Running title: Spatial models for distance sampling
Number of words: ∼4158
Number of tables: 0
Number of figures: 6
Number of references: 27

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

David L. Miller[1*], **Louise Burt**[2], **Eric 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, Scotland*

*Correspondence author. dave@ninepointeightone.net

Date of submission: 4th October 2012

Summary

Our understanding of a biological population can be greatly enhanced by knowledge of their distribution in space and as a function of environmental covariates. It may also be necessary to use model-based inference to obtain abundance estimates from non-randomly designed surveys. Density surface modelling achieves both of these aims, allowing for the spatial modelling of distance sampling data. This review focuses on advances that have occurred since Hedley & Buckland (2004), in particular with regard to spatial smoothing: alternative response distributions for count data, dealing with complex regions, and estimating uncertainty. We offer a comparison of the various options for the practitioner as well as an examples and software.

Keywords: Distance sampling; spatial modelling; generalized additive models; Poisson processes; abundance estimation.

Introduction

When surveying biological populations it is increasingly common to record spatially referenced data; for example: coordinates of observations, bathymetry or chlorophyll A levels. Mapping the spatial distribution of a population can be extremely useful for practitioners, especially when communicating results to non-experts. Spatial models allow for the vast databases spatially-referenced data to be harnessed, allowing for interactions between environmental covariates and population densities to be investigated. Including spatial covariates into the model (for example, latitude and longitude) can account for spatial autocorrelation. 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)). Note that here we use the term "spatial model" to include any model which includes spatially referenced covariates, not just those which contain smooths of location.

This article concerns combining spatial modelling techniques with distance sampling (Buckland *et al.* (2001), Buckland *et al.* (2004)). Distance sampling takes simple strip sampling and extends it to the case where detection is not certain, for example when animals are cryptic.

Observers travel along transect centre lines or stand at points and record the perpendicular distance from the centre line or point to the object of interest (y). These distances are used to estimate the *detection function* ($g(y)$) by modelling the decrease in detectability with increasing distance from the line or point. The detection function may also include animal/observer specific covariates (Marques *et al.* (2007)). From the fitted detection function, the probability of detection can be calculated, this gives the probability that an animal within the truncation distance is detected, which can then be used to calculate density and abundance (Buckland *et al.* (2001), Chapter 3).

In a distance sampling analysis one assumes that the objects of interest are distributed according to some process (Buckland *et al.* (2001), Section 2.1). If the objects' locations are not dependent on any spatially varying covariates (such as location, distance from coast, depth, etc) a homogenous process is assumed; so with respect to the line, the objects are distributed uniformly. It is often possible to design surveys such that this assumption holds (for example, ensuring that transect lines run perpendicular to geographical features that would attract or repel animals) or by post-stratification (Buckland *et al.* (2001), Section 3.7).

Hedley & Buckland (2004) were the first to address spatial modelling of distance sampling data, allowing for a relaxation of the homogeneity of the point process, by including a rate parameter which is a function of spatially varying covariates. Thinking of the underlying placement of the objects as an inhomogeneous point process allows us to think of the detection process as a "thinning" (Cox & Isham (1980), Section 4.3) of the process, resulting in another inhomogeneous point process. By assuming the object placement and detection processes are independent, it is possible to separate these two processes (placement and thinning) in the likelihood.

Modelling the spatial process not only permits the use of spatially referenced data, it also gives practitioners the opportunity to use data from opportunistic surveys, for example "incidental" data arising from "ecotourism" cruises can be included in analyses (Williams *et al.* (2006)). Although with such non-random designs, spatial placement is less important than placement with respect to the range of covariate values expected to be encountered within the area of interest.

The rest of the article is structured as follows: we describe two methods which take the point process approach before going on to describe the two-stage approach of Hedley & Buckland (2004). We then describes recent advances, along with some practical advice regarding the model fitting, formulation and checking. 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 pan-tropical spotted dolphins in the Gulf of Mexico. These data consist of 47 observations of groups of dolphins. The group size was recorded, as well as the Beaufort sea state at the time of the observation. Coordinates for each observation and depth at a series of points over the prediction area were also available as covariates for the analysis. A complete example analysis can be found at `http://www.github.com/dill/dsm/wiki/`.

Direct modelling of the process

From the point process description, two modelling procedures arise. One approach is to directly model the point process, estimating the observation process as the thinning of that point process (Niemi & Fernández (2010), Johnson *et al.* (2010)). A second approach consists of performing a distance analysis and using the fitted detection function as part of spatial model (Hedley & Buckland (2004)).

