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*], **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: 26th November 2012

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. Model-based inference may also be used to obtain abundance estimates from non-randomly designed surveys.
2. Density surface modelling achieves both of the above aims. DSMs combine distance sampling to account for uncertain detection and a spatial model for the effects of environmental covariates.
3. We offer a comparison of recent advances in the field and consider the likely directions of future research. In particular we consider spatial modelling techniques that may be advantageous to applied ecologists.
4. The methods discussed are freely available in R packages developed by the authors.

Keywords: distance sampling, line transect sampling, point transect sampling, population abundance, 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, altitude or (if at sea) bathymetry. Spatial models allow for the vast databases of spatially-referenced data to be harnessed, allowing for interactions between environmental covariates and population densities to be investigated. 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 include any model that includes spatially referenced covariates, not just smooths of location. This article concerns combining spatial modelling techniques with distance sampling (Buckland *et al.*, 2001, 2004).

Distance sampling takes plot sampling (counting the individuals or groups of objects in a strip or circle) and extends it to the case where detection is not certain. Observers travel along transect centre lines or stand at points and record the distance from the centre line or point to the object of interest (y). These distances are used to estimate the *detection function*, $g(y)$ (bottom left panel, figure 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 animal/observer specific covariates (multiple covariate distance sampling, MCDS; Marques *et al.*, 2007). From the fitted detection function, the probability of detection can be calculated.

The estimated probability that an animal is detected, $\hat{p}_i$, can then be used to calculate abundance as

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

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 individuals (Buckland *et al.*, 2001, Chapter 3). In general distance sampling is more efficient than plot sampling since all objects observed are recorded and only later discard observations deemed to far away (outside of the *truncation distance*).

When fitting the detection function 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). It is usually possible to design surveys such that a homogenous process can be assumed so, with respect to the line, objects are distributed uniformly. This can be achieved by e.g., 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).

Estimators such as eqn. 1 are referred to as *design-based* since they rely on the design of the study to ensure inference is valid. This article focusses on *model-based* inference. Using spatially explicit models one can investigate the response of biological populations to biotic and abiotic covariates which vary over the survey area. Modelling the spatial process also enables the use data from badly designed or opportunistic surveys, for example incidental data arising from "ecotourism" cruises can be included in analyses (Williams *et al.*, 2006).

Our aims in a DSM analysis are usually two-fold: (i) estimating overall abundance and (ii) investigating the relationship between abundance and environmental covariates. 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 terrestrial study, habitat may cause significant issues if there was not search effort in all habitats. 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.

The article focuses on those recent advances in spatial modelling of distance sampling data which are of most utility to applied ecologists. These new methods are available in the R packages `Distance` and `dsm`, and will soon be available in the popular Windows application Distance (Thomas *et al.*, 2010).

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. 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 bathymetry data were also available as covariates for the analysis. A complete example analysis is provided as an online appendix.

The rest of the article is structured as follows: we first describe the density surface modelling approach of Hedley & Buckland (2004), explain how to estimate abundance and uncertainty. We then describe recent advances, practical advice regarding the model fitting, formulation and checking. Before concluding, we look at two alternative (but less mature) methods which take a rather more direct approach to modelling spatial distance sampling data.

Density surface modelling

This section focuses on modelling the abundance/density estimation stage of distance sampling, using the "count model" of Hedley & Buckland (2004) which we refer to as *density surface modelling* (DSM). 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; small enough such that the density does not vary appreciably within a segment (usually making the segments approximately square, $2w \times 2w$, is sufficient). The general idea is to model the count or estimated abundance as a smooth function of covariates using a generalized additive model (GAM; Wood, 2006). For each segment or point, the response is modelled as a function of *covariates at the environmental level* (the z_{jk} with k indexing the covariates, e.g., location, sea surface temperature, weather conditions). The covered area enters the model as an offset: the area surveyed at segment j is $A_j = 2wl_j$ and at point j is $A_j = w\pi^2$ (where w is the truncation distance).

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 covered 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 (i.e., $\hat{p}_j = \hat{p}$, $\forall j$). The distribution of n_j can then be modelled as overdispersed Poisson, negative binomial, or Tweedie distribution (see *Recent developments*, below).

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.

