Running title: Spatial models for distance sampling
Number of words: ∼4925
Number of tables: 0
Number of figures: 6
Number of references: 46

Spatial models for distance sampling data: recent developments and future directions

David L. Miller[1*], M. Louise Burt[2],
Eric A. Rexstad[2], Len Thomas[2].

1. Department of Natural Resources Science, University of Rhode Island, Kingston, Rhode Island 02881, USA
2. Centre for Research into Ecological and Environmental Modelling, The Observatory, University of St. Andrews, St. Andrews KY16 9LZ, UK

*Correspondence author. `dave@ninepointeightone.net`

Date of submission: 4th July 2013

Summary

1. Our understanding of a biological population can be greatly enhanced by modelling their distribution in space and as a function of environmental covariates. Such models can be used to investigate the relationships between distribution and environmental covariates as well as reliably estimate abundances and create maps of animal/plant distribution.

2. Density surface models consist of a spatial model of the abundance of a biological population which has been corrected for uncertain detection via distance sampling methods.

3. We review recent developments in the field and consider the likely directions of future research before focussing on a popular approach based on generalized additive models. In particular we consider spatial modelling techniques that may be advantageous to applied ecologists such as quantification of uncertainty in a two-stage model and smoothing in areas with complex boundaries.

4. The methods discussed are available in an R package developed by the authors (`dsm`) and are largely implemented in the popular Windows software Distance.

Keywords: abundance estimation, Distance software, generalized additive models, line transect sampling, point transect sampling, population density, spatial modelling, wildlife surveys

Introduction

When surveying biological populations it is increasingly common to record spatially referenced data, for example: coordinates of observations, habitat type, elevation or (if at sea) bathymetry. Spatial models allow for vast databases of spatially-referenced data (e.g. OBIS-SEAMAP, Halpin *et al.*, 2009) to be harnessed, enabling investigation of interactions between environmental covariates and population densities. Mapping the spatial distribution of a population can be extremely useful, especially when communicating results to non-experts. Recent advances in both methodology and software have made spatial modelling readily available to the non-specialist (e.g., Wood, 2006; Rue *et al.*, 2009). Here we use the term "spatial model" to refer to any model that includes any spatially referenced covariates, not only those models that include location as a covariate. This article is concerned with combining spatial modelling techniques with distance sampling (Buckland *et al.*, 2001, 2004).

Distance sampling extends plot sampling to the case where detection is not certain. Observers move along lines or visit points and record the distance from the line or point to the object of interest (y). These distances are used to estimate the *detection function*, $g(y)$ (for example, Fig. 1), by modelling the decrease in detectability with increasing distance from the line or point (conventional distance sampling, CDS). The detection function may also include covariates (multiple covariate distance sampling, MCDS; Marques *et al.*, 2007) which affect the scale of the detection function. From the fitted detection function, the average probability of detection can be estimated by integrating out distance. The estimated average probability that an animal is detected given that it is in the area covered by the survey, $\hat{p}_i$, can then be used to estimate abundance as

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

where A is the area of the study region, a is the area covered by the survey (i.e., the sum of the areas of all of the strips/circles) and the summation takes place over the n observed clusters, each of size s_i (if individuals are observed, $s_i = 1 \forall i$) (Buckland *et al.*, 2001, Chapter 3). Often up to half the observations in a plot sampling data set are discarded to ensure the assumption of certain detection is met. In contrast, distance sampling uses observations that would have been discarded to model detection (although typically some detections are discarded beyond a given *truncation distance* during analysis).

Estimators such as eqn (1) rely on the design of the study to ensure that abundance estimates over the whole study area (scaling up from the covered region) are valid. This article focusses on *model-based* inference to extrapolate to a larger study area. Specifically, we consider the use of spatially explicit models to investigate the response of biological populations to biotic and abiotic covariates that vary over the study region. A spatially-explicit model can explain the between-transect variation (which is often a large component of the variance in design-based estimates) and so using a model-based approach can lead to smaller variance in estimates of abundance than design-based estimates. Model-based inference also enables the use of data from opportunistic surveys, for example, incidental data arising from "ecotourism" cruises (Williams *et al.*, 2006).

Our aims in creating a spatial model of a biological population are usually two-fold: (i) estimating overall abundance and (ii) investigating the relationship between abundance and environmental covariates. As with any predictions that are outside the range of the data, one should heed the usual warnings regarding extrapolation. For example, if a model contains elevation as a covariate, predictions at high, unsampled elevations are unlikely to be reliable. Frequently, maps of abundance or density are required and any spurious predictions can be visually assessed, as well as by plotting a histogram of the predicted values. A sensible definition of the region of interest avoids prediction outside the range of the data.

In this article we review the current state of spatial modelling of detection-corrected count data, illustrating some recent developments useful to applied ecologists. The methods discussed have been available in Distance software (Thomas *et al.*, 2010) for some time but the recent advances covered here have been implemented in a new R package, `dsm` (Miller *et al.*, 2013) and are to be incorporated into Distance.

Throughout this article a motivating data set is used to illustrate the methods. These data are sightings of pantropical spotted dolphins (*Stenella attenuata*) during April and May of 1996 in the Gulf of Mexico. Observers aboard the NOAA vessel Oregon II recorded sightings and environmental covariates (see `http://seamap.env.duke.edu/dataset/25` for survey details). A complete example analysis is provided in Appendix A. The data used in the analysis are available in the `dsm` package and Distance.

The rest of the article reviews approaches for the spatial modelling of distance sampling data before focussing on the density surface modelling approach of Hedley & Buckland (2004) to estimate abundance and uncertainty. We then describe recent advances and provide practical advice regarding model fitting, formulation and checking. Finally we discuss future directions for research in spatially modelling detection-corrected count data.

Approaches to spatial modelling of distance sampling data

Modelling of spatially referenced distance sampling data is equivalent to modelling spatially-referenced count data, with the additional information provided by collecting distances to account for imperfect detection. We review recent efforts to model such data; some consist of two steps (correction for imperfect detection, then spatial modelling), while others jointly estimate the relevant parameters.

