

1 **Running title:** Spatial models for distance sampling
2 **Number of words:** ~4107
3 **Number of tables:** 0
4 **Number of figures:** 5
5 **Number of references:** 31

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

8 **David L. Miller^{1*}, M. Louise Burt²,**
9 **Eric A. Rexstad², Len Thomas².**

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

14 ***Correspondence author. dave@ninepointeightone.net**

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.
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 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 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.
5. Density surface modelling enables applied ecologists to reliably estimate abundances and create maps of animal/plant distribution. Such models can also be used to investigate the relationships between distribution and environmental covariates.

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

38 Introduction

39 When surveying biological populations it is increasingly common to record
40 spatially referenced data, for example: coordinates of observations, habitat
41 type, elevation or (if at sea) bathymetry. Spatial models allow for vast data-
42 bases of spatially-referenced data (e.g. OBIS-SEAMAP, Halpin *et al.*, 2009)
43 to be harnessed, enabling investigation of interactions between environmental
44 covariates and population densities. Mapping the spatial distribution of a
45 population can be extremely useful, especially when communicating results
46 to non-experts. Recent advances in both methodology and software have
47 made spatial modelling readily available to the non-specialist (e.g., Wood,
48 2006; Rue *et al.*, 2009). Here we use the term “spatial model” to include any
49 model that includes spatially referenced covariates, not just smooths of loc-
50 ation. This article is concerned with combining spatial modelling techniques
51 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 travel 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*, $g(y)$ (bottom left panel, Fig. 1), by modelling the decrease in detectability with increasing distance from the line or point (conventional distance sampling, CDS). The detection function may also include covariates (multiple covariate distance sampling, MCDS; Marques *et al.*, 2007). 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

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

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 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 in order to ensure the assumption of certain detection is met. In contrast, distance sampling uses the observations that would have been discarded to model the detection (typically data are discarded beyond a given *truncation distance* during analysis).

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 that (with respect to the line) objects are distributed uniformly. This can be achieved by ensuring that transects randomly located.

Estimators such as eqn (1) rely on the design of the study to ensure that abundance estimates over the whole study area (scaling up from the covered region) are valid. This article focusses on *model-based* inference to extrapolate to a larger study area. Specifically, we consider the use spatially explicit models to investigate the response of biological populations to biotic and abi-

otic covariates that vary over the survey area. 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 lower variances. 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 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 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 avoid prediction outside the range of the data.

This article reviews the current landscape of spatial modelling of distance sampling data, illustrating some recent developments most useful to applied ecologists. The methods discussed have 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 (available on CRAN) `dsm` (and will be incorporated in to Distance).

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 pantropical spotted dolphins (*Stenella attenuata*) in the Gulf

98 of Mexico. 47 groups of dolphins were observed; group size was recorded, as
99 well as the Beaufort sea state at the time of the observation. Coordinates for
100 each observation and bathymetry data were also available as covariates for
101 the analysis. A complete example analysis is provided as an online appendix.
102 The data used in the analysis are available in the *dsm* package and Distance.

103 The rest of the article follows this structure: we first introduce the density
104 surface modelling approach of Hedley & Buckland (2004); explain how to
105 estimate abundance and uncertainty; describe recent advances and provide
106 practical advice regarding model fitting, formulation and checking. Before
107 concluding, we review alternative (but less mature) methods which take a
108 more direct approach to modelling spatial distance sampling data.

109 Density surface modelling

110 This section focuses on modelling the density/abundance estimation stage
111 of distance sampling, using the “count model” of Hedley & Buckland (2004),
112 which we refer to as *density surface modelling* (DSM). Both line and point
113 transects can be used but if lines are used then they are split into con-
114 tiguous *segments* (indexed by j), which are of length l_j . Segments should
115 be small enough such that the density does not vary appreciably within a
116 segment (usually making the segments approximately square, $2w \times 2w$, is
117 sufficient). Count or estimated abundance is then modelled as a smooth
118 function of covariates using a generalized additive model (GAM; e.g. Wood,
119 2006). For each segment or point, the response is modelled as a function of
120 environmental covariates that are measured at the segment/point level (the

121 z_{jk} with k indexing the covariates, e.g., location, sea surface temperature,
 122 weather conditions). The covered area enters the model as an offset: the
 123 area covered at segment j is $A_j = 2wl_j$ and at point j is $A_j = w\pi^2$ (where w
 124 is the truncation distance).

125 COUNT AS RESPONSE

The model for the count per segment is:

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

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

133 Fig. 1 (top panel) shows the raw observations of the dolphin data, along
 134 with the transect lines, overlaid on the depth data. Fig. 2 shows a GAM
 135 fitted to the dolphin data, the top panel shows predictions from a model
 136 where depth was the only covariate, the bottom panel shows predictions
 137 where a (bivariate) smooth of spatial location was also included.

