Running title: Spatial models for distance sampling
Number of words: ~5151 UPDATE
Number of tables: 0
Number of figures: 6
Number of references: 33 UPDATE

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: 7th May 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 and are largely implemented in the popular Windows package Distance (or are soon to be incorporated).

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 model which include location as a covariate. This article is concerned with combining spatial modelling techniques with distance sampling (Buckland *et al.*, 2001, 2004).

Distance sampling takes plot sampling (counting all the individuals or groups of objects within a strip or circle) and extends it to the case where detection is not certain. Observers move along lines or stand at 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. 2), 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 probability of detection can be estimated. The estimated probability that an animal is detected, $\hat{p}_i$, 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 groups, each of size s_i (if individuals are observed, $s_i = 1 \forall i$). (Buckland *et al.*, 2001, Chapter 3). In general, distance sampling is more efficient than plot sampling because a much higher proportion of observations can be used in the analysis. 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 the 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. In contrast 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. 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 landscape of spatial modelling of distance sampling data, illustrating some recent developments most useful to applied ecologists. The methods discussed have been available in the popular Windows application Distance (Thomas *et al.*, 2010) for some time but the recent advances covered here have been implemented in a new R package, `dsm` (?) and are soon to incorporated into Distance.

[[this needs to go somewhere]] Throughout this article a motivating data set is used to illustrate the methods. These data are from a combination of several shipboard surveys conducted on several cetacean species in the Gulf of Mexico. We investigate 47 observations of groups of pantropical spotted dolphins (*Stenella attenuata*); group size was recorded, as well as the Beaufort sea state

at the time of the observation. Coordinates for each observation and bathymetry data were available as covariates for the analysis. A complete example analysis is provided as an online appendix. The data used in the analysis are available in the `dsm` package and Distance.

The rest of the article follows this structure: we first review approaches for the spatial modelling of distance sampling data before focussing on the density surface modelling approach of Hedley & Buckland (2004); explain how to estimate abundance and uncertainty; describe recent advances and provide practical advice regarding model fitting, formulation and checking. Finally we discuss future directions for research in spatially modelling distance sampling data.

Approaches to spatial modelling of distance sampling data

Modelling of spatially referenced distance sampling data is in essence the same as modelling spatially-referenced count data, with the additional information provided by collecting distances in order to account for imperfect detection of the species in question. We now 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. We begin with two-stage approaches and then move on to one-stage approaches.

Two-stage approaches

The main focus for 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 stages: first a detection function is fitted to the distance data to obtain detection probabilities for groups (flocks, pods etc) or individuals. Counts are allocated to a series of segments (contiguous transect sections) which have their areas multiplied by the detection probabilities. A generalised additive model (GAM; e.g. Wood, 2006) is then constructed with the per-segment counts as the response. GAMs provide an extremely flexible class of models which include generalized linear models (GLMs; McCullagh & Nelder, 1989) but extend them with the addition of splines to create smooth functions of covariates, random effects terms and correlation structures (amongst other extensions). This article aims to cover other recent advances using this approach so we refer readers to the later sections of the paper for details of developments since Hedley & Buckland (2004).

Niemi & Fernández (2010) proposed a Bayesian point process approach. 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 (though an extension that incorporates uncertainty is, however, feasible), the model also does not account for group size this could be included by considering a marked point process (Cox & Isham, 1980, Section 5.5).

Ver Hoef *et al.* (2013) model seal populations in the Bering sea by combining a detection function and including additional information from a model of seal haul-outs on ice (both estimated using frequentist methods) with a Bayesian spatial model. The detection function and haul-out model correct the observed density estimates which are then modelled using a Bayesian hierarchical model for the spatial component which itself is split into a presence/absence part (to allow modelling of the large number of zeros in the data) and a density portion (which also accounts for spatial autocorrelation). The authors show that that when extra information is available (such as the haul-out data collected from tags on the seals), this can be incorporated into a model, giving additional insight and interpretability to results.

