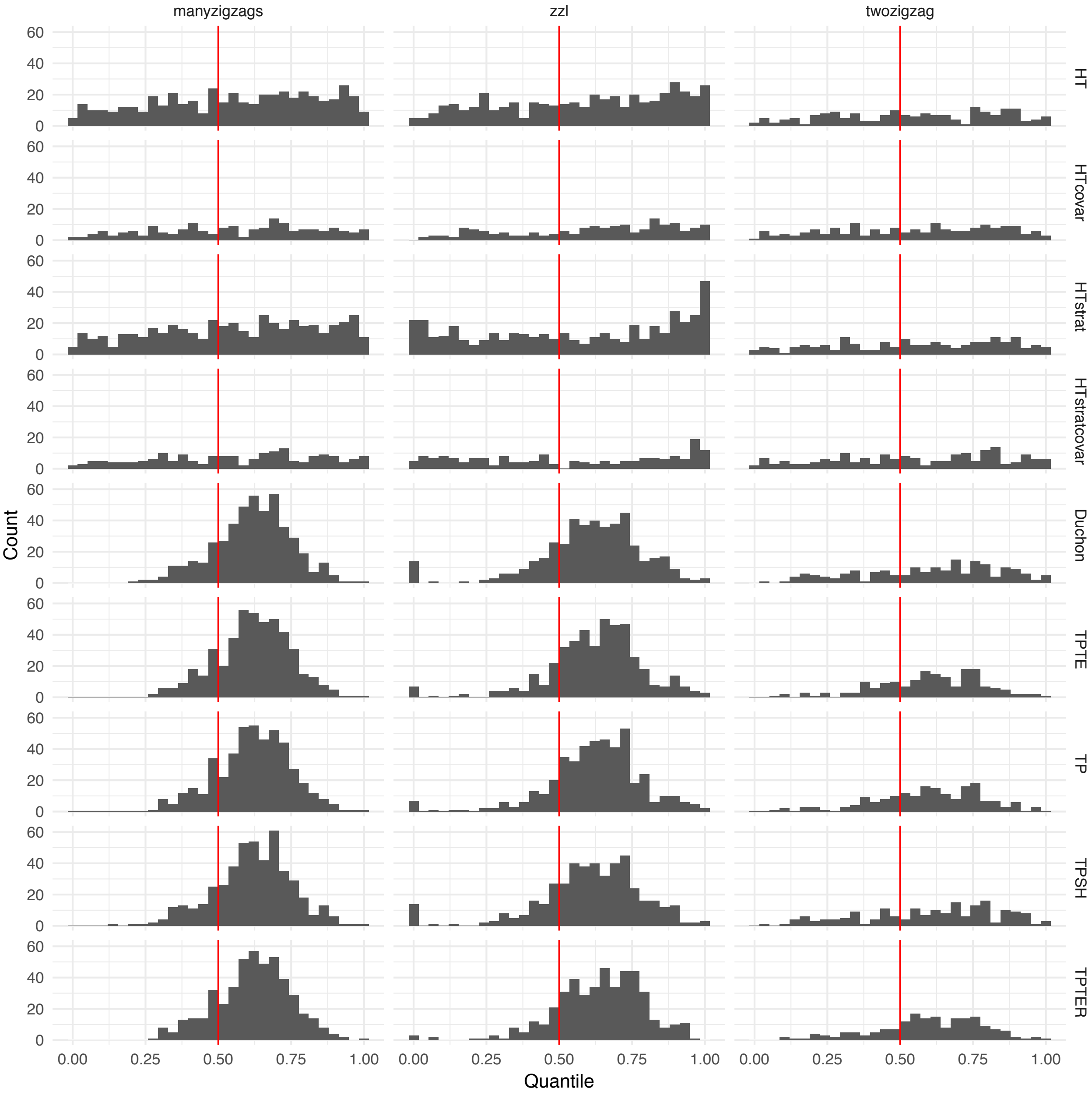

Figure 15: Histograms of the "self-confidence" measure for each model for the flat density surface when the weather covariate was used to generate data. Here we see generally good behaviour in all models, though the spatial models are slightly positively biased, which is likely down to model overparamatrisation (smooths should be estimated with zero effects, which is hard). Rows are the models (see Section 4.4) and columns give the design (see Section 4.3).