Johnson *et al.* (2010) propose a point process-based model for distance sampling data (henceforth referred to as DSpat). They first assume that the locations of all individuals in the survey area (not just those which were observed) are a realisation of an inhomogeneous Poisson process which is a function of space. The authors then take the novel approach of allowing for separate (disjoint) regions of the survey area to have different detection functions associated with them. The sum of these detection functions is then used as a thinning of the Poisson process. The parameters are then found via standard maximum likelihood methods for point processes (see, e.g. Baddeley & Turner (2000)). In contrast to Hedley & Buckland (2004), parameters are estimated jointly so uncertainty from both the spatial pattern and the observation process is incorporated into variance estimates for the abundance. Concurrent estimation of the parameters also ensures that interactions between the thinning and underlying point process are estimated correctly. The authors also address the issue of overdispersion (commonly a symptom of animals or groups clustering), unmodelled by spatial covariates in a manner similar to that for GLMs (see *Recent Developments*, below, for another approach).

Niemi & Fernández (2010) also use Poisson processes but incorporate it into a fully Bayesian approach. Their intensity function takes the form of a product of a parametric function of the covariates and a mixture of Gaussian kernels as a spatial smooth. An appropriate degree of smoothing could be selected by putting prior distributions on the number and locations of the "knots" of the spatial smooth (the means of the Gaussian kernels) and then using reversible jump MCMC (Green (1995)). However, because the authors only include a single precision parameter for all of the kernels, small and large scale variation cannot both be accommodated. As in Johnson *et al.* (2010), the detection function was used as a thinning of the process, although (unlike DSpat) only one detection function was used across the whole region with known parameters. This means that detection function uncertainty is not incorporated in the spatial model.

Both of the above Poisson process models do not account for group size, both stating that this could be included by considering a marked point process (Cox & Isham (1980), Section 5.5). Both methods offer direct modelling of the point process, although with some drawbacks compared to the methodology of Hedley & Buckland (2004). It should be noted that the loss of efficiency from using a two-stage approach is not large (Buckland *et al.* (2004), p. 313). For these reasons, the article focuses on method of Hedley & Buckland (2004) and the advances which can be applied to their methodology.

Density surface modelling

We refer to the approach of Hedley & Buckland (2004) as *density surface modelling* (DSM). Rather than modelling the point process directly, the DSM approach uses the estimated abundance (of individuals or groups) as response for a spatially explicit model. DSMs can therefore be thought of as an extension to spatial models for strip transects (where the response is simply a count). The DSM approach is incorporated into the popular software package Distance (Thomas *et al.* (2010)).

First, consider conducting a strip transect survey. Strips are divided into contiguous *segments* (indexed by j), which are of length l_j; small enough such that the density does not vary appreciably within a segment. For each segment, the number of individuals observed (n_j) is used as the response. The count can then be modelled as a function of spatial and environmental covariates (the $\mathbf{z}_{jk}$ for k indexing the covariates: e.g. location, sea surface temperature, weather conditions) using a generalized additive model (GAM; e.g. Wood (2006)). The covered area enters the model as an offset (the area of segment j, $A_j = 2wl_j$, where w is the truncation distance). The model for the count per segment is:

$$\mathbb{E}(n_j) = \exp\left[\log_e(A_j) + \beta_0 + \sum_k f_k(\boldsymbol{z}_{jk})\right], \tag{1}$$

where the f_ks are smooth functions of the covariates in the GAM case and β_0 is an intercept term. The distribution of n_j can then be modelled as over-dispersed Poisson, negative binomial, or Tweedie (see *Recent developments*, below) distribution.

DSM with environmental-level covariates

Strip transects assume that detection within the segment is certain, to relax this assumption, if perpendicular distance is recorded, the per-segment abundance can be estimated and used as the response. A detection function is fitted to the distances using CDS or MCDS methods and, having calculated the probability of detection, n_j is replaced by a Horvitz-Thompson type estimator (Thompson (2002)) of abundance in the segment:

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

where $\hat{p}_j$ is the probability of detection in segment j (although $\hat{p}_j = \hat{p}$, $\forall j$ if there are no covariates other than distance in the detection function). 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 $s_{jr} = 1$, $\forall j, r$). It is possible that a bias is incurred by the group size (since larger groups are more visible), *Practical advice*, gives one method that can be used to deal with size bias in grouped populations.

Having estimated the response for the GAM, the following model is fitted:

$$\mathbb{E}(\hat{N}_j) = \exp\left[\log_e(A_j) + \beta_0 + \sum_k f_k(\boldsymbol{z}_{jk})\right], \tag{2}$$

where $\hat{N}_j$, as with n_j, is assumed to follow an overdispersed Poisson, negative binomial, or Tweedie distribution.

The above definition of the smooth terms is rather general because several covariates could be included in single smooth terms via tensor products of univariate bases (see Wood (2006), Section 4.1.8) or via multivariate spline bases (e.g. thin plate regression splines; Wood (2003)), as well as simple linear terms or random effects. A typical use of a bivariate spline in this setting is to smooth with respect to spatial coordinates of the segment centroids. Basis choice for spatial smooths is covered below. Note that even if location is not used, the model is still spatial (in some sense), because the covariates used in the GAM are spatially referenced.