Two-stage models have the disadvantage of requiring that uncertainty from both parts of the model must be combined. Appropriately combining uncertainty from the detection function and the spatial model can be tricky and ignoring uncertainty from one of the sources can lead to falsely narrow confidence intervals for abundance estimates. More information regarding how this is addressed for DSMs is given in the section "Recent developments", below.

We note that there are many approaches to modelling spatially embedded count data (for example, random forests, Breiman (2001); boosted regression trees, Friedman (2002); Oppel *et al.* (2011) provide an overview of such methods for marine bird modelling). Also worthy of note is the approach of Barry & Welsh (2002) using a two-stage approach to model presence/absence then spatial pattern (both via GAMs) in order to account for zero-inflation.

Any of these methods could potentially be adapted into a two-stage approach for distance sampling data by adjusting the counts but that they all fall victim to the above issue of combining uncertainties.

One-stage approaches

Rather than fitting two separate models, many recent articles have combined these two steps (most via hierarchical Bayesian methods). The first of these chronologically was Royle *et al.* (2004), formulating an unconditional likelihood per-point/line that is a function of the unobserved transect abundances. These unobserved abundances were treated as (Poisson or negative binomial) random effects, which were then integrated out to give a per-transect likelihood which is a function only of detection function parameters and parameters of the random effects (linear functions of the environmental covariates). Due to the multinomial nature of the per-transect likelihood proposed, distance data must be binned, resulting in a loss of information (an arbitrarily large number of bins could be used as an approximation to continuous data, though this is potentially computationally intensive). Chelgren *et al.* (2011) proposed replacing the multinomial per-transect likelihood with a binomial distribution multiplied by a detection function. The binomial term is effectively the collapsing of the multinomial bins into one very large bin and gives the number of animals captured in the transect, thus allowing the use of exact distances.

The work of Schmidt *et al.* (2011) takes a somewhat similar approach to Royle & Dorazio (2008), building a presence/absence-type model for groups, augmenting the data with unobserved groups (similar to the approach taken in Royle & Dorazio (2008)). The authors then used a Poisson distribution to model group size (using a random effect to incorporate overdispersion), combining these parts to give a model of individual abundance. The authors used the Distance software (Thomas *et al.*, 2010) to determine the form of the detection function but conducted all parameter estimation (including detection function parameters) as part of one hierarchical Bayesian model (hence we consider this a one-step approach). Conn *et al.* (2012) also use 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) adopt a hierarchical Bayesian state-space model, separating the problem into an observation and process components. The process component describes the underlying population density as it changes over time and space (though the authors only include strata as a spatial component). The observation part of the model then links the process model to the data via the detection function.

A Bayesian formulation may be advantageous as it allows for easy inclusion of random effects as well as additional information from other sources and experiments. A one-step procedure means that variance estimates include observation and process uncertainty without need for additional calculations.

Outside of the Bayesian world, 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). In contrast to Hedley & Buckland (2004), 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 ensures that correlations between the detection function and underlying point process are estimated correctly (and do not falsely inflate or deflate variance estimates). The authors also addressed the issue of overdispersion unmodelled by spatial covariates (i.e. counts that do not follow a Poisson mean-variance relationship) using a post-hoc correction factor.

Density surface modelling

This section focuses on modelling the density/abundance estimation stage of the DSM approach introduced above. 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 or covariate values vary appreciably within a segment (usually making the segments approximately square, $2w \times 2w$, is sufficient). Count or estimated abundance is then modelled as a (sum of) smooth function(s) of covariates using a generalized additive model. For each segment or point, the response is modelled as a function of environmental covariates that are measured at the segment/point level (z_{jk} with k indexing the covariates, e.g., location, sea surface temperature, weather conditions).

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 $A_j = \pi w^2$ (where w is the truncation distance).

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

Count as response

The model for the count per segment is:

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

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 overdispersed Poisson, negative binomial, or Tweedie distribution (see *Recent developments*, below).