Two-stage approaches

The focus of this article is the "count model" of Hedley & Buckland (2004), we will henceforth refer to this approach as *density surface modelling* (DSM). Modelling proceeds in two steps: a detection function is fitted to the distance data to obtain detection probabilities for clusters (flocks, pods, etc.) or individuals. Counts are allocated to corresponding segments (contiguous transect sections). A generalised additive model (GAM; e.g. Wood, 2006) is then constructed with the per-segment counts as the response with either counts or segment areas corrected for detectability (see *Density surface modelling*, below). GAMs provide a flexible class of models that include generalized linear models (GLMs; McCullagh & Nelder, 1989) but extend them with the possible addition of splines to create smooth functions of covariates, random effects terms or correlation structures. We cover advances using this approach in *Recent developments*.

Niemi & Fernández (2010) proposed a Bayesian point process approach to spatial abundance modelling. The density of the objects is described by an intensity function, which included spatially-referenced covariates. Model fitting proceeded in two stages: first the detection function was fitted, then the spatial model (via MCMC) assuming the detection function parameters were known, so detection function uncertainty was not incorporated in the spatial model. A marked point process (Cox & Isham, 1980, Section 5.5) could be used to incorporate cluster size information.

Ver Hoef *et al.* (2013) modeled seal populations in the Bering Sea using a Bayesian spatial model, using a detection function to account for uncertain detection and incorporating additional information from a (frequentist) model of seal haul-outs on ice. The detection function and haul-out model corrected the observed density estimates which were modelled using a Bayesian hierarchical model for the spatial component. The Bayesian hierarchical model was itself was split into two parts (i) a presence/absence part to allow modelling of the large number of zeros in the data and (ii) a density part also used to account for spatial autocorrelation. The analysis shows that when extra information is available (such as telemetry data for the haul-out process) additional insight can be derived.

We note that there are many approaches to modelling spatially referenced count data (Oppel *et al.*, 2011, provides an overview of such methods for marine bird modelling). Also worthy of note is the approach of Barry & Welsh (2002) who used a two-stage approach to model presence/absence then spatial pattern (via two GAMs) to account for zero-inflation.

One-stage approaches

Rather than fitting two separate models, some authors have combined the detection function and spatial model fitted (mostly via hierarchical Bayesian methods). The first of these was Royle *et al.* (2004), who estimated the parameters of a specified detection function, formulating an unconditional likelihood per-point/line as a function of the unobserved transect abundances. These unobserved abundances were treated as random effects, integrated out to give a per-transect likelihood as a function of detection function and random effects parameters (linear functions of the environmental covariates). Due to the multinomial nature of the per-transect likelihood proposed, distance data must be allocated to bins (e.g. 0-5m, 5-15m, etc). Chelgren *et al.* (2011) proposed replacing the multinomial per-transect likelihood with a binomial distribution multiplied by a detection function. The binomial term collapses the multinomial bins into a single bin and gives the number of animals detected in the transect, thus allowing the use of exact distances.

The work of Schmidt *et al.* (2011) took a similar approach to Royle & Dorazio (2008), building a presence/absence-type model for clusters, augmenting the data with unobserved clusters. The authors then used a Poisson distribution to model cluster size (using a random effect to incorporate overdispersion), combining these parts gave a model of individual abundance. Conn *et al.* (2012) also used a hierarchical Bayesian model but in terms of abundance rather than density using a super-population/data augmentation approach (as in Link & Barker, 2009). In their formulation, the whole population within the study region is modelled, not just those animals observed during the survey.

Moore & Barlow (2011) adopted a hierarchical Bayesian state-space model, separating the problem into observation and process components. The process component described the underlying population density as it changes over time and space (though the authors only included strata as a spatial component). The observation part of the model then linked the process model to the data via the detection function.

Johnson *et al.* (2010) proposed a point process-based model for distance sampling data. They first assumed that the locations of all individuals in the survey area (not just those observed) form a realisation of a Poisson process. Parameters of the intensity function were then estimated via standard maximum likelihood methods for point processes (Baddeley & Turner, 2000). All parameters were estimated jointly so uncertainty from both the spatial pattern and the detection function was incorporated into variance estimates of the abundance. This also ensured that correlations between the detection function and underlying point process are estimated correctly (and do not falsely inflate or deflate variance estimates). A post-hoc correction factor was used to address overdispersion unmodelled by spatial covariates (i.e. counts that do not follow a Poisson mean-variance relationship).

One- vs. two-stage approaches

Generally very little information is lost by taking a two-stage approach. This is because transects are typically very narrow compared with the width of the study area so, provided no significant density variation takes place "across" the of the lines or within the point, there is no information in the distances about the spatial distribution of animals (this is an assumption of two-stage approaches).

Two-stage approaches are effectively divide and conquer techniques: concentrating on the detection function first, and then given the detection function, fitting the spatial model. One-stage models are more difficult to both estimate and check as both steps occur at once; models are potentially simpler from the perspective of the user and perhaps more mathematically elegant.

Two-stage models have the disadvantage that to accurately quantify model uncertainty one must appropriately combine uncertainty from the detection function and spatial models. This can be challenging; however the alternative of ignoring uncertainty from the detection process (e.g. Niemi & Fernández, 2010) can produce confidence or credible intervals for abundance estimates that have coverage below the nominal level. More information regarding how variance estimation is addressed for DSMs is given in *Recent developments*.

Density surface modelling

This section focuses on modelling the density/abundance estimation stage of the DSM approach introduced previously. Both line and point transects can be used, but if lines are used then they are are split into contiguous *segments* (indexed by j), which are of length l_j. Segments should be small enough such that neither density of objects nor covariate values vary appreciably within a segment (usually making the segments approximately square is sufficient; $2w \times 2w$, where w is the truncation distance). The area of each segment enters the model as (or as part of) an offset: the area of segment j is $A_j = 2wl_j$ and for point j is $A_j = \pi w^2$.

Count or estimated abundance (per segment or point) is then modelled as a sum of smooth functions of covariates (z_{jk} with k indexing the covariates, e.g., location, sea surface temperature, weather conditions; measured at the segment/point level) using a generalized additive model. Smooth functions are modelled as splines, providing flexible unidimensional (and higher-dimensional) curves (and surfaces, etc) that describe the relationship between the covariates and response. Wood (2006) and Ruppert *et al.* (2003) provide more in-depth introductions to smoothing and generalized additive models.