Data collected as point transects can also be analysed by setting $A_j =$

$w\pi^2$, $\forall j$.

Figure 1 (top panel) shows the raw observations from the dolphin data, along with the transect lines, overlaid on the depth data. Figure 2 shows a GAM 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. Further discussion of the plots follows in *Practical advice*, below.

DSM with covariates at the observation level

The above model only considers the case where the covariates are measured only at the segment/point level. Often covariates ($\boldsymbol{\zeta}_{ij}$, for individual/group i in segment j) are collected on the level of individuals; for example sex, length or observer identity. In this case the probability of detection is a function of the individual level covariates $\hat{p}(\boldsymbol{\zeta}_i)$. Individual level covariates can be incorporated into the model by adopting the following estimator of the per-segment abundance:

$$\hat{N}_j = \sum_{r=1}^{R_j} \frac{s_{jr}}{\hat{p}(\boldsymbol{\zeta}_{ij})}.$$

Estimating abundance and investigating relationships

Our aims in a DSM analysis are usually two-fold: estimating overall abundance and investigating the relationship between abundance and environmental covariates.

To calculate an abundance estimate for some region of interest, the necessary covariates (those included in the model) must be available for the whole of the region, and they must also be available at the required resolution (using prediction grid cells that are smaller than the resolution of the spatially referenced data will not have an effect on abundance/density estimates). Having acquired the relevant data and calculated the associated areas of the prediction cells, predictions can be made for the particular covariate levels and abundance estimates calculated from summing predicted values over the prediction grid cells.

As with any predictions which are outside of the range of the data, one should heed the usual warnings regarding extrapolation. For example, in an offshore study the effect of a continental shelf maybe cause significant issues if there was not search effort on both sides of the shelf. 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 is required to avoid prediction outside the range of the data.

Abundance estimation is not the only information contained in these models. By looking at plots of marginal smooths of the spatially referenced covariates, one can begin to understand the relationships between the covariates and abundance. Going back to the dolphin data, we can see the effect of depth on abundance in Figure 3. There we can see that there is a large depth effect between 0 and 500m which then seems to level off (a straight line could be drawn inside the confidence band (dashed line)), indicating that the dolphins prefer water deeper than 500m. 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.

Variance estimation

Estimating the variance of abundances calculated using DSM is not straight forward as 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, adjacent segments are likely to be correlated; failing to account for this spatial autocorrelation will lead to artificially low variance estimates and hence misleadingly narrow confidence intervals.

Resampling-based methods

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 then follows the following steps:

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 per-location and abundance variance can be calculated in the usual manner. The total variance of the abundance estimate 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; Seber (1982)). This assumes that the two components of the variance are independent and the method does not not take into account spatial autocorrelation (the individual segments are treated as independent).

The above procedure assumes that there is no correlation in space between segments and that residuals can be swapped around. Clearly if many animals are observed in a segment then we would expect there to be a relatively high level in the next segment (especially because the segments are defined after the survey). A moving block bootstrap (MBB) can account for some of the 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 $1, \ldots, 5$, the second block is segments $2, \ldots, 6$, and so on). Then, at step (2) above, resamples are taken of the blocks (i.e. groups of segments together) rather than individual segments within the transects. Using blocks should account for some of the autocorrelation between the segments, inflating the variances accordingly. The moving block bootstrap can also be modified to take into account detection function uncertainty by generating new distances from the fitted detection function and then recalculating the offset by fitting a detection function to the new data.

Variance propagation

Rather than using a bootstrap, Williams *et al.* (2011) calculate the variance without having to refit the model many times. Their method incorporates the uncertainty in the estimation of the detection function into the variance of the spatial model, albeit only in the case where covariates are measured at a point/segment level only. Their procedure is as follows:

1. Fit the model described in eqn 2.
2. Re-fit the model with an additional random effects term. This term characterises the uncertainty in the estimation of the detection function (via the uncertainty of the probability of detection, $\hat{P}_a$).
3. Variance estimates of the abundance calculated using standard GAM theory (Wood (2006), page 245) from the model will include uncertainty from the estimation of the detection function.

We consider propagating the uncertainty in this manner not only to be more computationally efficient but also preferable from a technical perspective. The bootstrap does not fully account for spatial autocorrelation, assuming that the residuals are exchangeable when they are not will lead to wider confidence intervals. The experience of the authors has been that in simulation the confidence intervals produced are narrower (than their bootstrap equivalents), while maintaining good coverage.

Visualising uncertainty

There are several ways to visualise the uncertainty measures calculated above. For the bootstrap methods, if at each round of the bootstrap the predicted values are stored per prediction grid cell, the coefficient of variation can be calculated per cell and then displayed. Figure 4 shows maps of the coefficient of variation for the model which includes both location and depth covariates. The top panel shows the result of running 1000 bootstrap replications including detection function uncertainty as above. The bottom panel shows the same plot but using the variance propagation method.