Abundance estimation is not the only information contained in these models. Plots of marginal smooths of the spatially referenced covariates show the relationships between the covariates and abundance. The effect of depth on abundance for the dolphin data can be seen in Figure 3. Between 0 and 500m there is a depth effect which then seems to level off (a straight line could be drawn inside the confidence band). This may indicate that the dolphins prefer water deeper than 500m, however the usual caveats inherent in interpreting results from observational studies apply.

Estimated abundance as response

An alternative to modelling counts would be to use the per-segment/circle abundance can be estimated using distance sampling methods and the estimated counts used 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.

DSM with covariates at the observation level

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

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

It is possible that bias is incurred by larger groups and therefore more visible groups. Including group size as a covariate in the detection function and fitting the above model is one solution. See *Practical advice*, below, for more information on grouped populations.

By not including an offset, but instead dividing the count (or estimated abundance) by the area, we can also model density rather than abundance. We concentrate on abundance here, see Hedley & Buckland (2004) for further details.

Prediction

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 that 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). The areas of the segments/points are included as an offset in the model, so the area of the prediction cells must be included in the prediction data. Predictions can be made for the particular covariate levels and abundance estimates calculated for a particular area by summing predicted values over corresponding grid cells.

VARIANCE ESTIMATION

Estimating the variance of abundances calculated using a DSM is not straight forward: 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.

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 is calculated 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 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 however, 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 (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. However, since 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.

Recent developments

GAM uncertainty and variance propagation

Rather than using a bootstrap, one can use GAM theory to construct uncertainty estimates for abundance estimates (and smooth terms) in a DSM. This merely requires that we take a Bayesian view and use the distribution of the parameters in the model (further information can found in Wood, 2006, page 245). Such an approach means that the variance can be calculated without having to refit the model many times.

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 with environmental level covariates. Their procedure is as follows:

1. Fit a density surface model.
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, $\hat{p}$).
3. Variance estimates of the abundance calculated as usual for the GAM 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. In simulation the confidence intervals produced are narrower (than their bootstrap equivalents), while maintaining good coverage.

A common way to visualise uncertainty in a DSM is to plot the per-cell coefficient of variation by dividing the standard error for each cell by its predicted abundance. Figure 4 shows a map of the coefficient of variation for the model which includes both location and depth covariates using the variance propagation method.

Edge effects

Recent work (Ramsay, 2002; Wang & Ranalli, 2007; Wood *et al.*, 2008; Scott-Hayward *et al.*; 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 mountain, say) are inappropriately linked by the model (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 fit to the data 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.* show that when using global smoothers, smoothing towards the plane can cause the fitted surface to "curl-up" 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 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). Although it is possible to perform optimization to find λ, this is generally seen as unnecessary as the distribution does not change appreciably when λ 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 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

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 obtain a detection function which fits the data as well as possible and only after a satisfactory detection function has been obtained, begin spatial modelling. A simple smooth of spatial location will given an idea of the distribution of the population, more covariates can then be added. A useful feature is the additional shrinkage available for GAMs which allow smooth terms to be removed from the model during fitting. Model selection can be performed for the detection function using AIC and model checking using goodness-of-fit tests given in Buckland *et al.* (2004). For the spatial model, smooth terms can be selected using as well as p-values. Generalized cross validation (GCV) score (or related metrics such as UnBiased Risk Estimator or REstricted Maximum Likelihood score; UBRE and REML, respectively) and percentage deviance explained are useful for model selection. We also highly recommend the use of standard GAM diagnostic plots. Wood (2006), Chapter 5, provides practical information on GAM model selection and fitting.

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. Caution should be employed when interpreting smooth relationships and abundance estimates, especially if there is poor coverage of covariate values. For example if there is a large agglomeration of individuals at a a high value of depth but no further observations occur at such a high value, then investigators should be skeptical of any relationship. For this reason a smooth of space is recommended for inclusion in candidate models. 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.

In the analysis we have converted from latitude and longitude to kilometres from (27.01, -88.3), 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. Moving 1 degree in latitude is not the same as moving 1 degree in longitude, so using kilometres from the centre of the study area is sensible (using SI units throughout also removes the need for conversion).

If animals occur in groups rather than individually, bias can be incurred due to larger groups being more visible than smaller groups. Bias due to group size can be assessed by regressing evaluations of the fitted detection function onto the logarithm of group size, then comparing the expected and observed values of the group size, if there is a large difference then it may be necessary to include size as a covariate in the detection function. The bottom right panel of figure 1 shows a such a plot with the regression line overlaid.