We begin by describing a formulation where only covariates measured per-segment (e.g. habitat, Beaufort sea state) are included in the detection function. We later expand this simple formulation to include observation level covariates (e.g., cluster size, species)

Count as response

The model for the count per segment is:

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

where the f_ks are smooth functions of the covariates and β_0 is an intercept term. Multiplying the segment area (A_j) by the probability of detection ($\hat{p}_j$) gives the *effective area* for segment j. If there are no covariates other than distance in the detection function then the probability of detection is constant for all segments (i.e., $\hat{p}_j = \hat{p}$, $\forall j$). The distribution of n_j can be modelled as an overdispersed Poisson, negative binomial, or Tweedie distribution (see *Recent developments*).

Fig. 2 shows the raw observations of the dolphin data, along with the transect lines, overlaid on the depth data. A half-normal detection function was fitted to the distances and is shown in Fig. 1. Fig. 3 shows a DSM fitted to the dolphin data. The top panel shows predictions from a model where depth was the only covariate, the bottom panel shows predictions where a (bivariate) smooth of spatial location was also included. Comparing the models using GCV score, the latter had a considerably lower score (39.12 vs 48.46) and so would be selected as our preferred model.

As well as simply calculating abundance estimates, relationships between covariates and abundance can be illustrated via plots of marginal smooths. The effect of depth on abundance (on the scale of the link function) for the dolphin data can be seen in Fig. 4.

An alternative to modelling counts is to use the per-segment/circle abundance using distance sampling estimates as the response. In this case we replace n_j by:

$$\hat{N}_j = \sum_{r=1}^{R_j} \frac{s_{jr}}{\hat{p}_j},$$

where R_j is the number observations in segment j and s_{jr} is the size of the r^{th} cluster in segment j (if the animals occur individually then $s_{jr} = 1, \forall j, r$).

The following model is then fitted:

$$\mathbb{E}(\hat{N}_j) = A_j \exp\left[\beta_0 + \sum_k f_k\left(\boldsymbol{z}_{jk}\right)\right],$$

where $\hat{N}_j$, as with n_j, is assumed to follow an overdispersed Poisson, negative binomial, or Tweedie distribution (see *Recent developments*, below). Note that the offset (A_j) is now the area of segment/point rather than effective area of the segment/point. Although $\hat{N}_j$ can always be modelled instead of n_j, it seems preferable to use n_j when possible, as one is then modelling actual (integer) counts as the response rather than estimates. Note that although $\hat{N}_j$ may take non-integer values, this does not present an estimation problem for the response distributions covered here.

DSM with covariates at the observation level

The above models consider the case where the covariates are measured at the segment/point level. Often covariates (z_{ij}, for individual/cluster i and segment/point j) are collected on the level of observations; for example sex or cluster size of the observed object or identity of the observer. In this case the probability of detection is a function of the object (individual or cluster) level covariates $\hat{p}(z_i)$. Object level covariates can be incorporated into the model by adopting the following estimator of the per-segment/point abundance:

$$\hat{N}_j = \sum_{r=1}^{R_j} \frac{s_{jr}}{\hat{p}(z_{rj})}.$$

Density can be modelled rather than abundance by not including an offset, but instead dividing the count (or estimated abundance) by the area of the segment/point (and weighting observations by the segment/point areas). We concentrate on abundance here; see Hedley & Buckland (2004) for further details on modelling density.

Prediction

A DSM can be used to predict abundance over a larger/different area than was originally surveyed. In that case the investigator must create a series of prediction cells over the prediction region. For each cell the covariates included in the DSM must be available; the area of each cell is also required. Having made predictions for each cell, these can be plotted as an abundance map (as in Fig. 3) or, summing over cells, an overall estimate of abundance can be calculated. It is worth noting that using prediction grid cells that are smaller than the resolution of the spatially referenced data has no effect on abundance/density estimates.

Variance estimation

Estimating the variance of abundances calculated using a DSM is not straightforward: uncertainty from the estimated parameters of the detection function must be incorporated into the spatial model. A second consideration is that in a line transect survey, abundances in adjacent segments are likely to be correlated; failure to account for this spatial autocorrelation will lead to artificially low variance estimates and hence misleadingly narrow confidence intervals.

Hedley & Buckland (2004) describe a method of calculating the variance in the abundance estimates using a parametric bootstrap, resampling from the residuals of the fitted model. The bootstrap procedure is as follows.

Denote the fitted values for the model to be $\hat{\boldsymbol{\eta}}$. For $b = 1, \ldots, B$ (where B is the number of resamples required).

1. Resample (with replacement) the per-segment/point residuals, store the values in $\mathbf{r}_b$.

2. Refit the model but with the response set to $\hat{\boldsymbol{\eta}} + \mathbf{r}_b$ (where $\hat{\boldsymbol{\eta}}$ are the fitted values from the orginal model).

3. Take the predicted values for the new model and store them.

From the predicted values stored in the last step the variance originating in the spatial part of the model can be calculated. The total variance of the abundance estimate (over the whole region of interest or sub-areas) can then be found by combining the variance estimate from the bootstrap procedure with the variance of the probability of detection from the detection function model using the delta method (which assumes that the two components of the variance are independent; Ver Hoef, 2012).

The above procedure assumes that there is no correlation in space between segments, which are usually contiguous along transects. If many animals are observed in a particular segment then we might expect there to be high numbers in the adjacent segments. A moving block bootstrap (MBB; Efron &

Tibshirani, 1993, Section 8.6) can account for some of this spatial autocorrelation in the variance estimation. The segments are grouped together into overlapping blocks (so if the block size is 5, block one is segments $1, \ldots, 5$, block two is segments $2, \ldots, 6$, and so on). Then, at step (2) above, resamples are taken at the block level (rather than individual segments within a transect). Using MMB will account for correlation between the segments at scales smaller than the block size, inflating the variances accordingly. Block size can be selected by plotting an autocorrelogram of the residuals from the DSM. Block size dictates the maximum amount of spatial autocorrelation accounted for, this may not fully account for the autocorrelation (testing sensitivity to block size by trying several different sizes can be time consuming).

Both bootstrap procedures can also be modified to take into account detection function uncertainty by simulating distances from the fitted detection function and then re-calculating the offset by fitting a detection function to the simulated distances.