Fig. 1 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. 2. 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 Generalized Cross Validation (GCV) score, the latter had a considerably lower score (39.12 vs 48.46) and so would be selected as our "best" 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 for the dolphin data can be seen in Fig. 4.

Estimated abundance as response

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} group in segment j (if the animals occur individually then $s_{jr} = 1, \forall j, r$).

The following model is then fitted:

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

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 is now the area rather than effective area of the segment/point.

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/group i and segment/point j) are collected on the level of observations; for example sex or group 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 group) level covariates $\hat{p}(z_i)$. Object level covariates can be incorporated into the model by adopting the following estimator of the per-segment 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 offset, but instead dividing the count (or estimated abundance) by the area of the segment (and weighting observations by the segment areas). We concentrate on abundance here, see Hedley & Buckland (2004) for further details on modelling density.

Prediction

Abundance can be predicted for the each cell in a grid over the region in question and by summing predicted values over corresponding grid cells. The areas of the prediction cells must be accounted for in the predictions. Environmental covariates included in the model must be available at each prediction cell at the required resolution (using prediction grid cells that are smaller than the resolution of the spatially referenced data have 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 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; Seber, 1982).

The above procedure assumes that there is no correlation in space between segments, 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 of the blocks (contiguous collections of segments) rather than individual segments within the transects. Using blocks should account for some of the autocorrelation between the segments, inflating the variances accordingly. However, because the block size dictates the maximum amount of spatial autocorrelation accounted for, this may not fully account for the autocorrelation. These bootstrap procedures can also be modified to take into account detection function uncertainty by generating new distances from the fitted detection function and then re-calculating the offset by fitting a detection function to the new distances.

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 when only segment level covariates are in the DSM. Their procedure is to fit the density surface model with an additional random effects 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 is not only more computationally efficient but also preferable to the moving block bootstrap from a technical perspective. A moving block bootstrap does not fully account for spatial autocorrelation because when it reallocates blocks of residuals, it does so without considering the dependence between blocks. This can then lead to wide confidence intervals. The confidence intervals produced via variance propagation are 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) 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 domain (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 can be "smoothed across" leading to incorrect inference. 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 appealing as the model jointly estimates boundary conditions for a complex study area along with the interior smooth. This can be helpful when uncertainty is estimated via a bootstrap as the model helps avoid 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 *et al.* show that global smoothers which have unpenalized plane components tend to cause the fitted surface to increase unrealistically as predictions move further away from the locations of survey effort. They suggest the use of Duchon splines (a generalisation of thin plate regression splines) to alleviate the problem.

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). Through the parameter λ, many common distributions arise; varying λ between 1 (Poisson) and 2 (gamma) leads to a random variable which is a sum of M gamma variables where M is Poisson distributed (Jørgensen, 1987). The distribution does not change appreciably when λ is changed by less than 0.1 therefore, a simple line search over the possible values of λ is usually reasonable. Mark Bravington (pers. comm.) suggested plotting the square root of the absolute value of the residuals against fitted values; a "flat" plot (points forming a horizontal line) give an indication of a "good" value for λ. We additionally suggest using the metrics described in the next section for model selection.

Practical advice

Fig. 6 shows a flow diagram of the modelling process for creating a DSM. 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 after a satisfactory detection function has been obtained, begin spatial modelling. 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 groups rather than individually, bias can be incurred due to the higher visibility of larger groups. 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 group size may change according to location, Ferguson *et al.* (2006) use two GAMs (one to model observed groups and one to model the group size) to deal with spatially-varying group size amongst delphinids, though the authors are not able calculate the variance of the resulting predictions.