Recent developments

Edge effects

Recent work (Ramsay (2002), Wang & Ranalli (2007), Wood *et al.* (2008), Scott-Hayward et al (in prep) and Miller and Wood (submitted)) has highlighted the need to take care when smoothing over areas with complicated boundaries; for example, if the survey area includes rivers, peninsulae or islands. If two parts of the domain (either side of a peninsula, say) are inappropriately linked by the model (the distance between the points is measured "as the crow flies", rather than "as the fish swims") then the boundary feature can be "smoothed across" leading to incorrect inference. Ensuring that a realistic spatial model has been fit to the data (and, for example, that whales have not been estimated to dwell on land) is essential for valid inference. The soap film smoother of Wood *et al.* (2008) is particularly appealing as the model jointly estimates boundary conditions for a complex study area along with the "interior" smooth. This can be particularly 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 (in prep.) show that when using global smoothers, smoothing towards the plane can cause the fitted surface to "curlup" as predictions move further away from the data. They suggest the use of *Duchon splines* (a generalisation of thin plate regression splines) to alleviate the problem by smoothing toward the intercept.

Tweedie distribution

The Tweedie distribution offers a very flexible alternative to the quasi-Poisson distribution is the usual response distribution when modelling count data (Candy (2004)). Through the parameter p, many common distributions arise; varying p 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)). Although it is possible to perform optimization to find p, this is generally seen as unnecessary as the distribution does not change appreciably when p is changed by less than 0.1 (therefore trial and error is usually reasonable). Mark Bravington (pers. comm.) suggested plotting the square root of the absolute value of the residuals and if this plot is flat a "correct" p has been found. Additionally he suggests a value of 1.5/1.6 for p for fisheries and 1.2 marine mammal work is generally acceptable.

Practical advice

Figure 5 shows a flow diagram of the modelling process for creating a density surface model for distance sampling data. 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 start with a detection function without covariates and a simple smooth of spatial location and then add in more complicated features such as covariates in the detection function, or using a soap film smoother (perhaps afterwards dropping the location term). Model discrimination can be performed for the detection function using goodness-of-fit tests (Buckland *et al.* (2004) and AIC. For the spatial model, generalized cross validation (GCV) score and percentage deviance explained are useful metrics, we also highly recommend the use of standard GAM diagnostic plots. An example of such plots is given in Figure 6 along with a description of their uses.

In the dolphin analysis, we include a smooth of location. This not only 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 Figure 2 and the plot of the depth in Figure 1, the plot of the smooth of depth alone looks very similar to the raw plot of the depth data. A smooth of an environmental-level covariate such as depth can be very useful for assessing the relationships between abundance and the covariate, although investigators should be cautious of interpreting artefacts of correlations between covariates and abundance especially if there is incomplete coverage of covariate values.

In the analysis we have converted from latitude and longitude to metres from the point (27.01, -88.3). This is because the bivariate smoother which we use (the thin plate spline, Wood (2003)) is isotropic: it treats the wigglyness of the smoother in each direction as equal: a move of 1 degree in latitude is not the same as a move of 1 degree in longitude, the move to meters from the centre of the study area is sensible (using SI units removes the need for conversion later). Limiting the “wigglyness” of smooths of spatial location can be a useful way of restricting their influence whilst still allowing them to “mop up” the residual spatial correlation in the data.

If animals occur in groups rather than individually a size bias can occur due to larger groups being more visible than smaller groups. Bias due to group size can be accounted for in a DSM analysis as in Buckland *et al.* (2001), Section 4.8.2.4: regressing evaluations of the fitted detection function onto the logarithm of group size. The bottom right panel of figure 1 shows a such a plot with the regression line.

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 depend upon a random sample design, and therefore data from "platforms of opportunity" (Williams *et al.* (2006)) can be used to make inference.

Unbiased estimates are dependent upon either a) distribution of sampling effort being random throughout the study area (for design-based inference) or b) the model is correct (for model-based inference). It is easier to have confidence in the former than in the latter because correctness cannot be demonstrated in biological systems. Nevertheless model-based inference will play an increasing role in population assessment as we attempt to squeeze more information from the data we gather.

The field is quickly evolving to allow modelling of more complex data however the basic principle remains as in Hedley & Buckland (2004), albeit with various additions to the modelling process. We expect to see large advances two areas: temporal inferences and the handling of spatial autocorrelation. These should become more mainstream as modern spatio-temporal modelling techniques are adopted. Petersen *et al.* (2011) provide a very basic framework for temporal modelling; their model includes extra smooth terms for their spatial and depth smooth terms after the construction of an offshore windfarm which are included via an indicator. 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 allows observations to cluster according to one of several states (e.g. "feeding" or "transit") taking into account short-term agglomerations ("hot spots").