Estimation of uncertainty for a given prediction region can be found by calculating the appropriate quantiles of the resulting abundance estimates (outlier removal may be required before quantile calculation). DSM uncertainty can be visualised via a plot of per-cell coefficient of variation obtained by dividing the standard error for each cell by its predicted abundance (as in Fig. 5).

Recent developments

GAM uncertainty and variance propagation

Rather than using a bootstrap, one can use GAM theory to construct uncertainty estimates for DSM abundance estimates. This requires that we use the distribution of the parameters in the GAM to simulate model coefficients, using them to generate replicate abundance estimates (further information can found in Wood, 2006, page 245). Such an approach removes the need to refit the model many times, making variance estimation much faster.

Williams *et al.* (2011) go a step further and incorporate the uncertainty in the estimation of the detection function into the variance of the spatial model, albeit only when segment level covariates are in the DSM. Their procedure is to fit the density surface model with an additional random effect term that characterises the uncertainty in the estimation of the detection function (via the derivatives of the probability of detection, $\hat{p}$, with respect to their parameters). Variance estimates of the abundance calculated using standard GAM theory will include uncertainty from the estimation of the detection function. A more complete mathematical explanation of this result is given in Appendix B.

We consider that propagating the uncertainty in this manner to be preferable to the MBB because it is more computationally efficient meaning investigators can easily and quickly estimate variances of complex models. The confidence intervals produced via variance propagation appear comparable (if not narrower) than their bootstrap equivalents, while maintaining good coverage (results of a small simulation study are given in Appendix C).

Fig. 5 shows a map of the coefficient of variation for the model which includes both location and depth covariates. Variance has been calculated using the variance propagation method.

Edge effects

Previous work (Ramsay, 2002; Wang & Ranalli, 2007; Wood *et al.*, 2008; Scott-Hayward *et al.*, 2013; Miller & Wood, submitted) has highlighted the need to take care when smoothing over areas with complicated boundaries, e.g., those with rivers, peninsulae or islands. If two parts of the study area (either side of a river or inlet, say) are inappropriately linked by the model (i.e. if the distance between the points is measured as a straight line, rather taking into account obstacles) then the boundary feature (river, etc) can be "smoothed across" so positive abundances are predicted in areas where animals could not possibly occur. Ensuring that a realistic spatial model has been fitted to the data is essential for valid inference. The soap film smoother of Wood *et al.* (2008) is an appealing solution: a bivariate smooth function of location that can be included in any GAM but that allows for boundary conditions to be estimated and obeyed for a complex study area. Such an approach can be helpful when uncertainty is estimated via a bootstrap as edge effects can also cause large, unrealistic predictions which can plague other smoothers (Bravington & Hedley, 2009).

Even if the study area does not have a complicated boundary, edge effects can still be problematic. Miller (2012) notes that some smoothers have plane components that tend to cause the fitted surface to increase unrealistically as predictions are made further away from the locations of survey effort. This problem can be alleviated by the using a different type of smoother (e.g. a generalisation of thin plate regression splines called *Duchon splines*).

Tweedie distribution

The Tweedie distribution offers a flexible alternative to the quasi-Poisson and negative binomial distributions as a response distribution when modelling count data (Candy, 2004). In particular it is useful when there are a high proportion of zeros in the data (Shono, 2008; Peel *et al.*, 2012) and avoids multiple-stage modelling of zero-inflated data (as in Barry & Welsh, 2002).

The distribution has three parameters parameters: a mean, dispersion and a third power parameter, which leads to additional flexibility. The distribution does not change appreciably when the power parameter is changed by less than 0.1 and therefore a simple line search over the possible values for the power parameter is usually a reasonable approach to estimating the parameter. Mark Bravington (pers. comm.) suggested plotting the square root of the absolute value of the residuals against fitted values; a "flatter" plot (points forming a horizontal line) give an indication of a "good" value. We additionally suggest using the metrics described in the next section for model selection.

Appendix D gives further details about the Tweedie distribution (including its probability density function and further references).

Practical advice

A flow diagram of the modelling process for creating a DSM is shown in Fig. 6. The diagram shows which methods are compatible with each other and what the options are for modelling a particular data set.

In our experience, it is sensible to obtain a detection function that fits the data as well as possible and only begin spatial modelling after a satisfactory detection function has been obtained. Model selection for the detection function can be performed using AIC and model checking using goodness-of-fit tests given in Burnham *et al.* (2004, Section 11.11). If animals occur in clusters rather than individually, bias can be incurred due to the higher visibility of larger clusters. It may then be necessary to include size as a covariate in the detection function (see Buckland *et al.*, 2001, Section 4.8.2.4). For some species cluster size may change according to location, Ferguson *et al.* (2006) use two GAMs (one to model observed clusters and one to model the cluster size) to deal with spatially-varying cluster size amongst delphinids, though the authors do not present the variance of the resulting predictions.

Smooth terms can be selected using (approximate) p-values (Wood, 2006, Section 4.8.5). An additional useful technique for covariate selection is to use an extra penalty for each term in the GAM allowing smooth terms to be removed from the model during fitting (illustrated in Appendix A; Wood, 2011). Smoothness selection is performed by generalized cross validation (GCV) score, unbiased risk estimator (UBRE) or restricted maximum likelihood (REML) score. When model covariates are effectively functions of one another (e.g. depth could be written as a function of location) GCV and

UBRE can suffer from optimisation problems (Wood, 2006, Section 4.5.3) which can lead to unstable models (Wood, 2011). REML provides a fitting criteria with a more pronounced optima which avoids some problems with parameter estimation, though caution should always be taken when dealing with highly correlated covariates. A significant drawback of REML is that scores cannot be used to compare models with different linear terms or offsets (Wood, 2011), though the p-value and additional penalty techniques described above can be used to select model terms. We highly recommend the use of standard GAM diagnostic plots; Wood (2006) provides further practical information on GAM model selection and fitting.

In the analysis of the dolphin data we included a smooth of location that nearly doubles the percentage deviance explained (27.3% to 52.7%). One can see this when comparing the two plots in Fig. 3 and the plot of the depth (Fig. 2), the plot of the model containing only a smooth of depth looks very similar to the raw plot of the depth data. Using a smooth of location can be a primitive way to account for spatial autocorrelation and/or as a proxy for other spatially varying covariates that are unavailable.

