Shortcomings of Horvitz-Thompson-like estimators for large-scale cetacean abundance estimation

David L. Miller[1] and Mark V. Bravington[2]

8th September 2016

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

[2]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 that these analyses can be inappropriate in simple situations. We reason that if an estimator shows poor properties in a simple situation, then in reality (when distribution, availability or detectability are more complex) 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, Anderson, Burnham, Borchers et al., 2001; Buckland, Anderson, Burnham, Laake et al., 2004; Buckland, Rexstad et al., 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.

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, observed detection-related covariates and estimated parameters), we can estimate the average (over distance) probability of detection for an observation (conditional on observed covariate values), $\hat{p}_i$, by integrating out distance for a given observation i.

2.1 Horvitz-Thompson-like estimators

We first describe the "Horvitz-Thompson-like" estimator (henceforth HT) (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, 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 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, 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 or regions 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 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 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 (though the encounter rate estimator itself is $\hat{N}/L$; Innes et al., 2002).

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 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, 2006) 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, 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].

[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.

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, 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.

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, Anderson, Burnham, Laake et al., 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, 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 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:

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

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

Intuitively, we say that 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 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 HT 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 ($g(0) = 1$). 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) 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. 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's likely that there will be

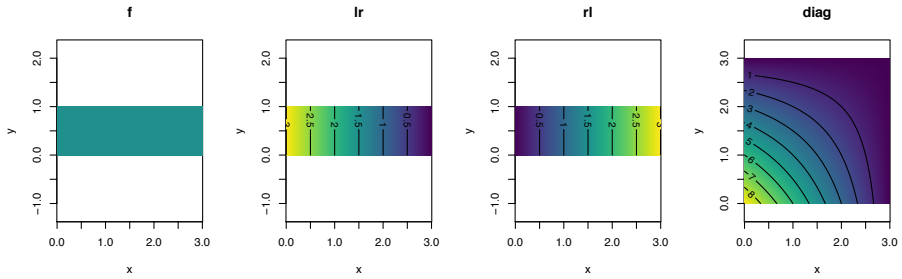

Figure 1: The proposed density surfaces. From left to right: flat, left-right gradient, right-left gradient, diagonal gradient.

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 detection functions was controlled by a logistic function. Due to it's change along the x axis, the covariate is confounded with the "lr" and "rl" density gradients – this is eminently possible in survey data and is a strong reason to always collect some kind of weather-type covariate and include this in models.

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.

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.

Detection function	Scale	Shape	Truncation
"good"	0.025	3	0.05
"bad"	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.

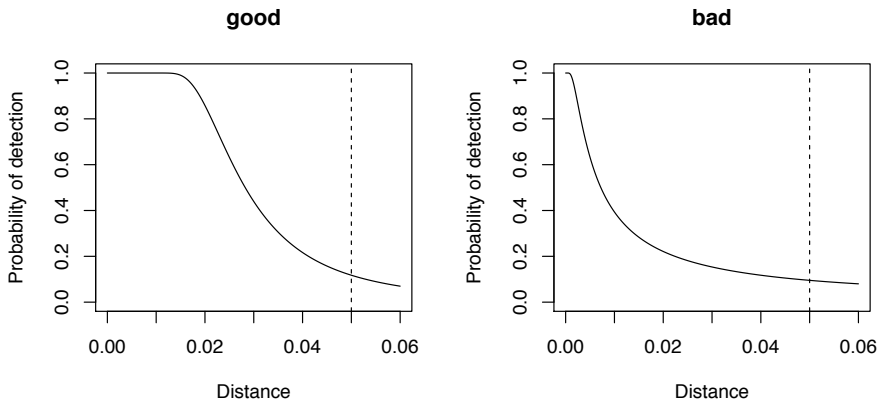

Figure 2: Detection functions used in the simulations. Dashed line indicates the truncation distance used. Parameters are as in Table 1.

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 strata used for stratified HT estimates.

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*

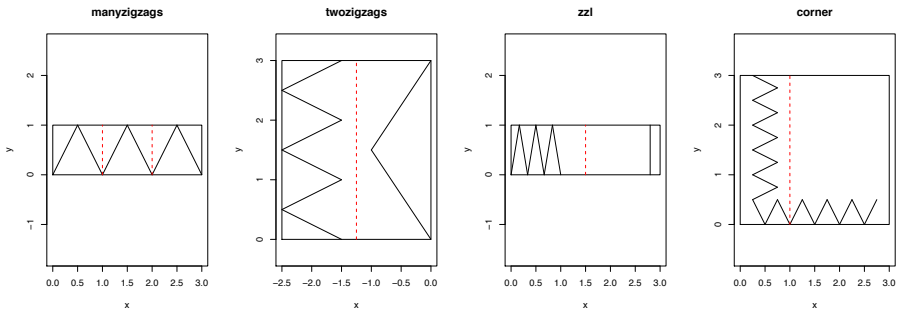

Figure 3: Realised effort for the designs used in the simulations. From left to right: good design with even coverage; an "iffy" design where effort high in one side and low in the other; an "iffy" design where effort is more sporadically allocated; the bad design, with most of the effort in the left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

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^{o} 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 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.

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 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.

4.6.1 Bias

We can simply calculate the bias ($\hat{N} - N_{\text{truth}}$) 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

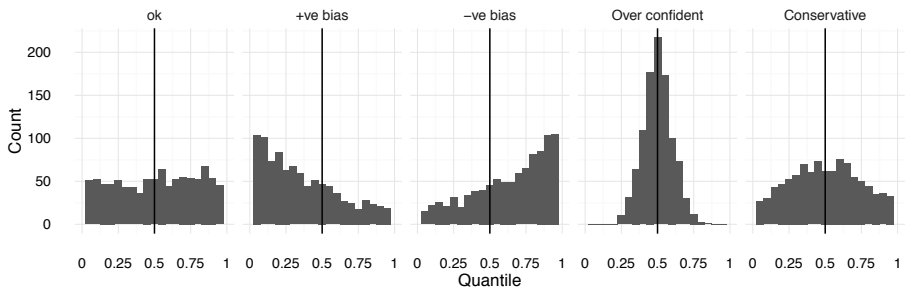

Figure 4: Illustration of the "self-confidence" measure. From left to right: "ok" denotes shows a plot with flat quantile distribution (no problems), "+ve bias" shows the spike at 1 indicating positive bias, "-ve bias" shows a large spike at zero indicating negative bias, "overconfident" shows a spike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

desirable, since 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 about. 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 (which we would like to be estimated as a zero effect) 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 in 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 when density is not as simple.