Smooth terms can be selected using (approximate) p-values, as one would usually for a GAM. 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 the example analysis; 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 failures (Wood, 2006, Section 4.5.3) which can lead to unstable models (Wood, 2011). To avoid these issues REML is recommended for smoothness selection when many spatially-referenced covariates are used. A significant drawback is that REML scores can only be used to compare models with the same fixed effects (i.e. linear terms; 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. This not only nearly doubles the percentage deviance explained (27.3% to 52.7%), it also allows us to account for spatial autocorrelation (in a primitive way). One can see this when comparing the two plots in Fig. 3 and the plot of the depth (Fig. 1), the plot of the model containing only a smooth of depth looks very similar to the raw plot of the depth data. 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 a high value of depth but if no further observations occur at such a high value, then investigators should be skeptical of any relationship. A smooth of location can be useful although limiting the "wigglyness" of smooths of spatial location (by limiting their basis size) can be a useful way of restricting their influence whilst still allowing them to "mop up" the residual spatial correlation in the data (see the example analysis).

In the analysis presented here we transform the covariates for spatial location from latitude and longitude to kilometres north and east of the centre of the survey region at $(27.01^{\circ}, -88.3^{\circ})$. 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 SI units throughout makes analysis easier.

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. Inference from a sample of sightings to a population in a study area does not have to depend upon a random sample design, and therefore data collected from "platforms of opportunity" (Williams *et al.*, 2006) can be used.

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 via mechanisms such as that of Skaug (2006), which allowed 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.

Recent 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 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.
URL http://linkinghub.elsevier.com/retrieve/pii/S0304380002001941

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.

Breiman, L. (2001) Random forests. *Machine learning.*
URL http://link.springer.com/article/10.1023/A:1010933404324

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.
URL http://www.herpconbio.org/Volume_6/Issue_2/Chelgren_etal_2011.pdf

Conn, P.B., Laake, J.L. & Johnson, D.S. (2012) A Hierarchical Modeling Framework for Multiple Observer Transect Surveys. *PloS one*, **7**, e42294.
URL http://dx.plos.org/10.1371/journal.pone.0042294

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. URL `http://books.google.com/books?id=gLlpIUxRntoC&dq=an+introduction+to+the+bootstrap&hl=&cd=1&source=gbs_api`

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. URL `http://www.sciencedirect.com/science/article/pii/S0304380005004898`

Friedman, J.H. (2002) Stochastic gradient boosting. *Computational Statistics and Data Analysis.* URL `http://www.sciencedirect.com/science/article/pii/S0167947301000652`

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.

Jørgensen, B. (1987) Exponential dispersion models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **49**, 127–162.

Link, W.A. & Barker, R.J. (2009) *Bayesian Inference: with ecological applications.* Academic Press. URL `http://books.google.com/books?hl=en&lr=&id=hecon2l2QPcC&oi=fnd&pg=PP2&dq=link+barker&ots=S8--1npyLq&sig=TLCsfPaN2IRM1q5Tnl84lTlye1M`

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. ISBN 9780412317606.

URL http://books.google.com/books?id=h9kFH2_FfBkC&printsec=frontcover&dq=mccullugh+nelder&cd=1&source=gbs_api

Miller, D.L., Jones, E. & Matthiopoulos, J. (????) Reliable spatial smoothing without edge effects. pp. 1–8.

Miller, D.L. & Wood, S.N. (????) Finite area smoothing with generalized distance splines. pp. 1–27.

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.
URL http://doi.wiley.com/10.1111/j.1365-2664.2011.02018.x

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*.
URL http://www.sciencedirect.com/science/article/pii/S0006320711004319

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. 2011-1.

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*. The Analysis of Data from Populations, Metapopulations and Communities. ISBN 9780123740977.
URL http://books.google.com/books?id=rDppWpVP6a0C&printsec=frontcover&dq=Hierarchical+modeling+and+inference+in+ecology+inauthor:royle&hl=&cd=1&source=gbs_api

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.