138 Abundance estimation is not the only information that can be derived
 139 from these models. Plots of marginal smooths of the spatially referenced
 140 covariates show the relationships between the covariates and abundance. The

141 effect of depth on abundance for the dolphin data can be seen in Fig. 3.

142 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:

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

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

The following model is then fitted:

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

145 where \hat{N}_j , as with n_j , is assumed to follow an overdispersed Poisson, negative
146 binomial, or Tweedie distribution (see *Recent developments*, below). Note
147 that the offset is now the area rather than effective area of the segment/point.

148 *DSM with covariates at the observation level*

149 The above models consider the case where the covariates are measured at
150 the segment/point level. Often covariates (z_{ij} , for individual/group i and
151 segment/point j) are collected on the level of observations; for example sex,
152 group size or observer identity. In this case the probability of detection is a
153 function of the individual level covariates $\hat{p}(z_i)$. Individual level covariates

154 can be incorporated into the model by adopting the following estimator of
 155 the per-segment abundance:

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

156 By not including an offset, but instead dividing the count (or estimated
 157 abundance) by the area of the segment, we can also model density rather
 158 than abundance. We concentrate on abundance here, see Hedley & Buckland
 159 (2004) for further details on modelling density.

160 PREDICTION

161 To calculate an abundance estimate for a region of interest, the environ-
 162 mental covariates included in the model must be available at each prediction
 163 point at the required resolution (using prediction grid cells that are smal-
 164 ler than the resolution of the spatially referenced data have no effect on
 165 abundance/density estimates). The areas of the prediction cells must also be
 166 included in the prediction data. Predictions are made for the each grid cell
 167 using covariate values associated with each cell and abundance estimates are
 168 produced for a given region by summing predicted values over corresponding
 169 grid cells.

170 VARIANCE ESTIMATION

171 Estimating the variance of abundances calculated using a DSM is not straight-
 172 forward: uncertainty from the estimated parameters of the detection function
 173 must be incorporated into the spatial model. A second consideration is that

174 in a line transect survey, adjacent segments are likely to be correlated; failure
175 to account for this spatial autocorrelation will lead to artificially low variance
176 estimates and hence misleadingly narrow confidence intervals.

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

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

- 182 1. Resample (with replacement) the per-segment residuals, store the val-
183 ues in \mathbf{r}_b .
- 184 2. Refit the model but with the response set to $\hat{\boldsymbol{\eta}} + \mathbf{r}_b$ (where $\hat{\boldsymbol{\eta}}$ are the
185 fitted values from the original model).
- 186 3. Take the predicted values for the new model and store them.

187 From the predicted values stored in the last step the variance originating in
188 the spatial part of the model can be calculated. The total variance of the
189 abundance estimate (over the whole region of interest or sub-areas) can then
190 be found by combining the variance estimate from the bootstrap procedure
191 with the variance of the probability of detection from the detection function
192 model (using the delta method which assumes that the two components of
193 the variance are independent; Seber, 1982).

194 The above procedure assumes that there is no correlation in space between
195 segments however, if many animals are observed in a particular segment then
196 we might expect there to be high numbers in the adjacent segments. A

197 moving block bootstrap (MBB; Efron & Tibshirani, 1993, Section 8.6) can
 198 account for some of this spatial autocorrelation in the variance estimation.
 199 The segments are grouped together into overlapping blocks, (so if the block
 200 size is 5, block one is segments 1, \dots , 5, block two is segments 2, \dots , 6, and so
 201 on). Then, at step (2) above, resamples are taken of the blocks (contiguous
 202 collections of segments) rather than individual segments within the transects.
 203 Using blocks should account for some of the autocorrelation between the
 204 segments, inflating the variances accordingly. However, because the block size
 205 dictates the maximum amount of spatial autocorrelation accounted for, this
 206 may not fully account for the autocorrelation. These bootstrap procedures
 207 can also be modified to take into account detection function uncertainty by
 208 generating new distances from the fitted detection function and then re-
 209 calculating the offset by fitting a detection function to the new distances.

210 Recent developments

211 *GAM uncertainty and variance propagation*

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

218 Williams *et al.* (2011) go a step further and incorporate the uncertainty
 219 in the estimation of the detection function into the variance of the spatial

220 model, albeit only when only environmental covariates are in the DSM. Their
221 procedure is as follows:

- 222 1. Fit a density surface model.
- 223 2. Re-fit the model with an additional term that characterises the uncer-
224 tainty in the estimation of the detection function (via the derivatives
225 of the probability of detection, \hat{p}).
- 226 3. Variance estimates of the abundance calculated using standard GAM
227 theory will include uncertainty from the estimation of the detection
228 function.