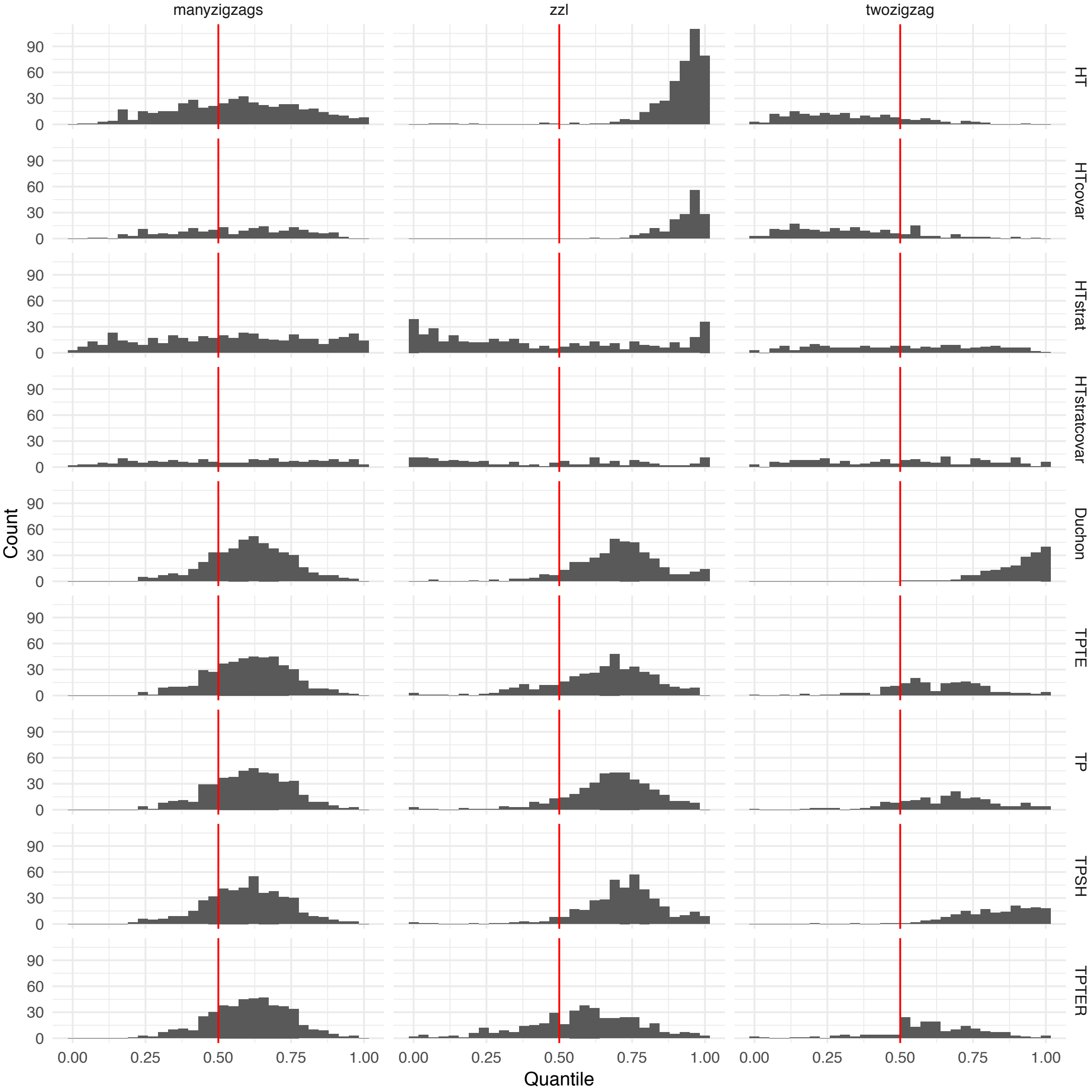

Figure 16: Histograms of the "self-confidence" measure for each model for the left-right density surface when the weather covariate was used to generate data. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). For the "zzl" design the stratification (which acts as spatial and covariate stratification) is quite important, giving much better results (again predicated on the correct stratification being selected). Spatial model results seem reasonable with the notable exception of the Duchon and shrinkage (TPSH) methods for the "twozigzag" design.