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 less biased, and more coherent way to think about modelling phenomena that are in 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 misspecified, 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. Even with this constraint, the spatial models performed well. Estimating a flat surface (equivalently estimating almost all of the coefficients of the bivariate smooths to be zero), with uneven sampling of the density surface is very difficult (not to mention an unrealistic situation).

The `ltdesigntester` package developed for this paper is quite general and can be used for any design or density surface. Incorporating more complex detectability covariates is more complicated, but achievable. We encourage investigators to input their current designs and create some simple density surfaces to test how well abundance can be estimated.

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.

References

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.

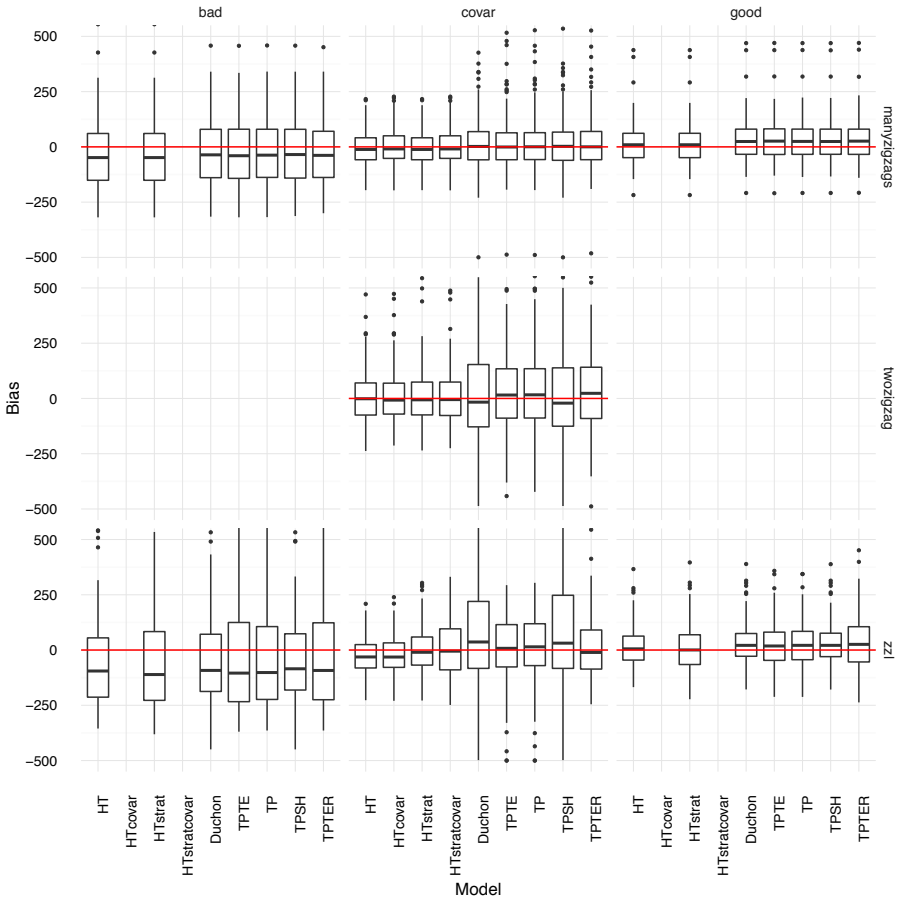

Figure 5: Bias in abundance for each of the models per simulation scenario for the flat density. Columns give the detection function type (see Section4.2) and rows give the design used (see Section 4.3). With a flat density all models perform relatively well, even when the detectability is low (though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