229 A more complete mathematical explanation of this result is given in Appendix
230 B.

231 We consider that propagating the uncertainty in this manner is not only
232 more computationally efficient but also preferable from a technical perspect-
233 ive. A moving block bootstrap does not fully account for spatial autocor-
234 relation. Assuming that the residuals are exchangeable when they are not
235 will lead to wider confidence intervals. The confidence intervals produced
236 via the above method are narrower than their bootstrap equivalents, while
237 maintaining good coverage (results of a small simulation study are given in
238 Appendix C.

239 DSM uncertainty can be visualised via a plot of per-cell coefficient of
240 variation obtained by dividing the standard error for each cell by its pre-
241 dicted abundance. Fig. 4 shows a map of the coefficient of variation for the
242 model which includes both location and depth covariates using the variance
243 propagation method.

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 river or inlet, 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 fitted 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 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.

266 The Tweedie distribution offers a very flexible alternative to the quasi-Poisson
 267 and negative binomial distributions as a response distribution when model-
 268 ling count data (Candy, 2004). Through the parameter λ , many common
 269 distributions arise; varying λ between 1 (Poisson) and 2 (gamma) leads to
 270 a random variable which is a sum of M gamma variables where M is Pois-
 271 son distributed (Jørgensen, 1987). The distribution does not change appre-
 272 ciably when λ is changed by less than 0.1 therefore, a simple line search
 273 over the possible values of λ is usually reasonable. Mark Bravington (pers.
 274 comm.) suggested plotting the square root of the absolute value of the re-
 275 siduals against fitted values; a “flat” plot (points forming a horizontal line)
 276 give an indication of a “good” value for λ . We additionally suggest using the
 277 metrics described in the next section for model selection.

278 Practical advice

279 Fig. 5 shows a flow diagram of the modelling process for creating a density
 280 surface model. The diagram shows which methods are compatible with each
 281 other and what the options are for modelling a particular data set.

282 In our experience, it is sensible obtain a detection function which fits the
 283 data as well as possible and only after a satisfactory detection function has
 284 been obtained, begin spatial modelling. Model selection can be performed
 285 for the detection function using AIC and model checking using goodness-of-
 286 fit tests given in Buckland *et al.* (2004). If animals occur in groups rather
 287 than individually, bias can be incurred due to the higher visibility of larger

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 at zero distance, if there is a large difference then it may be necessary to include size as a covariate in the detection function see (see Buckland *et al.*, 2001, Section 4.8.2.4). The bottom right panel of Fig. 1 shows a such a plot with the regression line overlaid.

For the spatial model, smooth terms can also be selected using (approximate) p -values. A useful technique for covariate selection is to have an additional penalty to each term in the GAM which allows smooth terms to be removed from the model during fitting (this is illustrated in the example analysis Wood, 2006, Section 4.1.6). Smoothness selection is performed by generalized cross validation (GCV) score, UnBiased Risk Estimator (UBRE) or REstricted Maximum Likelihood (REML) score, these scores can be used for model selection (although it should be noted that REML cannot be used to compare models with different fixed effects, see Wood (2011)). Percentage deviance explained is suitable replacement for the usual (adjusted) R^2 . We highly recommend the use of standard GAM diagnostic plots. Wood (2006), Chapter 5, provides 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. 2 and the plot of the depth in Fig. 1, the plot of the smooth of depth alone looks very

313 similar to the raw plot of the depth data. A smooth of an environment-level
314 covariate such as depth can be very useful for assessing the relationships
315 between abundance and the covariate (as in Fig. 2). Caution should be
316 employed when interpreting smooth relationships and abundance estimates,
317 especially if there are gaps over the range of covariate values; large counts
318 may occur at a high value of depth but if no further observations occur at
319 such a high value, then investigators should be skeptical of any relationship.
320 For this reason, a smooth of space is recommended for inclusion in candidate
321 models. Limiting the “wigglyness” of smooths of spatial location (by limiting
322 their basis size) can be a useful way of restricting their influence whilst still
323 allowing them to “mop up” the residual spatial correlation in the data (how
324 this can be achieved is shown in the example analysis).

325 In the analysis presented we have converted from latitude and longitude
326 to kilometres from the centre of the survey region (27.01, -88.3) because the
327 bivariate smoother used (the thin plate spline; Wood, 2003) is isotropic, i.e.
328 it treats the wigglyness of the smoother in each direction as equal. Moving 1
329 degree in latitude is not the same as moving 1 degree in longitude and so using
330 kilometres from the centre of the study region makes the covariates isotropic
331 (using SI units throughout would also remove the need for conversion).

332 Direct modelling of the spatial point process

333 Rather than use a GAM to model the spatially explicit part of the model,
334 two recent articles (Johnson *et al.*, 2010; Niemi & Fernández, 2010) have
335 modelled the process using point processes (Cox & Isham, 1980). In both

336 cases the density of objects described by an intensity function, which can
337 include spatially-referenced covariates.