Direct modelling of the spatial point process

Rather than use a GAM to model the spatially explicit part of the model, two recent articles have modelled the process using point processes (Cox & Isham, 1980). In both cases the density of object is governed by a spatially-varying *itensity function*, which can include covariates in a similar manner to the GAM.

Johnson *et al.* (2010) propose a point process-based model for distance sampling data (known as DSpat). They first assume 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 are then estimated via standard maximum likelihood methods for point processes (Baddeley & Turner, 2000). In contrast to Hedley & Buckland (2004), all parameters are estimated jointly so uncertainty from both the spatial pattern and the detection function is incorporated into variance estimates for 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 address the issue of overdispersion unmodelled by spatial covariates using a post-hoc correction factor.

Niemi & Fernández (2010) also use Poisson processes but incorporate them into a fully Bayesian approach. Unlike Johnson *et al.* (2010) model fitting proceeds in two stages: first the detection function is fitted, then the spatial model (via MCMC) assuming the detection function parameters are known, so 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 DSM is not large (Buckland *et al.*, 2004, p. 313) because distances contain little information about spatial variation because transects are very thin compared to their lengths and circles are very small compared with study area.

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.

Unbiased estimates are dependent upon either (i) distribution of sampling effort being random throughout the study area (for design-based inference) or (ii) the model is correct (for model-based inference). It is easier to have confidence in the former than in the latter because our models are always wrong. 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 building on the basic ideas of density surface modelling. We expect to see large advances in two areas: temporal inferences and the handling of 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 smooth terms both before and after the construction of an offshore windfarm. 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 (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.

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.
URL `http://www.iwcoffice.org/_documents/sci_com/sc61docs/SC-61-IA14.pdf`

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.

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.

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.

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.

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.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.

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.

Scott-Hayward, L.A.S., MacKenzie, M.L., Donovan, C.R., Walker, C.G. & Ashe, E. (????) Complex Region Spatial Smoother (CReSS). pp. 1–31. URL `http://research-repositoryst-andrewsacuk/handle/10023/2048`

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.

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.

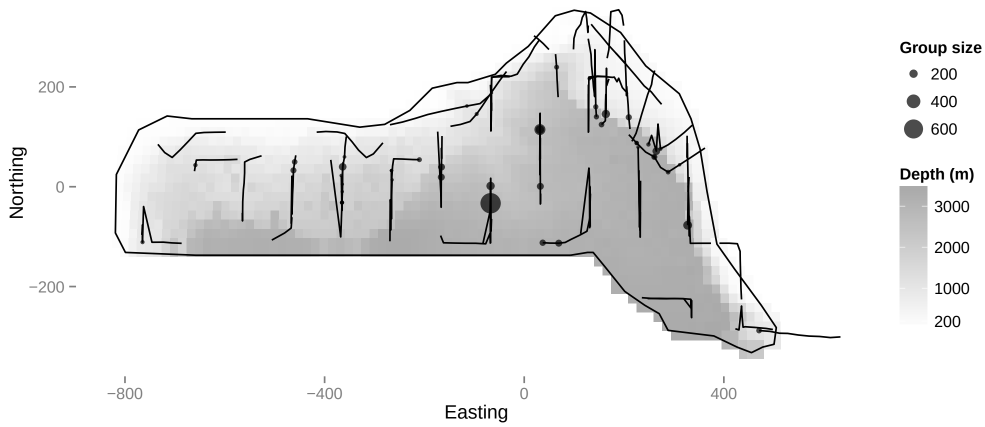
Wood, S.N. (2003) Thin plate regression splines. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **65**, 95–114.

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.