Acknowledgments

DLM wishes to thank Mark Bravington and Sharon Hedley for their help and patience in explaining and providing code for their variance propagation method and alerting him to the existence of the Markov modulated Poisson process.

References

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

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

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

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

Cleveland, W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, pp. 829–836.

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

Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

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.

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.

Niemi, A. & Fernández, C. (2010) Bayesian Spatial Point Process Modeling of Line Transect Data. *Journal of Agricultural, Biological, and Environmental Statistics*, **15**, 327–345.

Petersen, I.K., MacKenzie, M., Rexstad, E., 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, CREEM Technical Report.

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

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.

Seber, G.A.F. (1982) *The Estimation of Animal Abundance and Related Parameters.* Blackburn Pr. ISBN 9781930665552.

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.

Thompson, S.K. (2002) *Sampling.* Wiley, 2nd edn. ISBN 9781118162965.

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., 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 Top: the survey area, transect centrelines and observations with size of circle corresponding to the group size overlaid onto depth data; bottom left, histogram of observed distances with fitted detection function; bottom right, plot of evaluations of the fitted detection function at given distances versus the logarithm of group size with linear trend showing the relation between probability of detection (given distance) and group size.

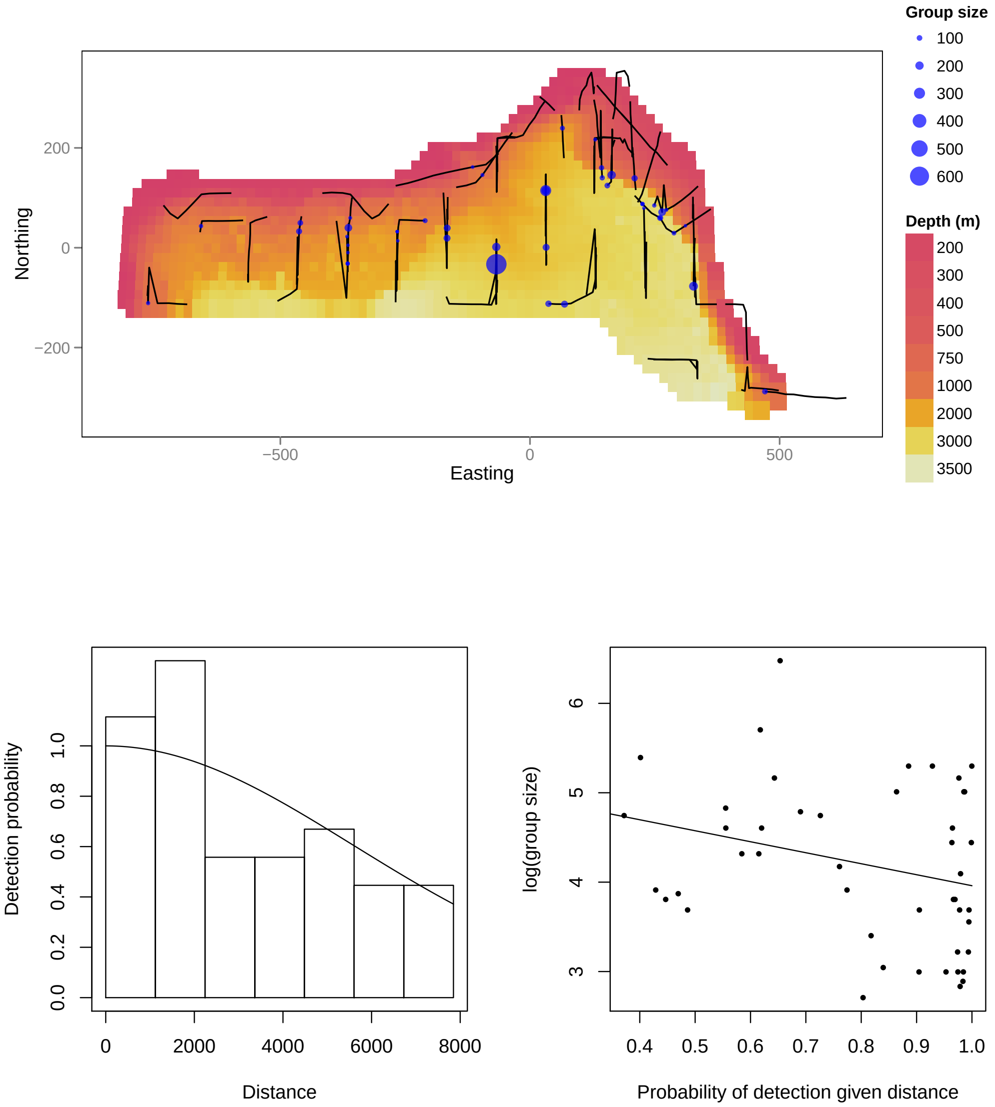


Fig. 2 Predictions for the dolphin data. Top: Predictions from the model using only depth as an explanatory variable, bottom: the model using both depth and location.