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.
URL `http://doi.wiley.com/10.1002/jwmg.216`

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.*

Seber, G.A.F. (1982) *The Estimation of Animal Abundance and Related Parameters.* ISBN 9781930665552.
URL `http://books.google.com/books?id=bnGaPQAACAAJ&dq=seber&cd=10&source=gbs_api`

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., 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.
URL `http://dx.doi.org/10.1016/j.stamet.2013.03.001`

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 The region, transect centrelines and location of detected pantropical dolphin groups, where size of circle corresponds to the group size, overlaid onto depth data.

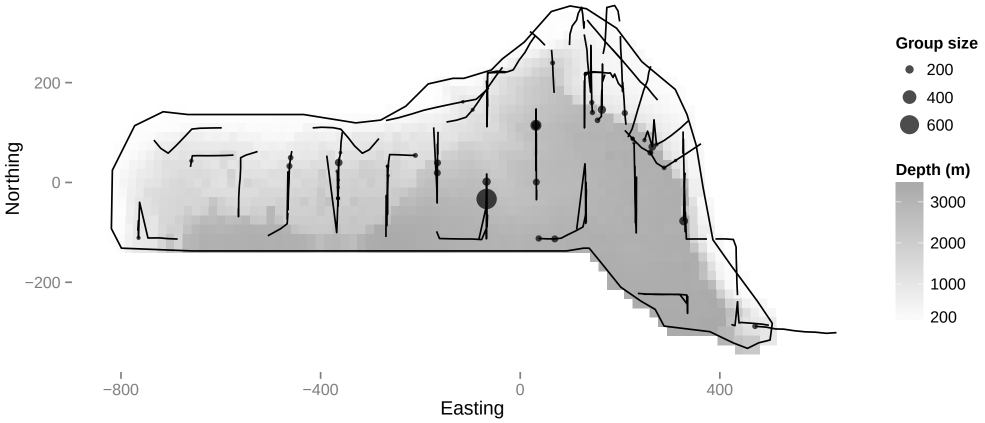


Fig. 2 Estimated detection function for pantropical dolphin groups overlaid onto the scaled histogram of observed distances. Distances are recorded in metres.