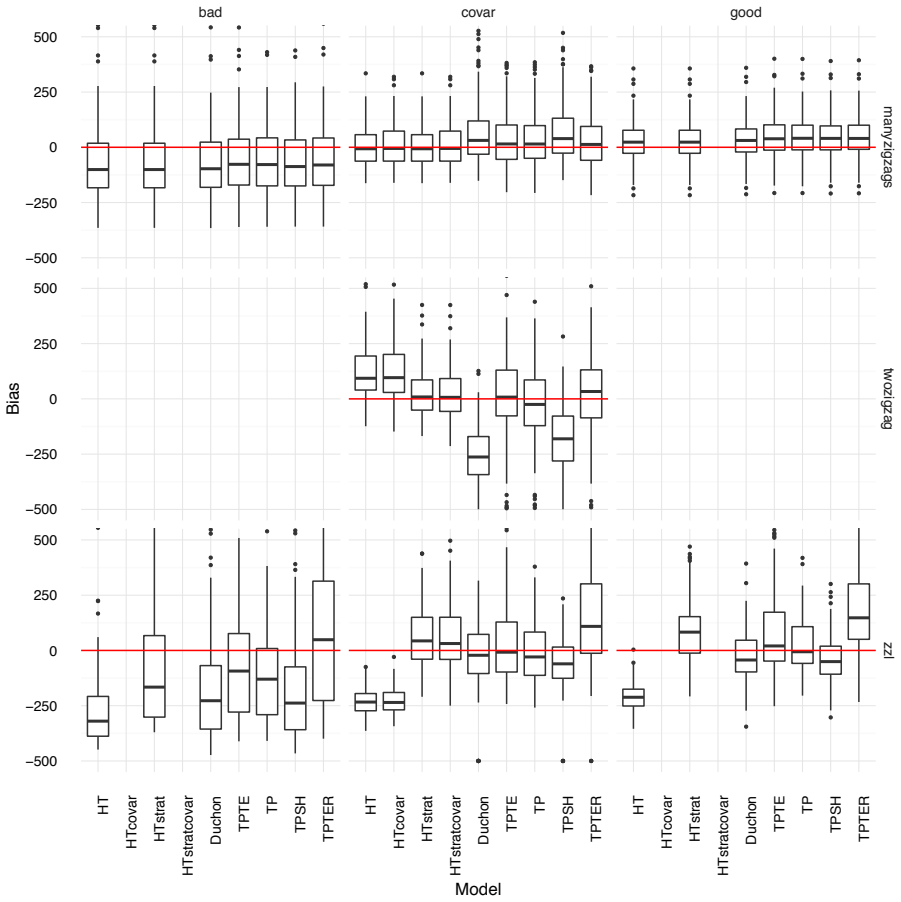

Figure 6: Bias in abundance for each of the models per simulation scenario for the left-right gradient density. Columns give the detection function type (see Section4.2) and rows give the design used (see Section 4.3). In this case for the "zzl" design, most of the effort is concentrated on the higher 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).

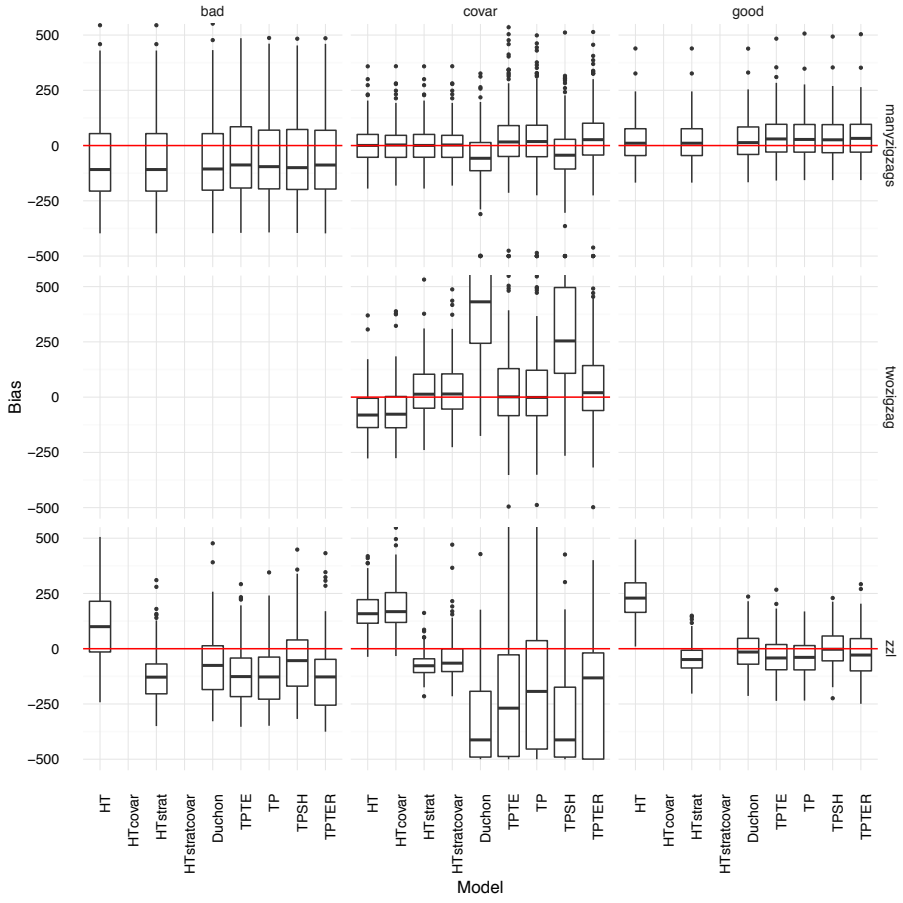

Figure 7: Bias in abundance for each of the models per simulation scenario for the right-left gradient density. Columns give the detection function type (see Section4.2) and rows give the design used (see Section 4.3). Here the "zzl" design allocates more effort to the low density areas, this leads to 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.