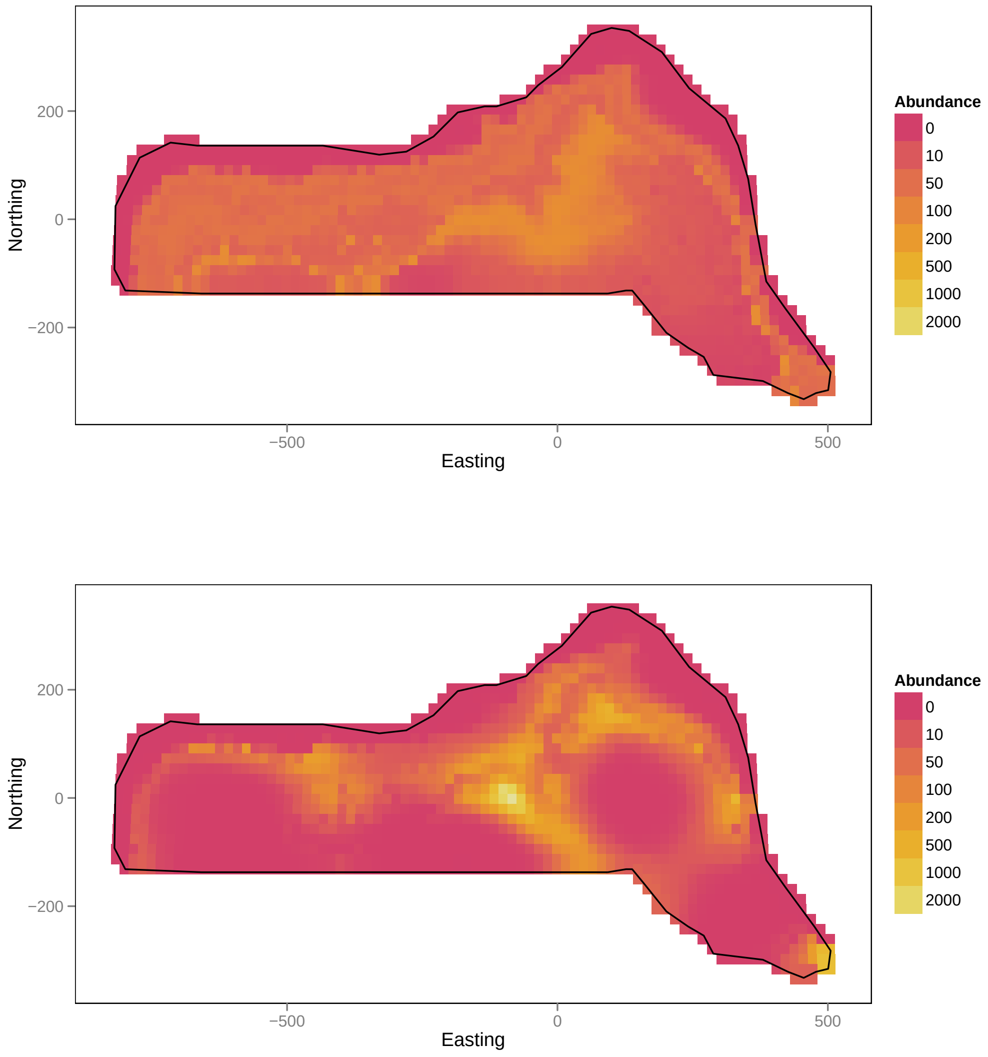


Fig. 3 Plot of the effect on the response of depth, note that it is possible to draw a straight line between 750m and 3000m within the confidence band, so the wiggles in the smooth may not be indicative of any relationship. What is clear is that there is some effect up to about 500m. The number in brackets on the y axis indicates the effective degrees of freedom of the smooth term. The rug ticks at the bottom of the plot indicate we have good coverage of the range of depth values in the survey area.