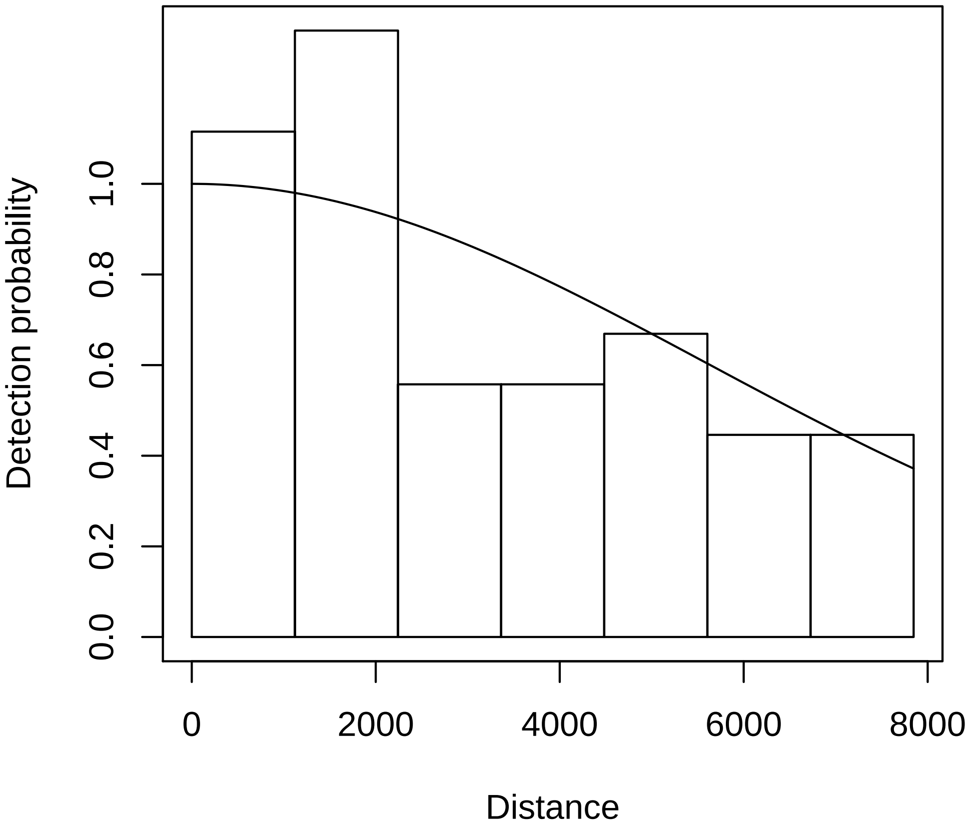


Fig. 3 Predicted abundance of dolphins from the model using only depth as an explanatory variable (top) and the model using both depth and location (bottom).

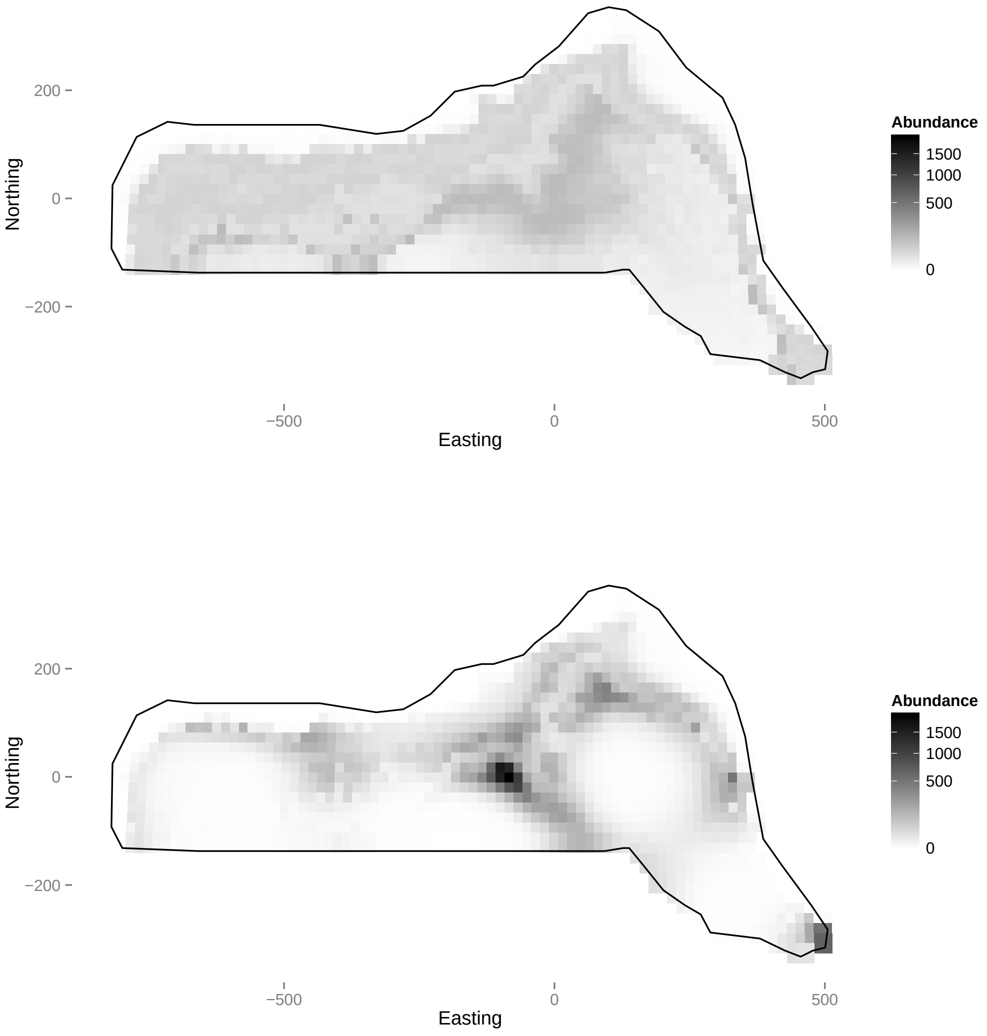


Fig. 4 Plot of the effect on the response of depth (from the model with both depth and location smooths), note that it is possible to draw a straight line between 750m and 3000m within the confidence band (between the dashed lines), so the wiggles in the smooth may not be indicative of any relationship. What is clear is that there is some effect 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.