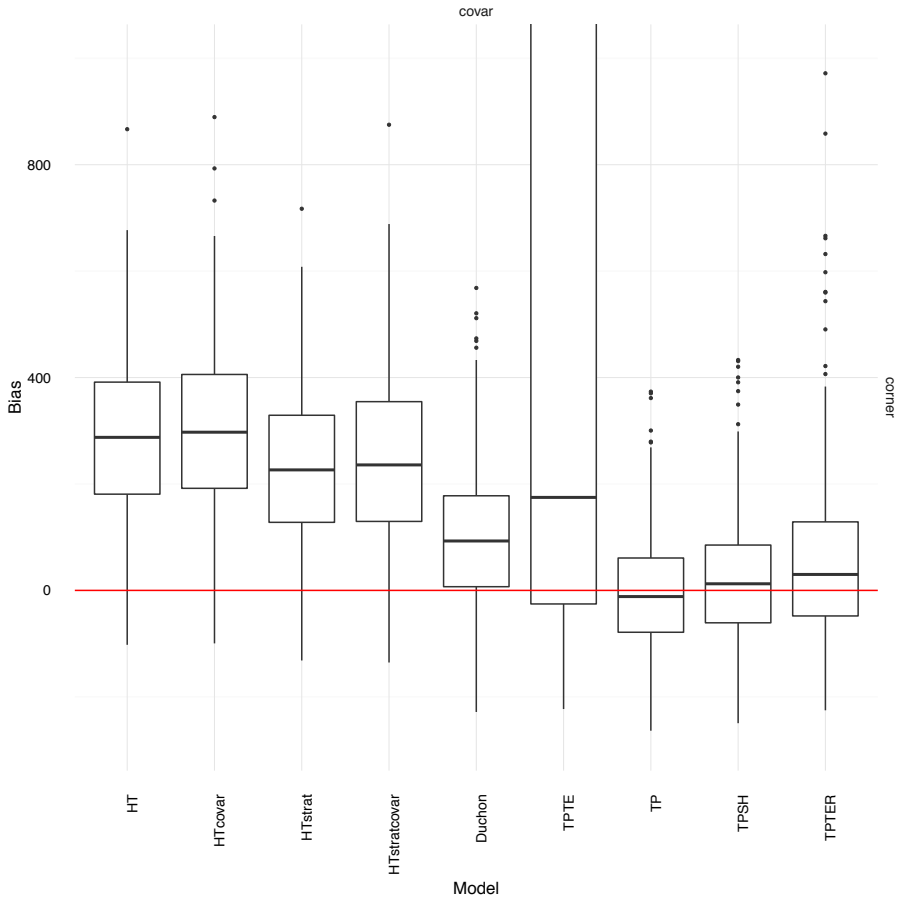

Figure 8: Bias in abundance for each of the models per simulation scenario for the diagonal gradient density. In this case the stratification of the corner design is somewhat arbitrary and there is a large extrapolation to the top right of the survey area (see Figure 3) where the flat density assumption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

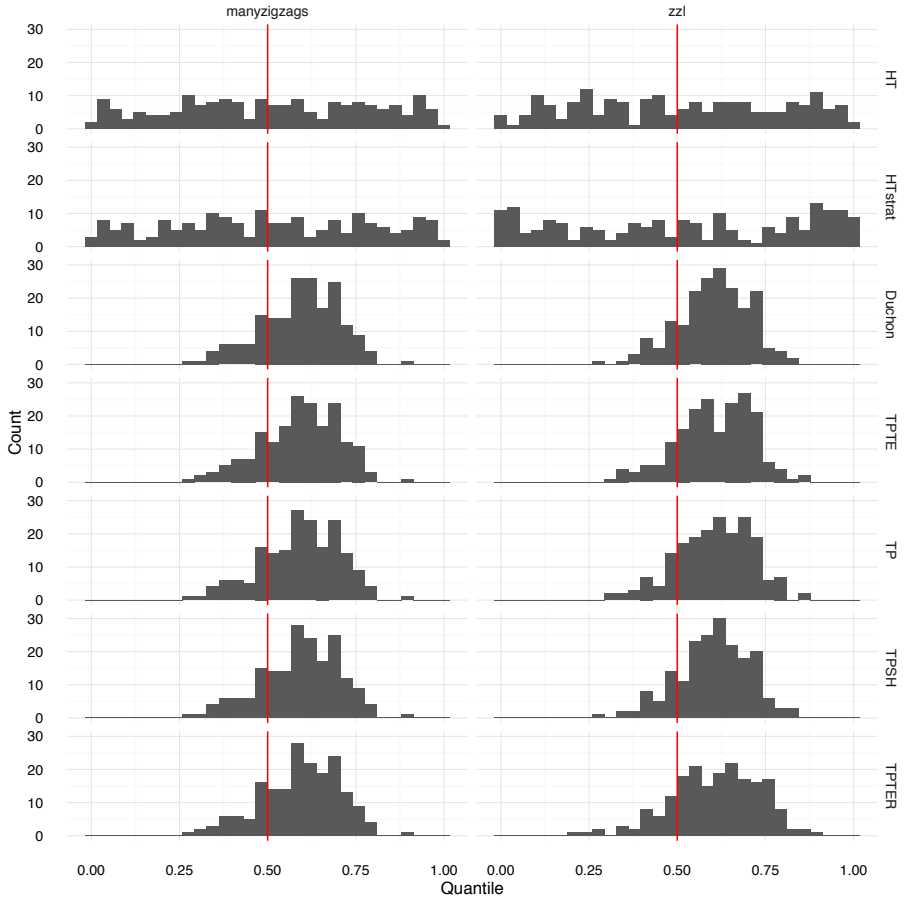

Figure 9: Histograms of the "self-confidence" measure for each model for the flat density surface with a "good" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). Here we see generally good behaviour in all models, though the spatial models are slightly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