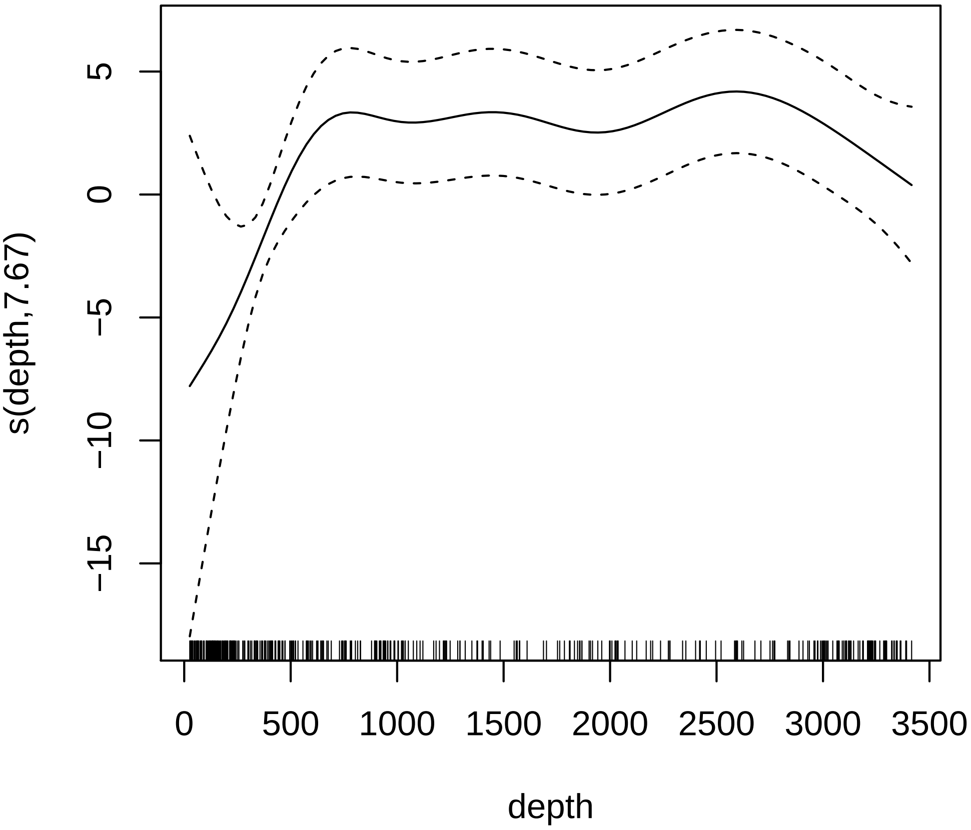


Fig. 4 Plot of coefficient of variation map, showing the uncertainty in the fitted model with observations and transect lines overlaid. Uncertainty was estimated using the variance propagation method of Williams *et al.* (2011).

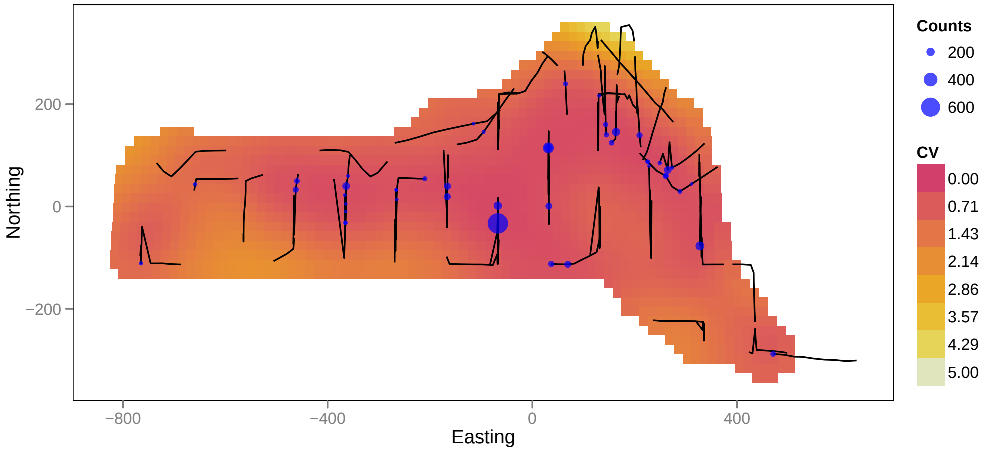


Fig. 5 Flow diagram showing the modelling process for creating a density surface model.