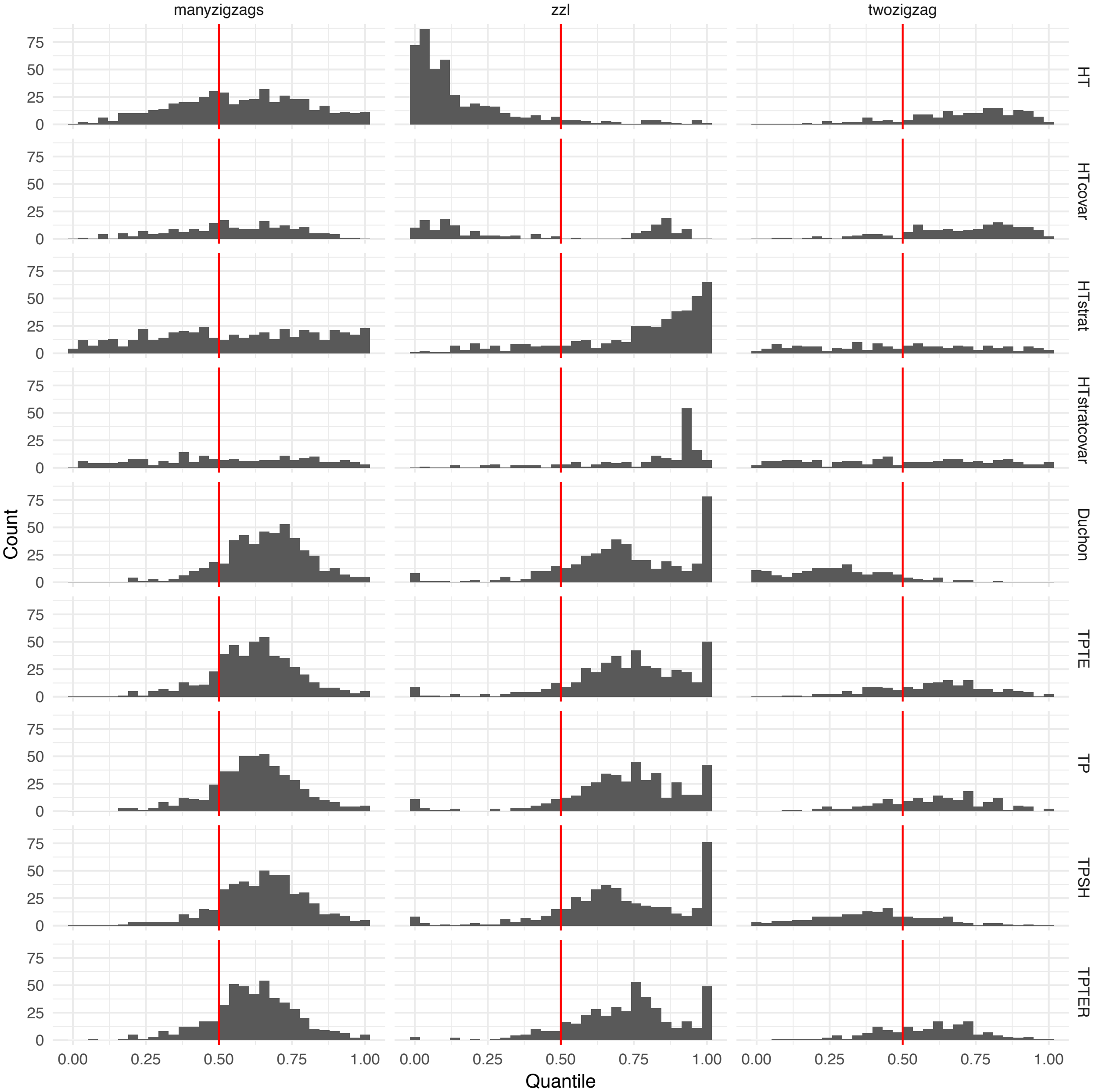

Figure 17: Histograms of the "self-confidence" measure for each model for the right-left density surface when the weather covariate was used to generate data. Rows are the models (see Section 4.4) and columns give the design (see 4.3). Here the "zzl" design allocates more effort to the low density areas, since there are many detections in the high effort area, the HT estimator overestimates abundance, the stratification improves this a little but causes underestimation, even including the covariate doesn't improve things much. We again see more erratic behaviour from the Duchon and shrinkage smoothers (TPSH) for the "twozigzag" design.

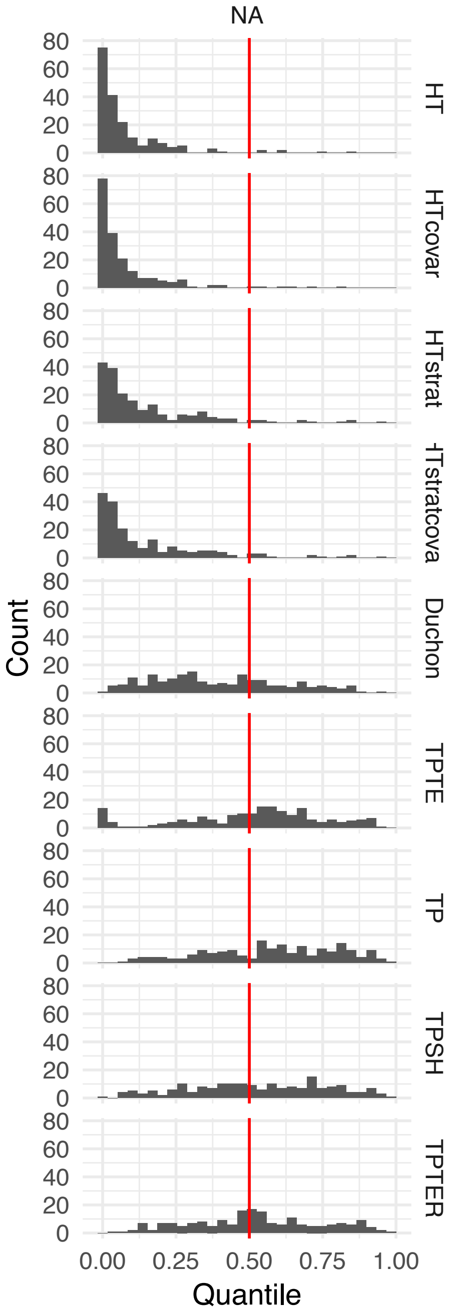

Figure 18: Histograms of the “self-confidence” measure for each model for the diagonal density surface when the weather covariate was used to generate data. Rows are the models (see Section 4.4) and columns give the design (see 4.3). As in Figure 8, the HT-based methods perform poorly, generally overestimating abundance, the spatial models (now not suffering from model overparamatrisation) perform very well, exhibiting the “conservative” behaviour referred to in Figure 4.

References

Augustin, N. H., M. Musio, et al. (2009). "Modeling Spatiotemporal Forest Health Monitoring Data". In: *Journal of the American Statistical Association* 104.487, pp. 899–911.

Augustin, N. H., E.-A. Sauleau, and S. N. Wood (2012). "On quantile quantile plots for generalized linear models". In: *Computational Statistics and Data Analysis* 56.8, pp. 2404–2409.

Augustin, N. H., V. M. Trenkel, et al. (2013). "Space-time modelling of blue ling for fisheries stock management". In: *Environmetrics* 24.2, pp. 109–119.

Boor, C. de (1978). *A Practical Guide to Splines*. Springer.

Borchers, D. L., S. T. Buckland, P. W. Goedhart, et al. (1998). "Horvitz-Thompson Estimators for Double-Platform Line Transect Surveys". In: *Biometrics* 54.4, p. 1221.