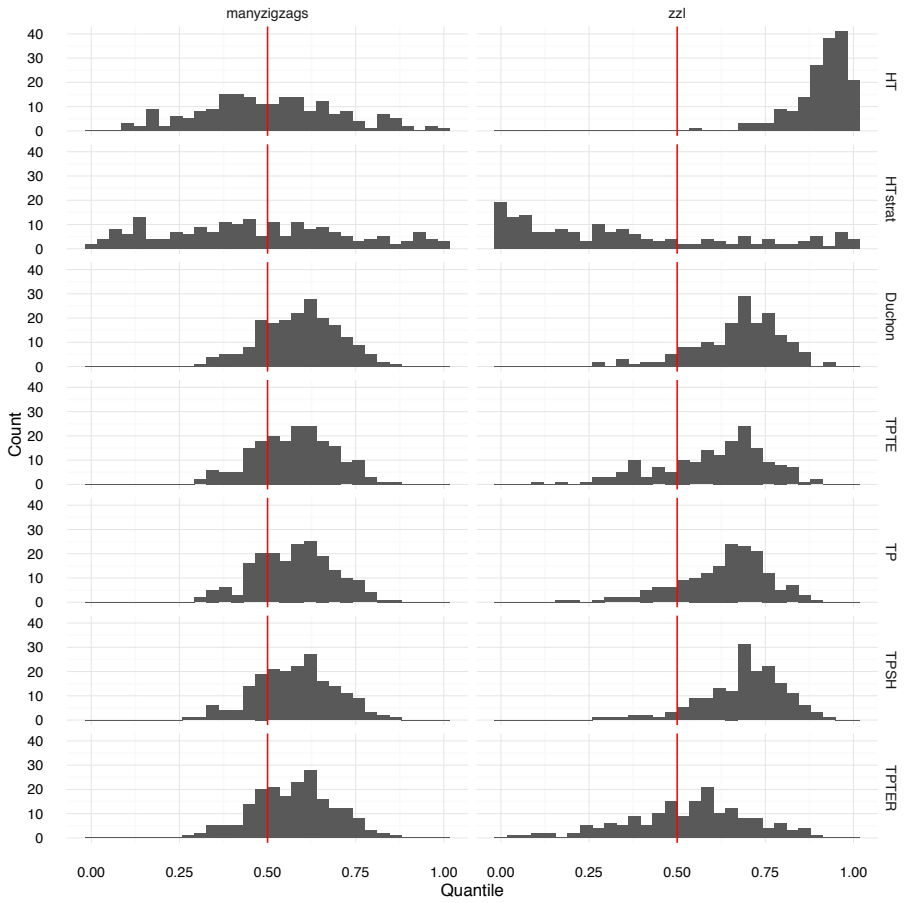

Figure 10: Histograms of the "self-confidence" measure for each model for the left-right density surface with a "good" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). In this case for the "zzl" design, most of the effort is concentrated on the 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).

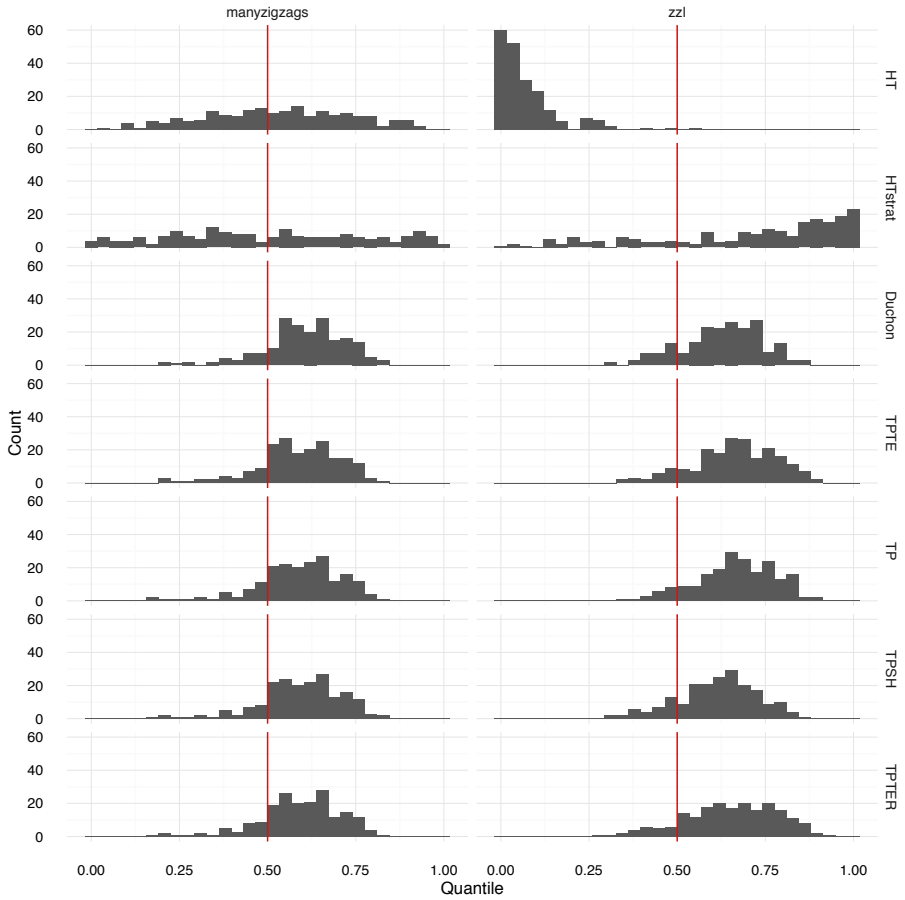

Figure 11: Histograms of the "self-confidence" measure for each model for the right-left density surface with a "good" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). Here the "zzl" design allocates more effort to the low density areas, as in Figure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