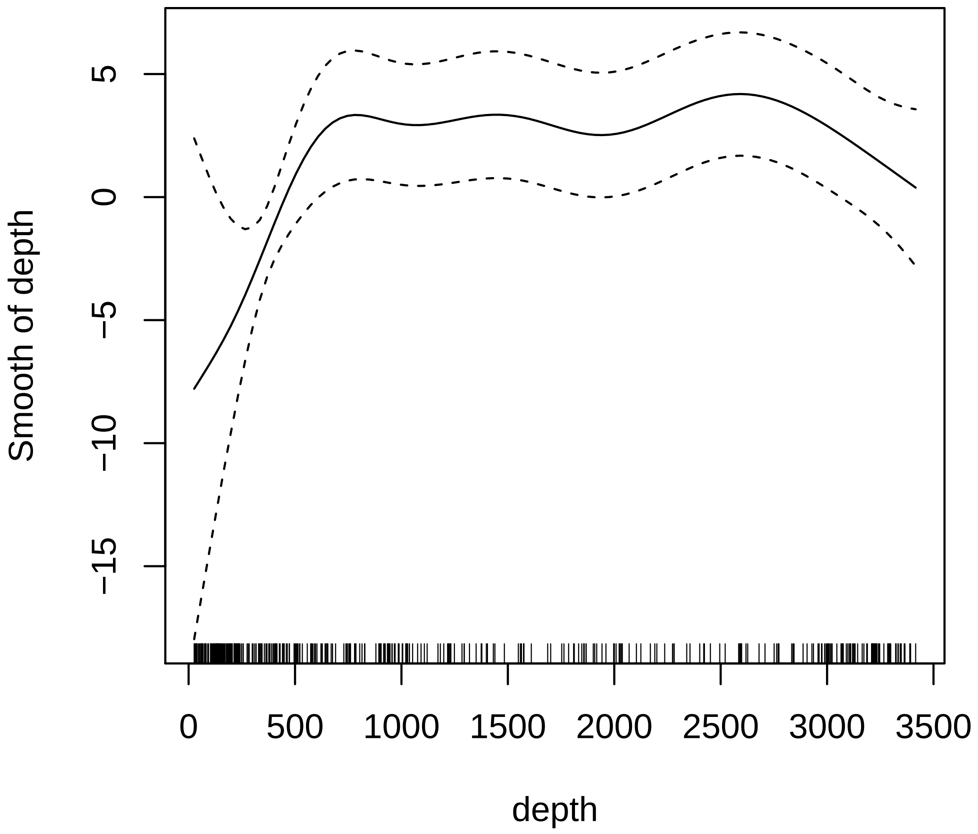


Fig. 5 Map of the coefficients of variation for the model with smooths of both depth and location. Uncertainty was estimated using the variance propagation method of Williams *et al.* (2011). As might be expected, there is high uncertainty where there is low sampling effort (Fig. 1).