Borchers, D. L., S. T. Buckland, and W. Zucchini (2002). *Estimating Animal Abundance: Closed populations*. Springer.

Borchers, D. L. and K. P. Burnham (2004). "General formulation for distance sampling". In: *Advanced Distance Sampling*. Oxford University Press, Oxford, UK, pp. 6–30.

Borchers, D. L., J. L. Laake, et al. (2005). "Accommodating Unmodeled Heterogeneity in Double-Observer Distance Sampling Surveys". In: *Biometrics* 62.2, pp. 372–378.

Borchers, D. L., W. Zucchini, and R. M. Fewster (1998). "Mark-Recapture Models for Line Transect Surveys". In: *Biometrics* 54.4, p. 1207.

Buckland, S. T., D. R. Anderson, K. P. Burnham, D. L. Borchers, et al. (2001). *Introduction to Distance Sampling*. Estimating Abundance of Biological Populations. Oxford University Press, Oxford, UK.

Buckland, S. T., D. R. Anderson, K. P. Burnham, J. L. Laake, et al. (2004). *Advanced Distance Sampling*. Estimating abundance of biological populations. Oxford University Press, Oxford, UK.

Buckland, S. T., K. B. Newman, et al. (2004). "State-space models for the dynamics of wild animal populations". In: *Ecological Modelling* 171.1-2, pp. 157–175.

Buckland, S. T., E. A. Rexstad, et al. (2015). *Distance Sampling: Methods and Applications*. Methods in Statistical Ecology. Springer International Publishing.

Burnham, K. P., D. R. Anderson, and J. L. Laake (1980). *Estimation of density from line transect sampling of biological populations*.

Burt, M. L. et al. (2014). "Using mark-recapture distance sampling methods on line transect surveys". In: *Methods in Ecology and Evolution* 5.11, pp. 1180–1191.

Chandler, R. E. (2005). "On the use of generalized linear models for interpreting climate variability". In: *Environmetrics* 16.7, pp. 699–715.

Fewster, R. M. et al. (2009). "Estimating the Encounter Rate Variance in Distance Sampling". In: *Biometrics* 65.1, pp. 225–236.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Taylor & Francis.

Hedley, S. L. and S. T. Buckland (2004). "Spatial models for line transect sampling". In: *Journal of Agricultural, Biological, and Environmental Statistics* 9.2, pp. 181–199.

Innes, S. et al. (2002). "Surveys of belugas and narwhals in the Canadian High Arctic in 1996". In: *NAMMCO Scientific Publications* 4, pp. 169–190.

Marra, G. and S. N. Wood (2011). "Practical variable selection for generalized additive models". In: *Computational Statistics and Data Analysis* 55.7, pp. 2372–2387.

Miller, D. L., M. L. Burt, et al. (2013). "Spatial models for distance sampling data: recent developments and future directions". In: *Methods in Ecology and Evolution* 4.11, pp. 1001–1010.

Miller, D. L. and S. N. Wood (2014). "Finite area smoothing with generalized distance splines". In: *Environmental and Ecological Statistics* 21.4, pp. 715–731.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.

Thomas, L. et al. (2010). “Distance software: design and analysis of distance sampling surveys for estimating population size”. In: *Journal of Applied Ecology* 47.1, pp. 5–14.

Winiarski, K. J. et al. (2014). “Integrating aerial and ship surveys of marine birds into a combined density surface model: A case study of wintering Common Loons”. In: *The Condor* 116.2, pp. 149–161.

Wood, S. N. (2003). “Thin plate regression splines”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.1, pp. 95–114.

— (2006a). *Generalized Additive Models*. An Introduction with R. CRC Press.

— (2006b). “ON CONFIDENCE INTERVALS FOR GENERALIZED ADDITIVE MODELS BASED ON PENALIZED REGRESSION SPLINES”. In: *Australian & New Zealand Journal of Statistics* 48.4, pp. 445–464.

Wood, S. N., N. Pya, and B. Säfken (2016). “Smoothing parameter and model selection for general smooth models”. In: *Journal of the American Statistical Association*, pp. 1–45.