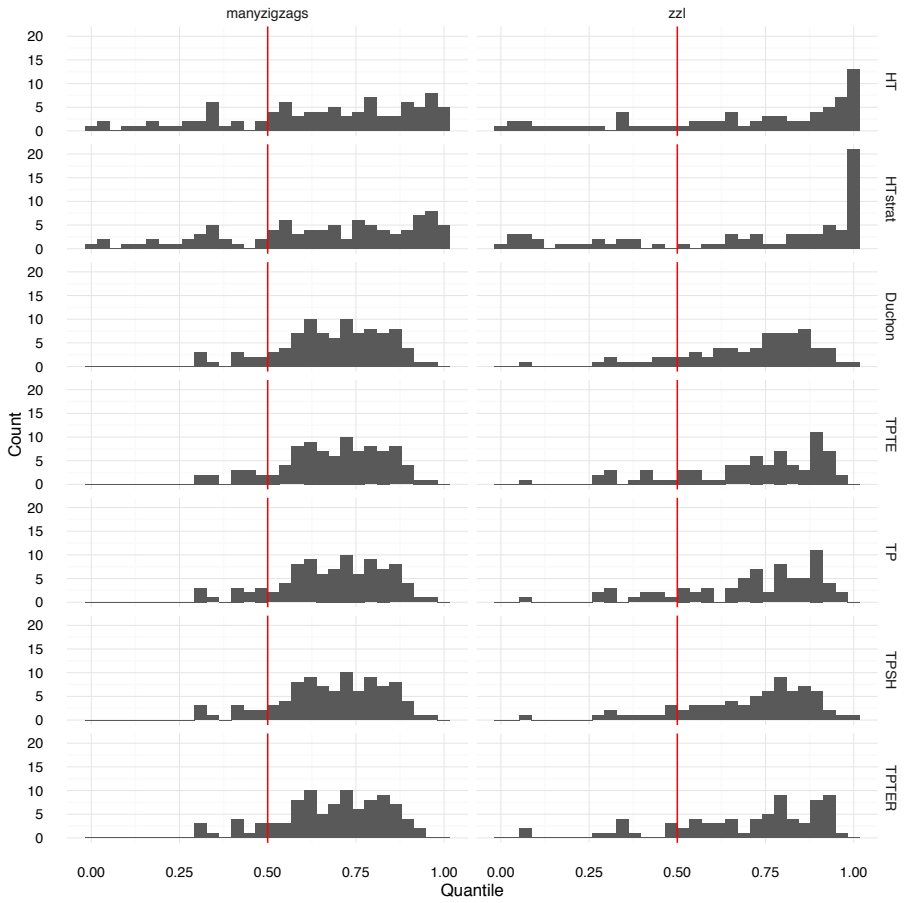

Figure 12: Histograms of the "self-confidence" measure for each model for the flat density surface with a "bad" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). Here we see generally good behaviour in all models (some negative bias across the board, 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 misspecification (smooths should be estimated with zero effects, which is hard).

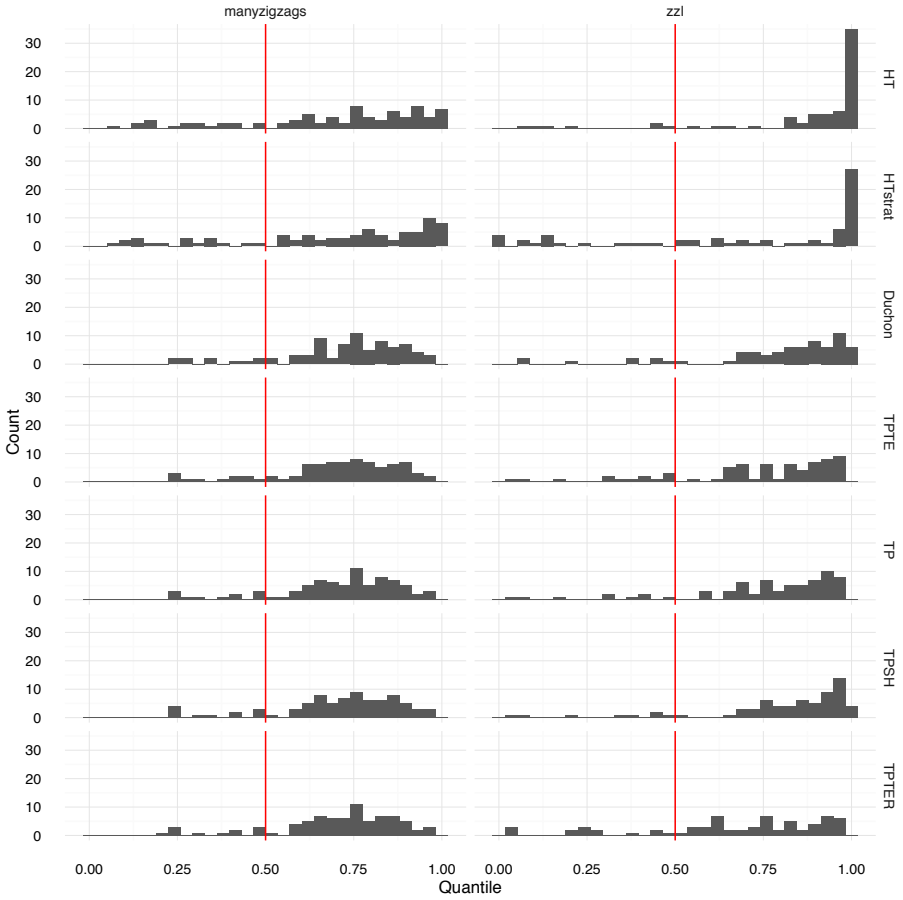

Figure 13: Histograms of the "self-confidence" measure for each model for the left-right density surface with a "bad" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). As in Figure 5 we see negative bias in the results, though again this is more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

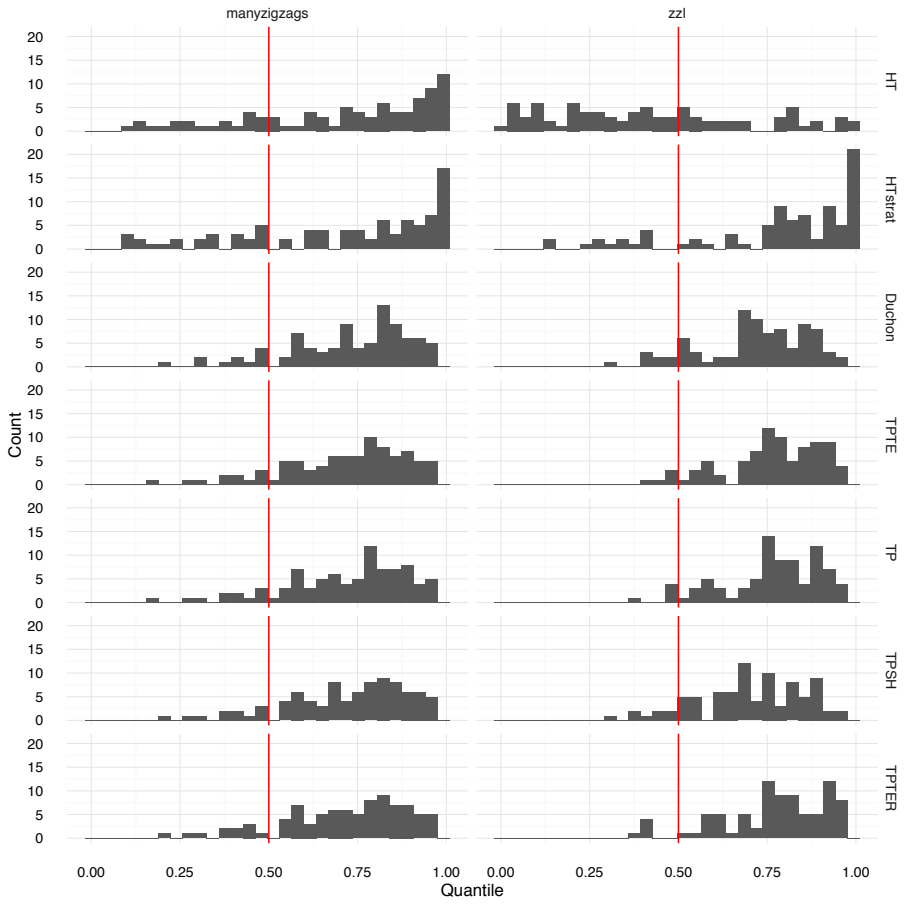

Figure 14: Histograms of the "self-confidence" measure for each model for the right-left density surface with a "bad" detection function. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). As in Figure 5 we see negative bias in the results, though again this is more 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.