338 Johnson *et al.* (2010) proposes a point process-based model for distance
339 sampling data. They first assume that the locations of all individuals in the
340 survey area (not just those observed) form a realisation of a Poisson process.
341 Parameters of the intensity function are then estimated via standard max-
342 imum likelihood methods for point processes (Baddeley & Turner, 2000). In
343 contrast to Hedley & Buckland (2004), all parameters are estimated jointly
344 so uncertainty from both the spatial pattern and the detection function is
345 incorporated into variance estimates for the abundance. This also ensures
346 that correlations between the detection function and underlying point process
347 are estimated correctly (and do not falsely inflate or deflate variance estim-
348 ates). The authors also addressed the issue of overdispersion unmodelled by
349 spatial covariates (i.e. counts that do not follow a Poisson mean-variance
350 relationship) using a post-hoc correction factor.

351 Niemi & Fernández (2010) also use Poisson processes but incorporate
352 them into a fully Bayesian approach. Model fitting proceeds in two stages:
353 first the detection function is fitted, then the spatial model (via MCMC)
354 assuming the detection function parameters are known, so detection func-
355 tion uncertainty is not incorporated in the spatial model (an extension that
356 incorporates uncertainty is, however, feasible).

357 Both of the above Poisson process models do not account for group size,
358 but both state that this could be included by considering a marked point
359 process (Cox & Isham, 1980, Section 5.5). Both methods offer direct mod-
360 elling 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 due to how thin transects are compared to their lengths and how small circles are compared to the study area.

Other examples of direct modelling of density are Royle & Dawson (2004) and Royle & Dorazio (2008).

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) model correctness (for model-based inference). It is easier to have more 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 the availability of spatially-references 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 two areas: temporal inferences and the handling of spatial correlation. These should become more mainstream as modern spatiotemporal 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”). These advances should assist both modellers and wildlife managers to make optimal conservation decisions.

Riding on the back of the advances of Royle & Dawson (2004), Niemi & Fernández (2010) and Johnson *et al.* (2010), direct modelling of the process should be possible via use of integrated nested Laplace approximation (INLA; Rue *et al.* (2009)). Such an advance would make computation both fast and allow for a flexible modelling.

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 a very active area of current research, and we look forward to further improvements and extensions in the near future.

408 **Acknowledgments**

409 DLM wishes to thank Mark Bravington and Sharon Hedley for their help
410 and patience in explaining and providing code for their variance propagation
411 method. Funding for the implementation of the recent advances into the
412 mrds package and Distance software came from the US Navy, Chief of Naval
413 Operations (Code N45), grant number N00244-10-1-0057.

414 References

- 415 Baddeley, A. & Turner, R. (2000) Practical maximum pseudolikelihood for spatial
416 point patterns. *Australian & New Zealand Journal of Statistics*, **42**, 283–322.
- 417 Bravington, M. & Hedley, S.L. (2009) Antarctic minke whale abundance estimates
418 from the second and third circumpolar IDCR/SOWER surveys using the
419 SPLINTR model.
420 URL [http://www.iwcoffice.org/_documents/sci_com/sc61docs/](http://www.iwcoffice.org/_documents/sci_com/sc61docs/SC-61-IA14.pdf)
421 [SC-61-IA14.pdf](http://www.iwcoffice.org/_documents/sci_com/sc61docs/SC-61-IA14.pdf)
- 422 Buckland, S.T., Anderson, D., Burnham, K.P., Laake, J.L., Borchers, D.L. &
423 Thomas, L. (2001) *Introduction to Distance Sampling*. Oxford University Press.
- 424 Buckland, S.T., Anderson, D., Burnham, K.P., Laake, J.L., Borchers, D.L. &
425 Thomas, L. (2004) *Advanced Distance Sampling*. Oxford University Press.
- 426 Candy, S. (2004) Modelling catch and effort data using generalised linear models,
427 the Tweedie distribution, random vessel effects and random stratum-by-year
428 effects. *Ccamlr Science*, **11**, 59–80.
429 URL http://www.ccamlr.org/ccamlr_science/Vol-11-2004/04candy.pdf
- 430 Cox, D.R. & Isham, V. (1980) *Point Processes*. Monographs on Applied Probability
431 and Statistics. Chapman and Hall. ISBN 9780412219108.
- 432 Efron, B. & Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Chapman
433 & Hall/CRC. ISBN 9780412042317.
434 URL [http://books.google.com/books?id=gLlpIUxRntoC&dq=an+](http://books.google.com/books?id=gLlpIUxRntoC&dq=an+introduction+to+the+bootstrap&hl=&cd=1&source=gbp_api)
435 [introduction+to+the+bootstrap&hl=&cd=1&source=gbp_api](http://books.google.com/books?id=gLlpIUxRntoC&dq=an+introduction+to+the+bootstrap&hl=&cd=1&source=gbp_api)
- 436 Halpin, P., Read, A., Fujioka, E., Best, B., Donnelly, B., Hazen, L., Kot, C.,
437 Urian, K., LaBrecque, E., Dimatteo, A., Cleary, J., Good, C., Crowder, L. &
438 Hyrenbach, K.D. (2009) OBIS-SEAMAP: The World Data Center for Marine
439 Mammal, Sea Bird, and Sea Turtle Distributions. *Oceanography*, **22**, 104–115.
440 URL http://www.tos.org/oceanography/archive/22-2_halpin.html
- 441 Hardin, J. & Hilbe, J. (2003) *Generalized Estimating Equations*. Chapman and
442 Hall/CRC, London, UK.
- 443 Hedley, S.L. & Buckland, S.T. (2004) Spatial models for line transect sampling.
444 *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 181–199.
- 445 Johnson, D.S., Laake, J.L. & Ver Hoef, J.M. (2010) A model-based approach for
446 making ecological inference from distance sampling data. *Biometrics*, **66**, 310–
447 318.