A more sophisticated way to account for spatial autocorrelation between segments (within transects) is to use an autocorrelation structure within the DSM (e.g. autoregressive models). Appendix A shows an example using generalized additive mixed model (GAMMs; Wood, 2006, Section 6.6, see Appendix A for an example) to construct an autoregressive (lag 1) correlation structure. This gives a significant reduction in variance, tightening the confidence interval around the abundance estimate.

In the analysis presented here, spatial location has been transformed from latitude and longitude to kilometres north and east of the centre of the survey region at $(27.01°, -88.3°)$. This is because the bivariate smoother used (the thin plate spline; Wood, 2003) is isotropic: there is only one parameter controlling the smoothness in both directions. Moving one degree in latitude is not the same as moving one degree in longitude and so using kilometres from the centre of the study region makes the covariates isotropic. Using metric units rather than non-standard units of measure such as degrees or feet throughout makes analysis much easier.

A smooth of an environment-level covariate such as depth can be very useful for assessing the relationships between abundance and the covariate (as in Fig. 4). Caution should be employed when interpreting smooth relationships and abundance estimates, especially if there are gaps over the range of covariate values. Large counts may occur at large values of depth but if no further observations occur at such a large value, then investigators should be skeptical of any relationship.

Discussion

The use of model-based inference for determining abundance and spatial distribution from distance sampling data presents new opportunities in the field of population assessment. Spatial models can be particularly useful when it comes to prediction: making predictions for some subset of the study area relies on stratification in design-based methods and as such can be rather limited. Our models also allow inference from a sample of sightings to a population in a study area without depending upon a random sample design, and therefore data collected from "platforms of opportunity" (Williams *et al.*, 2006) can be used (although a well designed survey is always preferable).

Unbiased estimates are dependent upon either (i) distribution of sampling effort being random throughout the study area (for design-based inference) or (ii) model correctness (for model-based inference). It is easier to have confidence in the former rather than in the latter because our models are always wrong. Nevertheless model-based inference will play an increasing role in population assessment as the availability of spatially-referenced data increases.

The field is quickly evolving to allow modelling of more complex data building on the basic ideas of density surface modelling. We expect to see large advances in temporal inferences and the handling of zero-inflated data and spatial correlation. These should become more mainstream as modern spatio-temporal modelling techniques are adopted. Petersen *et al.* (2011) provided a very basic framework for temporal modelling; their model included "before" and "after" smooth terms to quantify the impact of the construction of an offshore windfarm. Zero-inflation in count data may be problematic and two-stage approaches such as Barry & Welsh (2002) as well as more flexible response distributions made possible by Rigby & Stasinopoulos (2005) have yet to be exploited by those using distance sampling data. Spatial autocorrelation can be accounted for via approaches that explicitly introduce correlations such as generalized estimating equations (GEEs; Hardin & Hilbe, 2003) or generalized additive mixed models or via mechanisms such as that of Skaug (2006), which allow observations to cluster according to one of several states (such as high vs low density patches, possibly in response to temporary agglomerations of prey, although the mechanism is unimportant). These advances should assist both modellers and wildlife managers to make optimal conservation decisions.

Advances in Bayesian computation (INLA; Rue *et al.*, 2009), make one-step, Bayesian, density surface models computationally feasible (as INLA is an alternative to MCMC). We anticipate that such a direct modelling technique will dominate future developments in the field.

Density surface modelling allows wildlife managers to make best use of the available spatial data to understand patterns of abundance, and hence make better conservation decisions (e.g., about reserve or development placement). The recent advances mentioned here increase the reliability of the outputs from a modelling exercise, and hence the efficacy of these decisions. Density surface modelling from survey data is an active area of research, and we look forward to further improvements and extensions in the near future.

Acknowledgments

We wish to thank Paul Conn, another anonymous reviewer, and the associate editor for their helpful comments. DLM wishes to thank Mark Bravington and Sharon Hedley for their detailed discussions and for providing code for their variance propagation method. Funding for the implementation of the recent advances into the `dsm` package and Distance software came from the US Navy, Chief of Naval Operations (Code N45), grant number N00244-10-1-0057.

References

Baddeley, A. & Turner, R. (2000) Practical maximum pseudolikelihood for spatial point patterns. *Australian & New Zealand Journal of Statistics*, **42**, 283–322.

Barry, S.C. & Welsh, A.H. (2002) Generalized additive modelling and zero inflated count data. *Ecological Modelling*, **157**, 179–188.

Bravington, M.V. & Hedley, S.L. (2009) Antarctic minke whale abundance estimates from the second and third circumpolar IDCR/SOWER surveys using the SPLINTR model. Paper SC/61/IA14, IWC Scientific Committee.

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

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. & Thomas, L. (2004) *Advanced Distance Sampling.* Oxford University Press.

Burnham, K.P., Buckland, S.T., Laake, J.L., Borchers, D.L., Marques, T.A., Bishop, J.R. & Thomas, L. (2004) Further topics in distance sampling. *Advanced Distance Sampling* (eds. S.T. Buckland, D.R. anderson, K.P. Burnham, J.L. Laake, D.L. Borchers & L. Thomas). Oxford University Press.

Candy, S. (2004) Modelling catch and effort data using generalised linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects. *CCAMLR Science*, **11**, 59–80.

Chelgren, N.D., Samora, B., Adams, M.J. & McCreary, B. (2011) Using spatiotemporal models and distance sampling to map the space use and abundance of newly metamorphosed western toads (*Anaxyrus boreas*). *Herpetological Conservation and Biology*, **6**, 175–190.

Conn, P.B., Laake, J.L. & Johnson, D.S. (2012) A hierarchical modeling framework for multiple observer transect surveys. *PLoS ONE*, **7**, e42294.

Cox, D.R. & Isham, V. (1980) *Point Processes.* Monographs on Applied Probability and Statistics. Chapman and Hall. ISBN 9780412219108.

Efron, B. & Tibshirani, R.J. (1993) *An Introduction to the Bootstrap.* Chapman & Hall/CRC. ISBN 9780412042317.