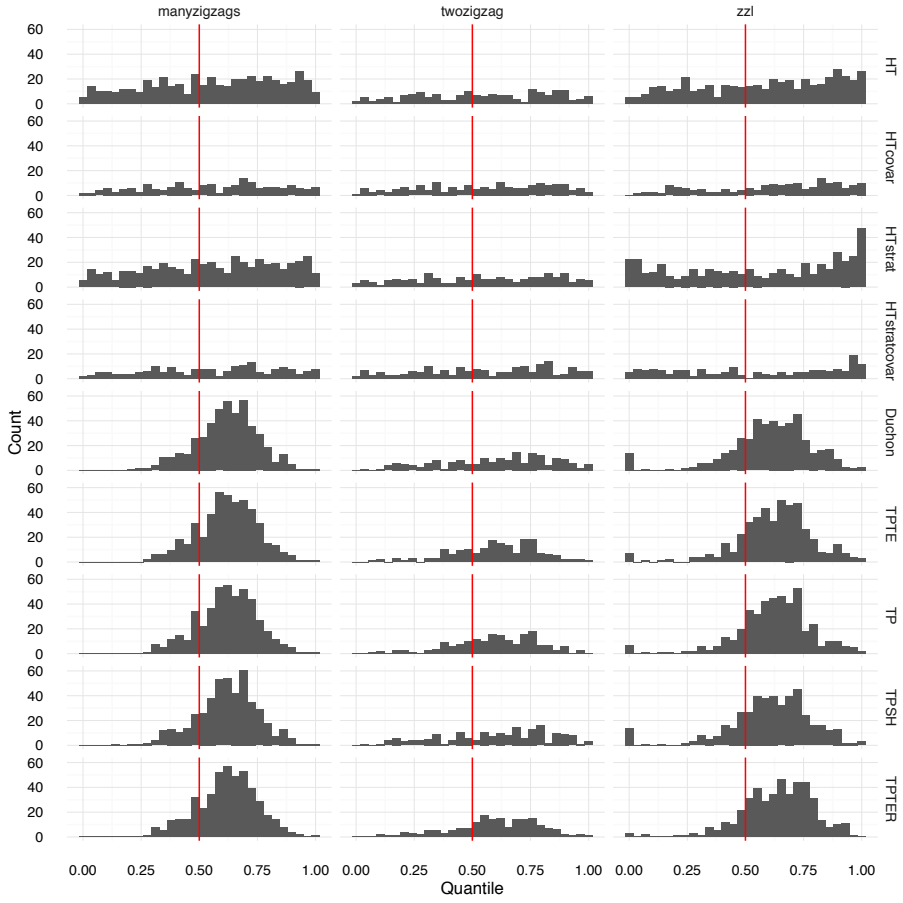

Figure 15: Histograms of the "self-confidence" measure for each model for the flat density surface when the weather covariate was used to generate data. Here we see generally good behaviour in all models, though the spatial models are slightly positively biased, which is likely down to model misspecification (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).

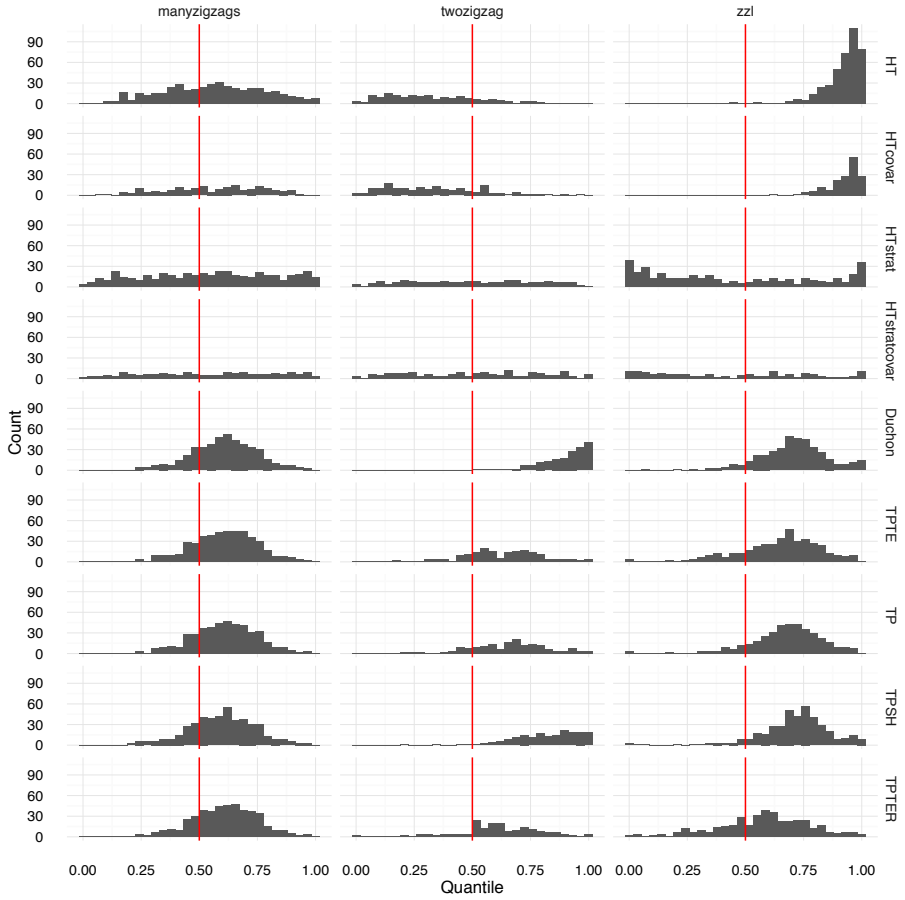

Figure 16: Histograms of the "self-confidence" measure for each model for the left-right density surface when the weather covariate was used to generate data. Rows are the models (see Section 4.4) and columns give the design (see Section 4.3). For the "zzl" design the stratification (which acts as spatial 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.