- 448 Jørgensen, B. (1987) Exponential dispersion models. *Journal of the Royal Statist-*
449 *ical Society. Series B, Statistical Methodology*, **49**, 127–162.
- 450 Marques, T.A., Thomas, L., Fancy, S. & Buckland, S.T. (2007) Improving estimates
451 of bird density using multiple-covariate distance sampling. *The Auk*, **124**, 1229–
452 1243.
- 453 Miller, D.L., Jones, E. & Matthiopoulos, J. (????) Reliable spatial smoothing
454 without edge effects. pp. 1–8.
- 455 Miller, D.L. & Wood, S.N. (????) Finite area smoothing with generalized distance
456 splines. pp. 1–27.
- 457 Niemi, A. & Fernández, C. (2010) Bayesian Spatial Point Process Modeling of Line
458 Transect Data. *Journal of Agricultural, Biological, and Environmental Statistics*,
459 **15**, 327–345.
- 460 Petersen, I.K., MacKenzie, M.L., Rexstad, E.A., Wisz, M.S. & Fox, A.D. (2011)
461 Comparing pre- and post-construction distributions of long-tailed ducks *Clan-*
462 *gula hyemalis* in and around the Nysted offshore wind farm, Denmark: a quasi-
463 designed experiment accounting for imperfect detection, local surface features
464 and autocorrelation. 2011-1.
- 465 Ramsay, T. (2002) Spline smoothing over difficult regions. *Journal of the Royal*
466 *Statistical Society. Series B, Statistical Methodology*, **64**, 307–319.
467 URL [http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00339/](http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00339/full)
468 [full](http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00339/full)
- 469 Royle, J.A. & Dawson, D. (2004) Modeling abundance effects in distance sampling.
470 *Ecology*.
471 URL <http://www.esajournals.org/doi/pdf/10.1890/03-3127>
- 472 Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical Modeling and Inference in*
473 *Ecology*. The Analysis of Data from Populations, Metapopulations and Com-
474 munities. ISBN 9780123740977.
475 URL [http://books.google.com/books?id=rDppWpVP6a0C&printsec=](http://books.google.com/books?id=rDppWpVP6a0C&printsec=frontcover&dq=Hierarchical+modeling+and+inference+in+ecology+inauthor:royle&hl=&cd=1&source=gbs_api)
476 [frontcover&dq=Hierarchical+modeling+and+inference+in+ecology+](http://books.google.com/books?id=rDppWpVP6a0C&printsec=frontcover&dq=Hierarchical+modeling+and+inference+in+ecology+inauthor:royle&hl=&cd=1&source=gbs_api)
477 [inauthor:royle&hl=&cd=1&source=gbs_api](http://books.google.com/books?id=rDppWpVP6a0C&printsec=frontcover&dq=Hierarchical+modeling+and+inference+in+ecology+inauthor:royle&hl=&cd=1&source=gbs_api)
- 478 Rue, H., Martino, S. & Chopin, N. (2009) Approximate Bayesian inference for
479 latent Gaussian models by using integrated nested Laplace approximations. *J.*
480 *R. Statist. Soc. B*, **71**, 319–392.
- 481 Scott-Hayward, L.A.S., MacKenzie, M.L., Donovan, C.R., Walker, C.G. & Ashe, E.
482 (????) Complex Region Spatial Smoother (CReSS). *Journal of Computational*
483 *and Graphical Statistics*.