Ferguson, M.C., Barlow, J., Fiedler, P., Reilly, S.B. & Gerrodette, T. (2006) Spatial models of delphinid (family Delphinidae) encounter rate and group size in the eastern tropical Pacific Ocean. *Ecological Modelling*, **193**, 645–662.

Halpin, P., Read, A., Fujioka, E., Best, B., Donnelly, B., Hazen, L., Kot, C., Urian, K., LaBrecque, E., Dimatteo, A., Cleary, J., Good, C., Crowder, L. & Hyrenbach, K.D. (2009) OBIS-SEAMAP: The world data center for marine mammal, sea bird, and sea turtle distributions. *Oceanography*, **22**, 104–115.

Hardin, J. & Hilbe, J. (2003) Generalized Estimating Equations. Chapman and Hall/CRC, London, UK.

Hedley, S.L. & Buckland, S.T. (2004) Spatial models for line transect sampling. *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 181–199.

Johnson, D.S., Laake, J.L. & Ver Hoef, J.M. (2010) A model-based approach for making ecological inference from distance sampling data. *Biometrics*, **66**, 310–318.

Link, W.A. & Barker, R.J. (2009) *Bayesian Inference: with ecological applications.* Academic Press, London, UK.

Marques, T.A., Thomas, L., Fancy, S. & Buckland, S.T. (2007) Improving estimates of bird density using multiple-covariate distance sampling. *The Auk*, **124**, 1229–1243.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models.* Chapman & Hall/CRC.

Miller, D.L. (2012) *On smooth models for complex domains and distances.* Ph.D. thesis, University of Bath.

Miller, D.L., Rexstad, E.A., Burt, M.L., Bravington, M.V. & Hedley, S.L. (2013) *dsm: Density surface modelling of distance sampling data.* URL `http://github.com/dill/dsm`

Miller, D.L. & Wood, S.N. (submitted) Finite area smoothing with generalized distance splines.

Moore, J.E. & Barlow, J. (2011) Bayesian state-space model of fin whale abundance trends from a 1991-2008 time series of line-transect surveys in the California Current. *Journal of Applied Ecology*, **48**, 1195–1205.

Niemi, A. & Fernández, C. (2010) Bayesian spatial point process modeling of line transect data. *Journal of Agricultural, Biological, and Environmental Statistics*, **15**, 327–345.

Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'Connell, A., Miller, P. & Louzao, M. (2011) Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. *Biological Conservation*, **156**, 94–104.

Peel, D., Bravington, M.V., Kelly, N., Wood, S.N. & Knuckey, I. (2012) A Model-Based Approach to Designing a Fishery-Independent Survey. *Journal of Agricultural, Biological, and Environmental Statistics*, **18**, 1–21.

Petersen, I.K., MacKenzie, M.L., Rexstad, E.A., Wisz, M.S. & Fox, A.D. (2011) Comparing pre- and post-construction distributions of long-tailed ducks *Clangula hyemalis* in and around the Nysted offshore wind farm, Denmark: a quasi-designed experiment accounting for imperfect detection, local surface features and autocorrelation. Technical report 2011-1, Centre for Research into Enivironmental and Ecological Modelling.

Ramsay, T. (2002) Spline smoothing over difficult regions. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **64**, 307–319.

Rigby, R. & Stasinopoulos, D. (2005) Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society-Series C Applied Statistics*, **54**, 507–554.

Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical Modeling and Inference in Ecology.* Academic Press, London, UK.

Royle, J., Dawson, D. & Bates, S. (2004) Modeling abundance effects in distance sampling. *Ecology*, **85**, 1591–1597.

Rue, H., Martino, S. & Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Statist. Soc. B*, **71**, 319–392.

Ruppert, D., Wand, M. & Carroll, R.J. (2003) *Semiparametric Regression.* Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press.

Schmidt, J.H., Rattenbury, K.L., Lawler, J.P. & Maccluskie, M.C. (2011) Using distance sampling and hierarchical models to improve estimates of Dall's sheep abundance. *The Journal of Wildlife Management*, **76**, 317–327.

Scott-Hayward, L.A.S., MacKenzie, M.L., Donovan, C.R., Walker, C.G. & Ashe, E. (2013) Complex region spatial smoother (CReSS). *Journal of Computational and Graphical Statistics.*

Shono, H. (2008) Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research*, **93**, 154–162.

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

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

Ver Hoef, J.M. (2012) Who invented the delta method? *The American Statistician*, **66**, 124–127.

Ver Hoef, J.M., Cameron, M.F., Boveng, P.L., London, J.M. & Moreland, E.E. (2013) A spatial hierarchical model for abundance of three ice-associated seal species in the eastern Bering Sea. *Statistical Methodology*, pp. 1–44.

Wang, H. & Ranalli, M. (2007) Low-rank smoothing splines on complicated domains. *Biometrics*, **63**, 209–217.

Williams, R., Hedley, S.L., Branch, T.A., Bravington, M.V., Zerbini, A.N. & Findlay, K.P. (2011) Chilean blue whales as a case study to illustrate methods to estimate abundance and evaluate conservation status of rare species. *Conservation Biology*, **25**, 526–535.

Williams, R., Hedley, S.L. & Hammond, P. (2006) Modeling distribution and abundance of Antarctic baleen whales using ships of opportunity. *Ecology and Society*, **11**, 1.

Wood, S.N. (2003) Thin plate regression splines. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **65**, 95–114.

Wood, S.N. (2006) *Generalized Additive Models: An introduction with R* . Chapman & Hall/CRC.

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **73**, 3–36.

Wood, S.N., Bravington, M.V. & Hedley, S.L. (2008) Soap film smoothing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **70**, 931–955.

Figures

Fig. 1 Estimated detection function for pantropical dolphin clusters overlaid onto the scaled histogram of observed distances. Distances are recorded in metres.