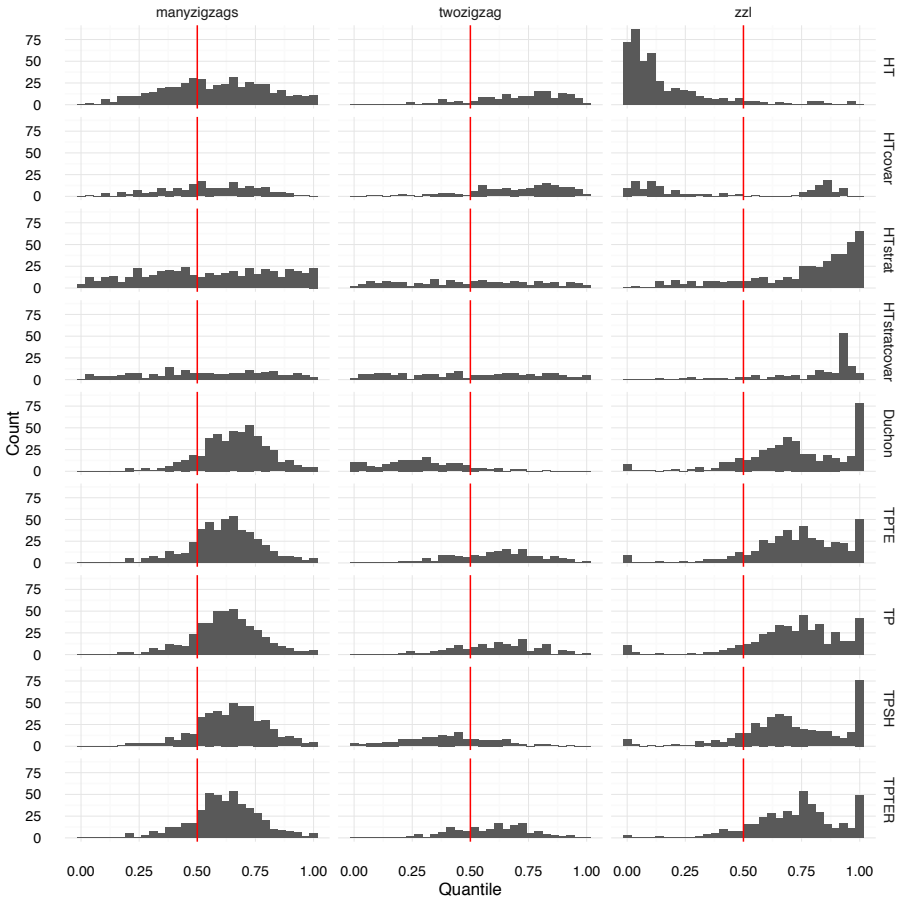

Figure 17: Histograms of the "self-confidence" measure for each model for the right-left density surface when the weather covariate was used to generate data. Rows are the models (see Section 4.4) and columns give the design (see 4.3). Here the "zzl" design allocates more effort to the low density areas, 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.

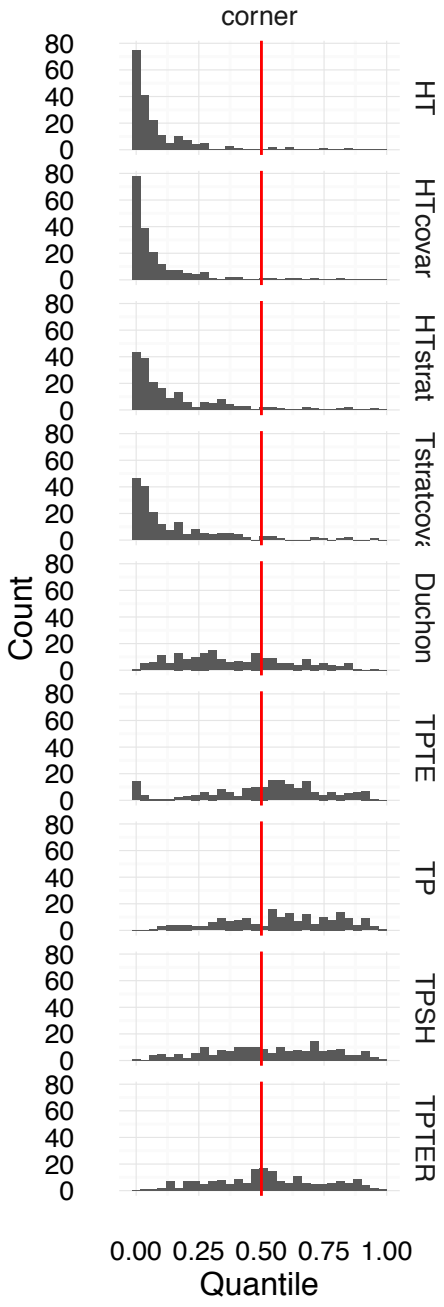

Figure 18: Histograms of the "self-confidence" measure for each model for the diagonal density surface when the weather covariate was used to generate data. Rows are the models (see Section 4.4) and columns give the design (see 4.3). As in Figure 8, the HT-based methods perform poorly, generally overestimating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

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.

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.

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.

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

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.