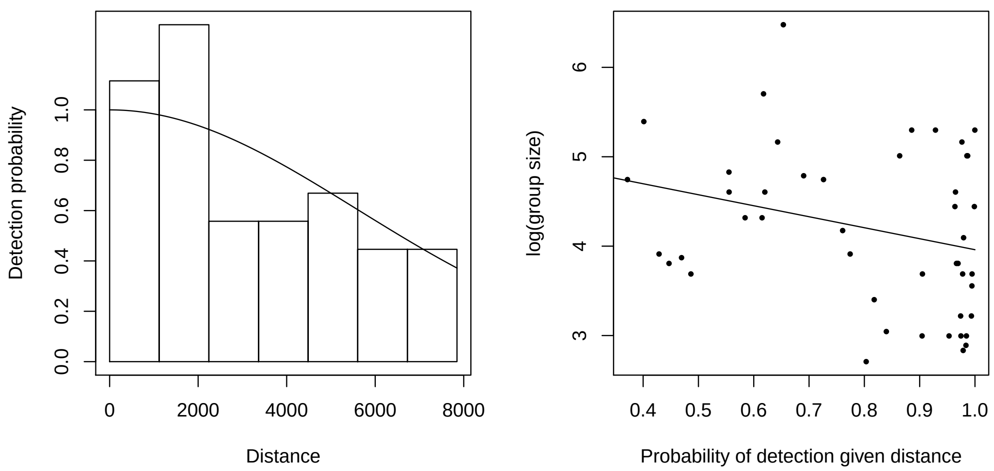
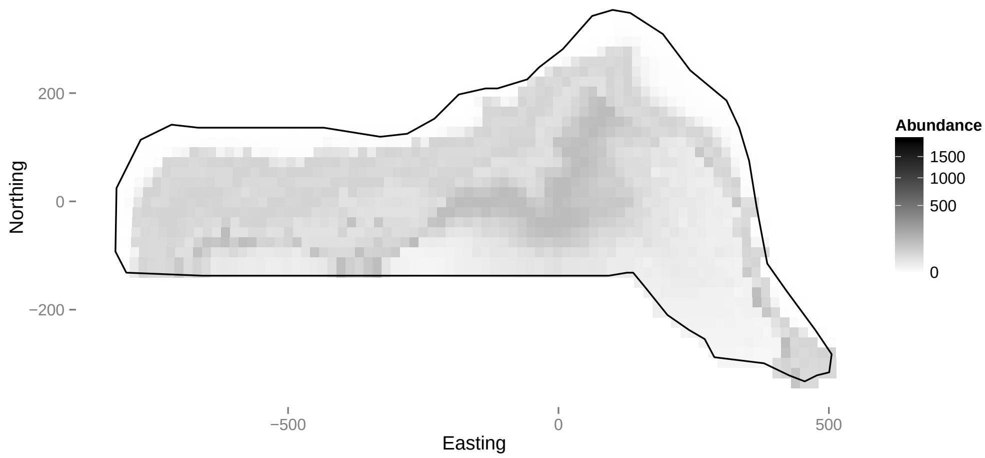


Fig. 2 Predictions for the dolphin data. Top: Predictions from the model using only depth as an explanatory variable, bottom: the model using both depth and location.



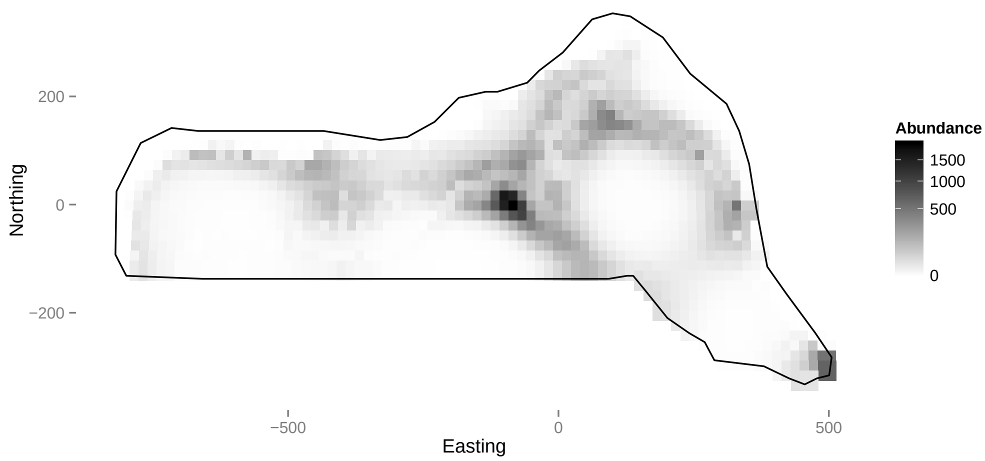


Fig. 3 Plot of the effect on the response of depth, note that it is possible to draw a straight line between 750m and 3000m within the confidence band (between the dashed lines), so the wiggles in the smooth may not be indicative of any relationship. What is clear is that there is some effect up to about 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.