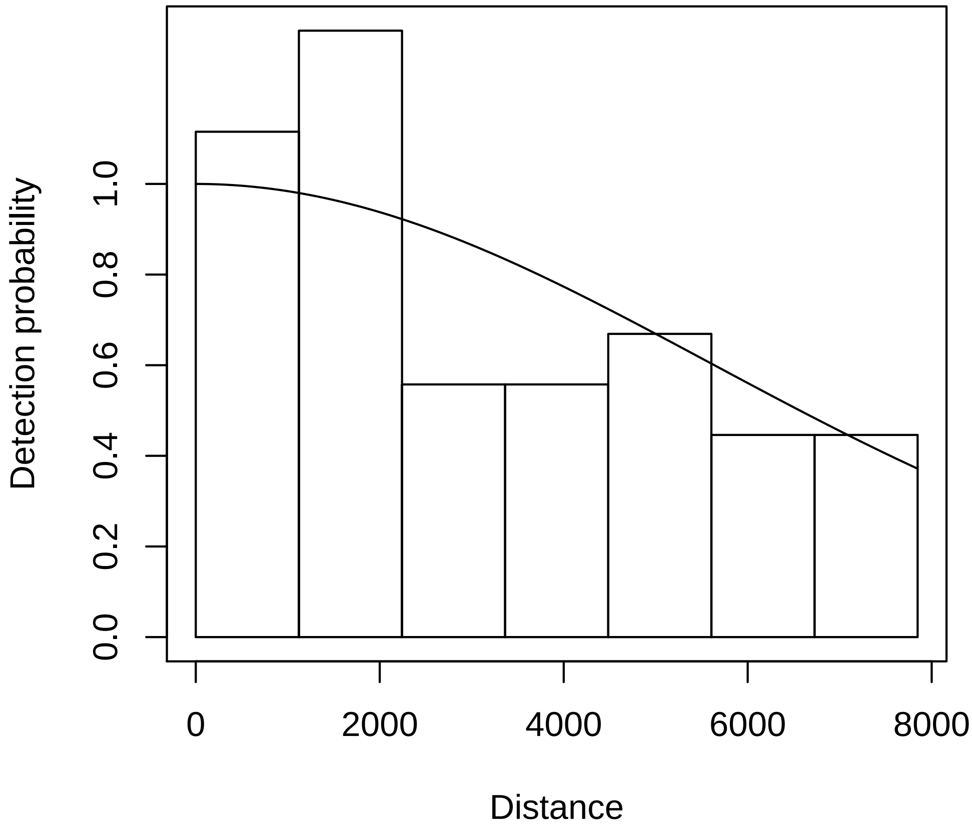

Fig. 2 The region, transect centrelines and location of detected pantropical dolphin clusters, where size of circle corresponds to the cluster size, overlaid onto depth data.

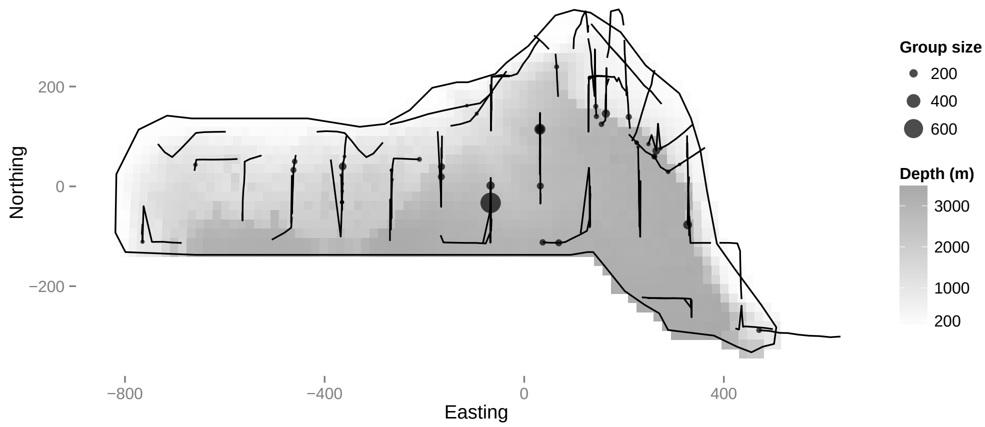

Fig. 3 Predicted abundance of dolphins from the DSM using only depth as an explanatory variable (top) and the model using both depth and location (bottom).