- 484 Seber, G.A.F. (1982) *The Estimation of Animal Abundance and Related Paramet-*
 485 *ers*. Blackburn Pr. ISBN 9781930665552.
- 486 URL [http://books.google.com/books?id=bnGaPQAACAAJ&dq=seber&cd=](http://books.google.com/books?id=bnGaPQAACAAJ&dq=seber&cd=10&source=gbp_api)
 487 [10&source=gbp_api](http://books.google.com/books?id=bnGaPQAACAAJ&dq=seber&cd=10&source=gbp_api)
- 488 Skaug, H.J. (2006) Markov modulated Poisson processes for clustered line transect
 489 data. *Environmental and Ecological Statistics*, **13**, 199–211.
- 490 Thomas, L., Buckland, S.T., Rexstad, E.A., Laake, J.L., Strindberg, S., Hedley,
 491 S.L., Bishop, J.R., Marques, T.A. & Burnham, K.P. (2010) Distance software:
 492 design and analysis of distance sampling surveys for estimating population size.
 493 *Journal of Applied Ecology*, **47**, 5–14.
- 494 Wang, H. & Ranalli, M. (2007) Low-rank smoothing splines on complicated do-
 495 mains. *Biometrics*, **63**, 209–217.
- 496 Williams, R., Hedley, S.L., Branch, T.A., Bravington, M.V., Zerbini, A.N. & Find-
 497 lay, K.P. (2011) Chilean blue whales as a case study to illustrate methods to
 498 estimate abundance and evaluate conservation status of rare species. *Conserva-*
 499 *tion Biology*, **25**, 526–535.
- 500 Williams, R., Hedley, S.L. & Hammond, P. (2006) Modeling distribution and
 501 abundance of Antarctic baleen whales using ships of opportunity. *Ecology and*
 502 *Society*, **11**, 1.
- 503 Wood, S.N. (2003) Thin plate regression splines. *Journal of the Royal Statistical*
 504 *Society. Series B, Statistical Methodology*, **65**, 95–114.
- 505 Wood, S.N. (2006) *Generalized Additive Models: An introduction with R*. Chapman
 506 & Hall/CRC.
- 507 Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal
 508 likelihood estimation of semiparametric generalized linear models. *Journal of*
 509 *the Royal Statistical Society. Series B, Statistical Methodology*, **73**, 3–36.
 510 URL [http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2010.](http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2010.00749.x/full)
 511 [00749.x/full](http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2010.00749.x/full)
- 512 Wood, S.N., Bravington, M.V. & Hedley, S.L. (2008) Soap film smoothing. *Journal*
 513 *of the Royal Statistical Society. Series B, Statistical Methodology*, **70**, 931–955.

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.