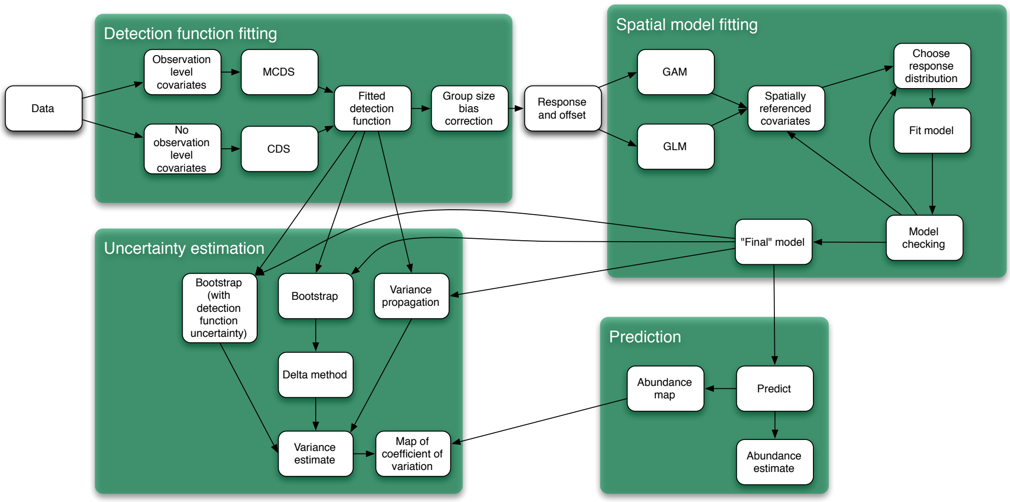
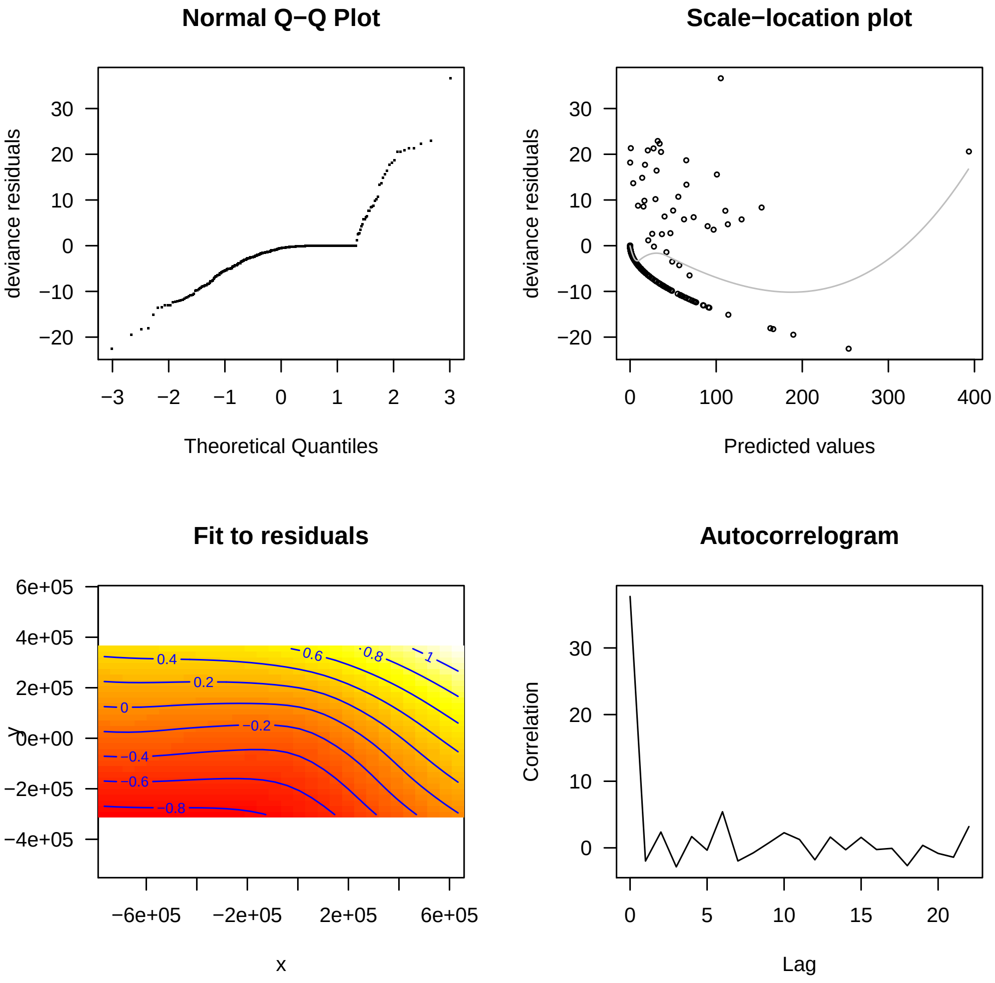

Detection function fitting
Data
Observation level covariates
No observation level covariates
MCDS
CDS
Fitted detection function
Group size bias correction
Response and offset
Spatial model fitting
GAM
GLM
Spatially referenced covariates
Choose response distribution
Fit model
Model checking
"Final" model
Uncertainty estimation
Bootstrap (with detection function uncertainty)
Bootstrap
Variance propagation
Delta method
Variance estimate
Map of coefficient of variation
Prediction
Abundance map
Predict
Abundance estimate

Fig. 6 Example of model diagnostics for the model which included both location and depth covariates for the dolphin data when a quasi-Poisson response distribution was specified. From top left clockwise: 1) normal Q-Q plot showing a problematic fit (the "elbow" in the points), 2) plot of (deviance) residuals against predicted values highlighting outliers and LOESS smooth (Cleveland (1979)) through the point overlaid, 3) a smooth of location fitted to the residuals showing some pattern left in the data and 4) the autocorrelogram.


Normal Q−Q Plot
deviance residuals
Theoretical Quantiles
Scale−location plot
deviance residuals
Predicted values
Fit to residuals
6e+05
4e+05
2e+05
0e+00
−2e+05
−4e+05
y
−6e+05
−2e+05
2e+05
6e+05
x
Autocorrelogram
Correlation
Lag