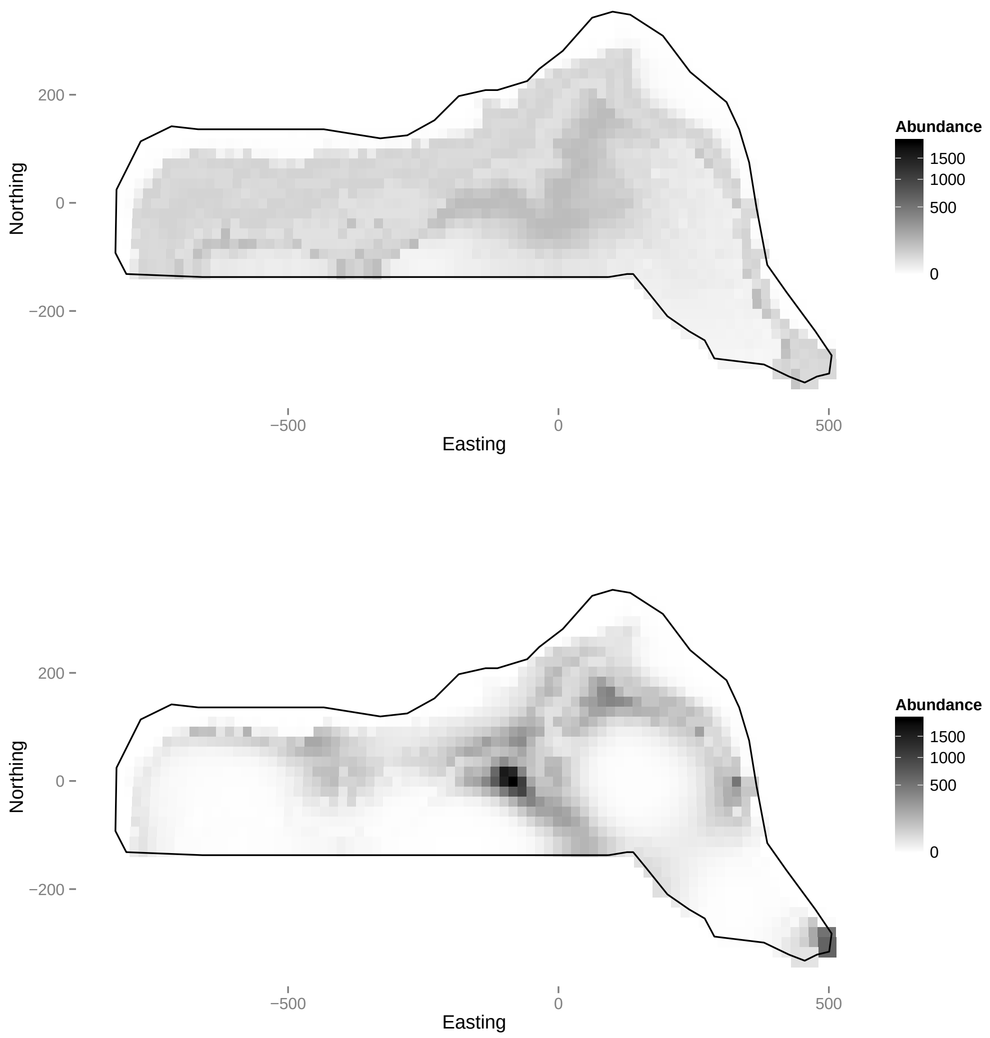

Fig. 4 Plot of the effect on the response of depth (from the model with both depth and location smooths), note that it is possible to draw a straight line between 750m and 3000m within the confidence band (between the dashed lines), so the wiggles in the smooth may not be indicative of any relationship. What is clear is that there the estimated number of dolphins increases up to about 500m. The rug ticks at the bottom of the plot indicate we have good coverage of the range of depth values in the survey area. Note that the y axis in such plots is on the scale of the link function (log in this case), so care should be taken in their interpretation.

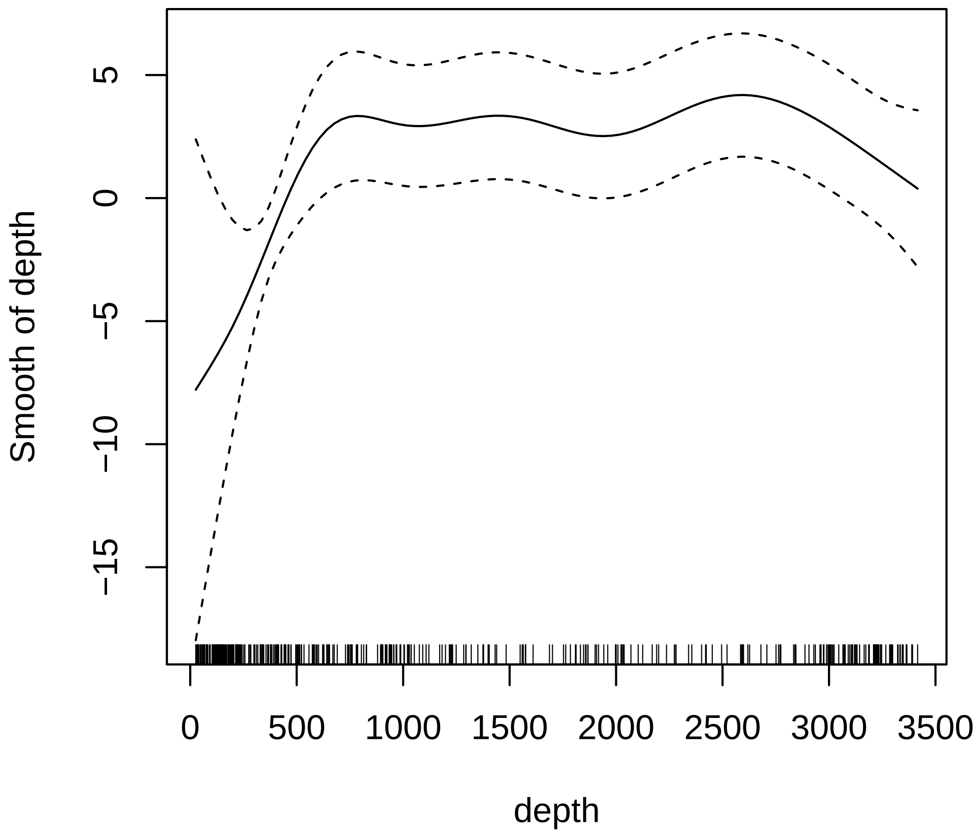

Fig. 5 Map of the coefficients of variation for the model with smooths of both depth and location. Uncertainty was estimated using the variance propagation method of Williams *et al.* (2011). As might be expected, there is high uncertainty where there is low sampling effort (Fig. 2).

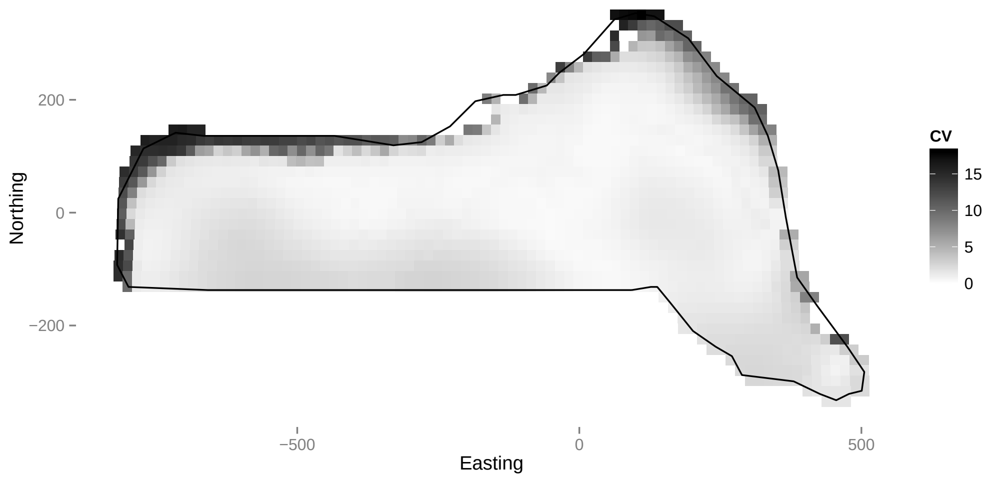

Fig. 6 Flow diagram showing the modelling process for creating a density surface model.