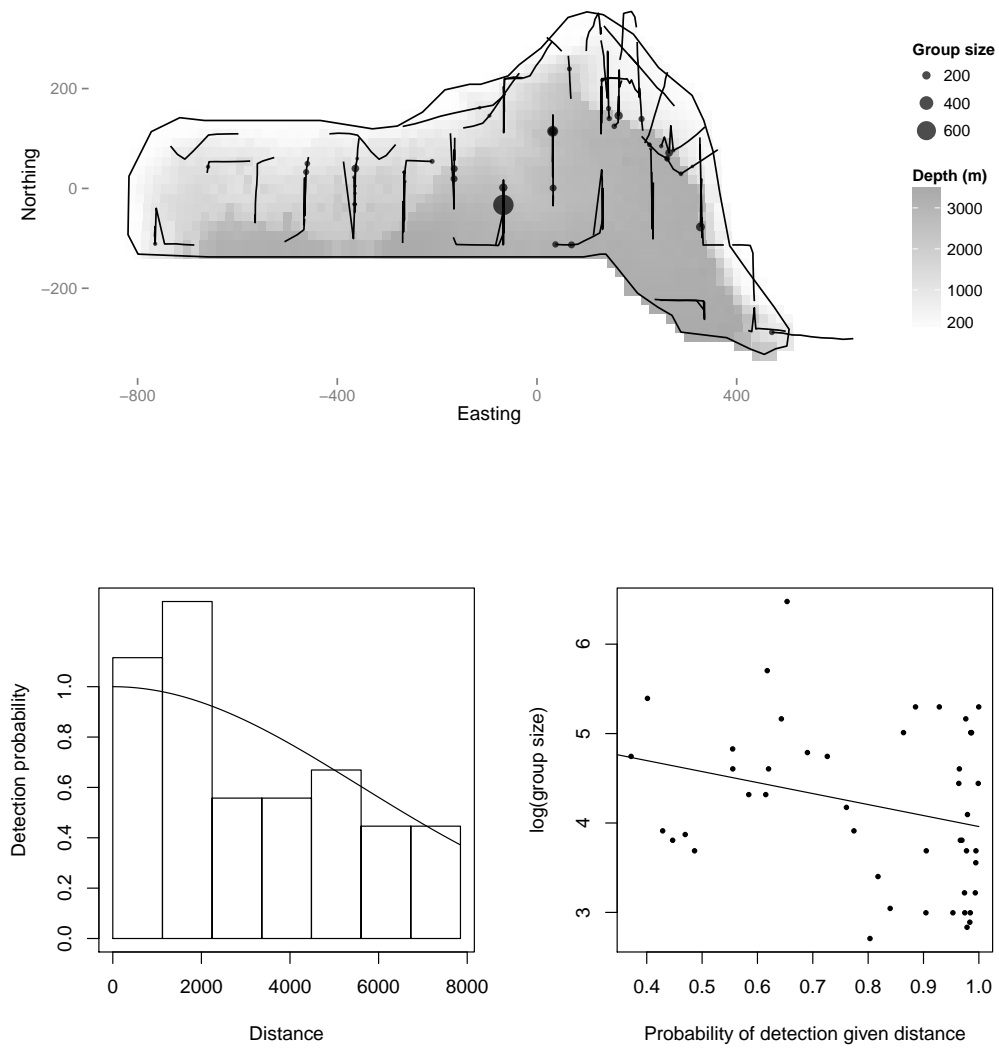


Fig. 2 Predictions for the dolphin data. Top: Predictions from the model using only depth as an explanatory variable, bottom: the model using both depth and location.

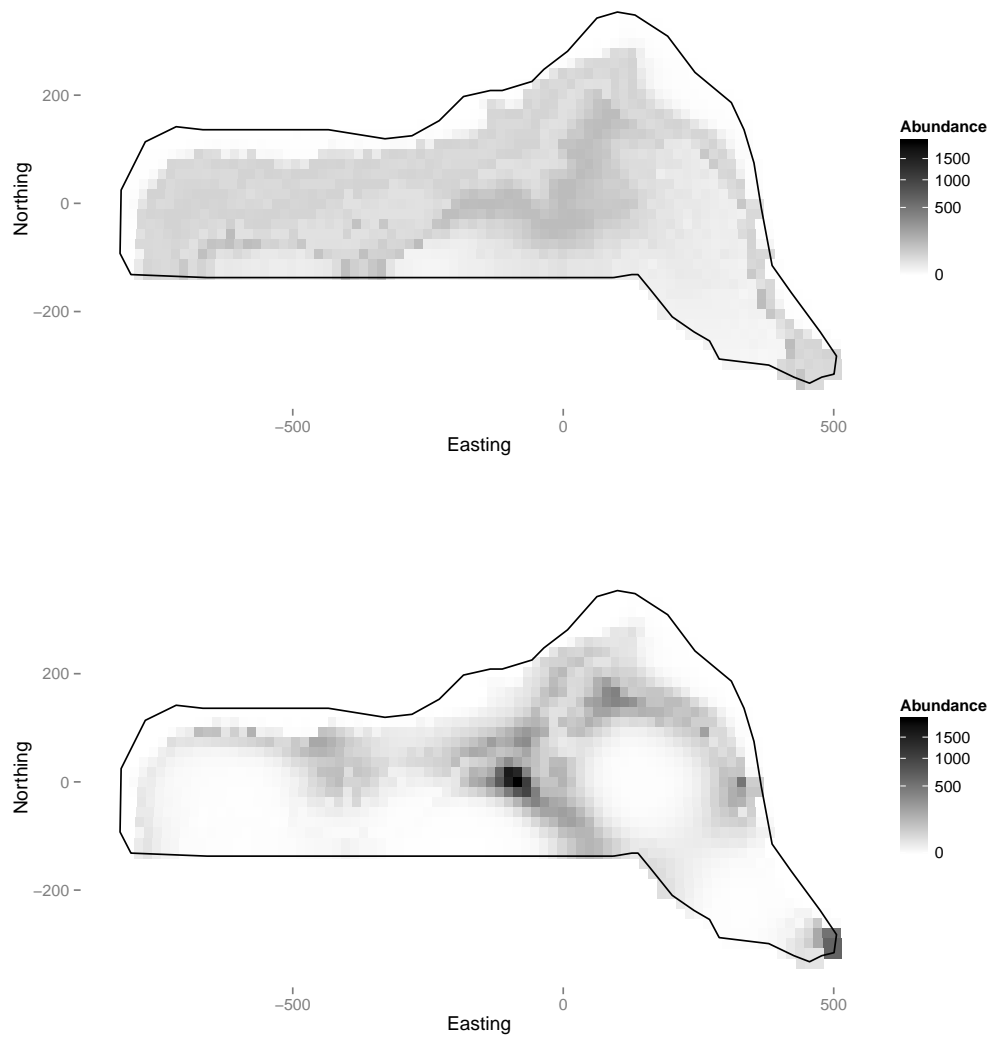


Fig. 3 Plot of the effect on the response of depth (from the model with both depth and location smooths), note that it is possible to draw a straight line between 750m and 3000m within the confidence band (between the dashed lines), so the wiggles in the smooth may not be indicative of any relationship. 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.