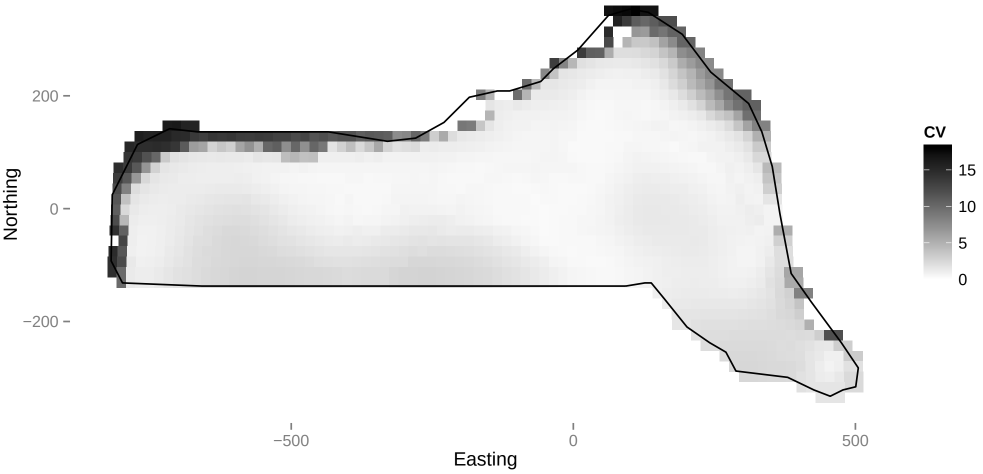
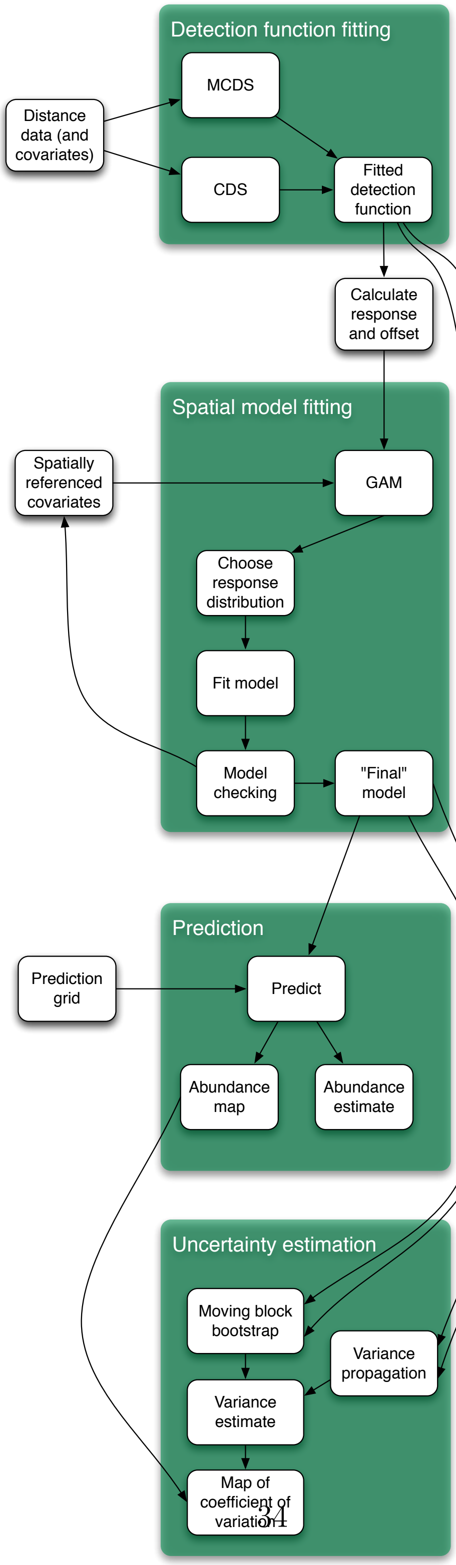


Fig. 6 Flow diagram showing the modelling process for creating a density surface model.


Detection function fitting
MCDS
Distance data (and covariates)
CDS
Fitted detection function
Calculate response and offset
Spatial model fitting
Spatially referenced covariates
GAM
Choose response distribution
Fit model
Model checking
"Final" model
Prediction
Prediction grid
Predict
Abundance map
Abundance estimate
Uncertainty estimation
Moving block bootstrap
Variance propagation
Variance estimate
Map of coefficient of variation