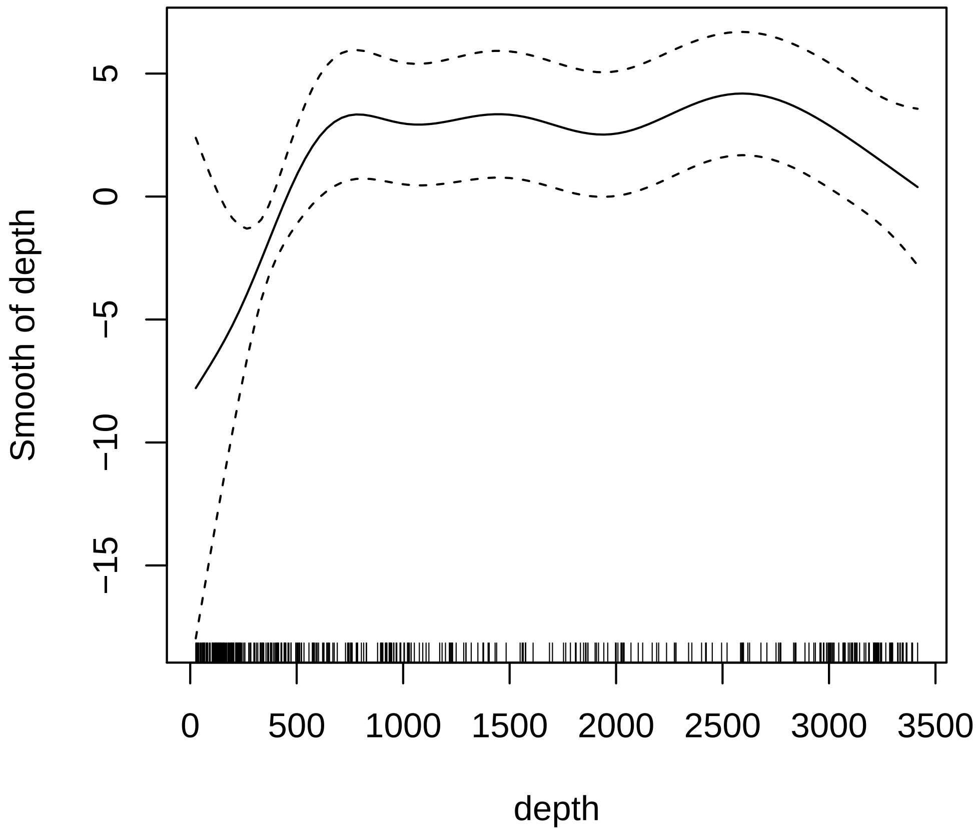


Fig. 4 Plot of coefficient of variation map for the model with smooths of both depth and location. Uncertainty was estimated using the variance propagation method of Williams *et al.* (2011).

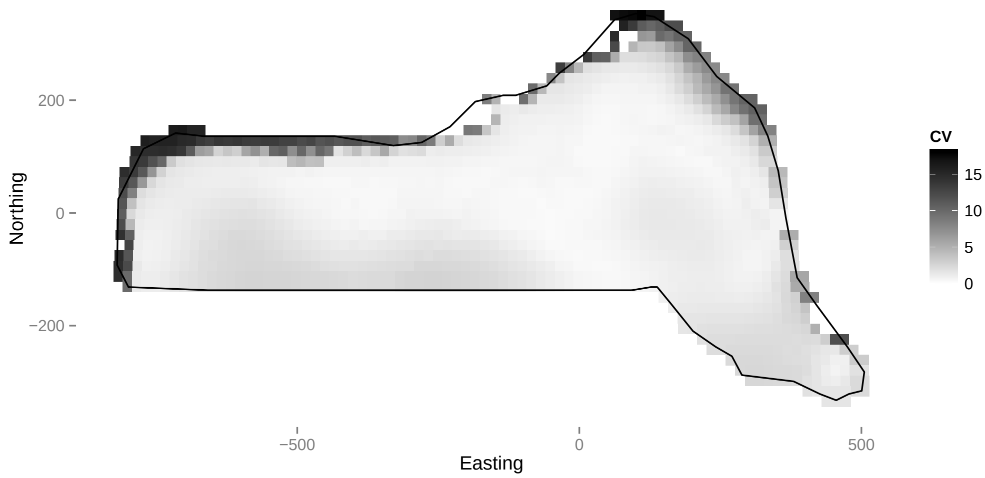


Fig. 5 Flow diagram showing the modelling process for creating a density surface model.