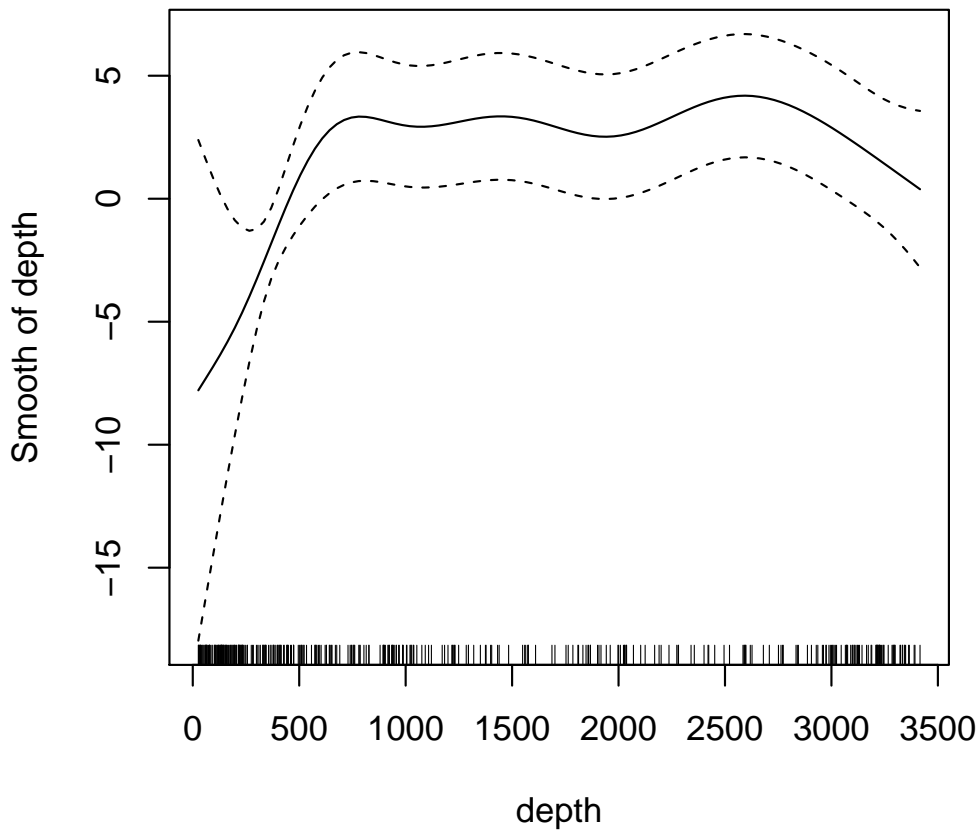


Fig. 4 Plot of coefficient of variation map for the model with smooths of both depth and location. Uncertainty was estimated using the variance propagation method of Williams *et al.* (2011). As might be expected, there is high uncertainty where there is low sampling effort (comparing to Fig. 1).

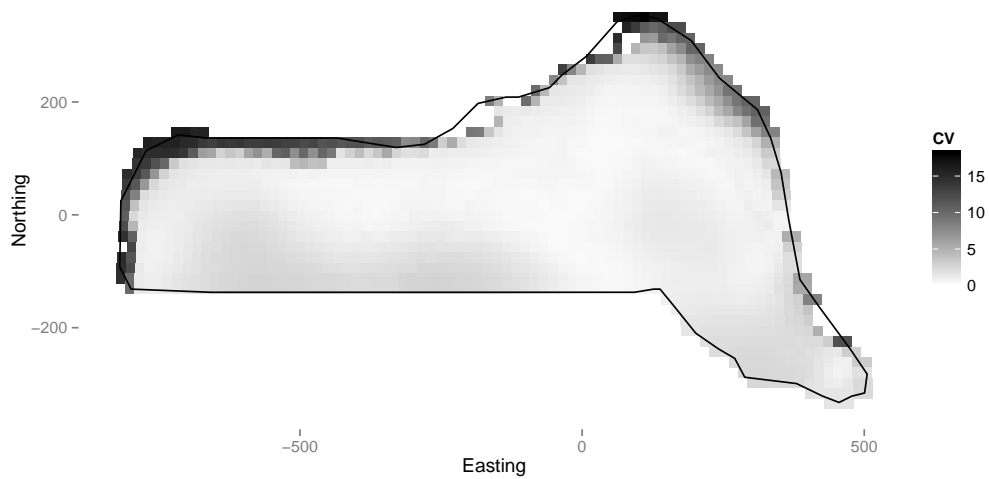


Fig. 5 Flow diagram showing the modelling process for creating a density surface model.

