

1 **Running title:** Spatial models for distance sampling  
2 **Number of words:** ~4309  
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<sup>1\*</sup>, M. Louise Burt<sup>2</sup>,**  
9 **Eric A. Rexstad<sup>2</sup>, Len Thomas<sup>2</sup>.**

- 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](mailto: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).

52 Distance sampling takes plot sampling (counting all the individuals or  
53 groups of objects within a strip or circle) and extends it to the case where  
54 detection is not certain. Observers move along lines or stand at points and  
55 record the distance from the line or point to the object of interest ( $y$ ). These  
56 distances are used to estimate the *detection function*,  $g(y)$  (bottom left panel,  
57 Fig. 1), by modelling the decrease in detectability with increasing distance  
58 from the line or point (conventional distance sampling, CDS). The detection  
59 function may also include covariates (multiple covariate distance sampling,  
60 MCDS; Marques *et al.*, 2007) which affect the scale of the detection function.  
61 From the fitted detection function, the probability of detection can be cal-

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

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

64 where  $A$  is the area of the study region,  $a$  is the area covered by the survey  
 65 (i.e., the sum of the areas of all of the strips/circles) and the summation  
 66 takes place over the  $n$  observed individuals (Buckland *et al.*, 2001, Chapter  
 67 3). In general distance sampling is more efficient than plot sampling because  
 68 a much higher proportion of observations can be used in the analysis. Often  
 69 up to half the observations in a plot sampling data set are discarded in order  
 70 to ensure the assumption of certain detection is met. In contrast, distance  
 71 sampling uses the observations that would have been discarded to model the  
 72 detection (typically data are discarded beyond a given *truncation distance*  
 73 during analysis).

74 When fitting the detection function in a distance sampling analysis, one  
 75 assumes that the objects of interest are distributed according to some process  
 76 (Buckland *et al.*, 2001, Section 2.1). It is usually possible to design surveys  
 77 such that a homogenous process can be assumed so that (with respect to  
 78 the line) objects are distributed uniformly. This can be achieved by ensuring  
 79 that transects randomly located.

80 Estimators such as eqn (1) rely on the design of the study to ensure that  
 81 abundance estimates over the whole study area (scaling up from the covered  
 82 region) are valid. This article focusses on *model-based* inference to extrapolate  
 83 to a larger study area. Specifically, we consider the use spatially explicit

84 models to investigate the response of biological populations to biotic and abi-  
85 otic covariates that vary over the survey area. A spatially-explicit model can  
86 explain the between transect variation (which is often a large component of  
87 the variance in design-based estimates) and so using a model-based approach  
88 can lead to lower variances. Model-based inference also enables the use of  
89 data from opportunistic surveys, for example, incidental data arising from  
90 “ecotourism” cruises (Williams *et al.*, 2006).

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

101 This article reviews the current landscape of spatial modelling of distance  
102 sampling data, illustrating some recent developments most useful to applied  
103 ecologists. The methods discussed have available in the popular Windows  
104 application Distance (Thomas *et al.*, 2010) for some time but the recent  
105 advances covered here have been implemented in a new R package, `dsm` (which  
106 is available on CRAN and will be incorporated in to Distance).

107 Throughout this article a motivating data set is used to illustrate the  
108 methods. These data are from a combination of several shipboard surveys

109 conducted on pantropical spotted dolphins (*Stenella attenuata*) in the Gulf  
110 of Mexico. 47 groups of dolphins were observed; group size was recorded, as  
111 well as the Beaufort sea state at the time of the observation. Coordinates for  
112 each observation and bathymetry data were available as covariates for the  
113 analysis. A complete example analysis is provided as an online appendix.  
114 The data used in the analysis are available in the `dsm` package and Distance.

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

## 121 Density surface modelling

122 This section focuses on modelling the density/abundance estimation stage  
123 of distance sampling, using the “count model” of Hedley & Buckland (2004),  
124 which we refer to as *density surface modelling* (DSM). Both line and point  
125 transects can be used but if lines are used then they are split into con-  
126 tiguous *segments* (indexed by  $j$ ), which are of length  $l_j$ . Segments should  
127 be small enough such that the density does not vary appreciably within a  
128 segment (usually making the segments approximately square,  $2w \times 2w$ , is  
129 sufficient). Count or estimated abundance is then modelled as a smooth  
130 function of covariates using a generalized additive model (GAM; e.g. Wood,  
131 2006). For each segment or point, the response is modelled as a function of

132 environmental covariates that are measured at the segment/point level (the  
 133  $z_{jk}$  with  $k$  indexing the covariates, e.g., location, sea surface temperature,  
 134 weather conditions). The covered area enters the model as an offset: the  
 135 area covered at segment  $j$  is  $A_j = 2wl_j$  and at point  $j$  is  $A_j = w\pi^2$  (where  $w$   
 136 is the truncation distance).

137 COUNT AS RESPONSE

138 The model for the count per segment is:

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

139 where the  $f_k$ s are smooth functions of the covariates and  $\beta_0$  is an intercept  
 140 term. Multiplying the covered area ( $A_j$ ) by the probability of detection ( $\hat{p}_j$ )  
 141 gives the *effective area* for segment  $j$ . If there are no covariates other than  
 142 distance in the detection function then the probability of detection is constant  
 143 (i.e.,  $\hat{p}_j = \hat{p}$ ,  $\forall j$ ). The distribution of  $n_j$  can be modelled as overdispersed  
 144 Poisson, negative binomial, or Tweedie distribution (see *Recent developments*,  
 145 below).

146 Fig. 1 (top panel) shows the raw observations of the dolphin data, along  
 147 with the transect lines, overlaid on the depth data. Fig. 2 shows a DSM  
 148 fitted to the dolphin data, the top panel shows predictions from a model  
 149 where depth was the only covariate, the bottom panel shows predictions  
 150 where a (bivariate) smooth of spatial location was also included.

151 As well as simply calculating abundance estimates, relationships between  
 152 covariates and abundance can be illustrated via plots of marginal smooths.

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

#### 154 ESTIMATED ABUNDANCE AS RESPONSE

155 An alternative to modelling counts would be to use the per-segment/circle  
 156 abundance can be estimated using distance sampling methods and the es-  
 157 timated counts used as the response. In this case we replace  $n_j$  by:

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

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

160 The following model is then fitted:

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

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

#### 164 *DSM with covariates at the observation level*

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



171 the per-segment abundance:

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

172 By not including an offset, but instead dividing the count (or estimated  
173 abundance) by the area of the segment, we can also model density rather  
174 than abundance. We concentrate on abundance here, see Hedley & Buckland  
175 (2004) for further details on modelling density.

## 176 PREDICTION

177 To calculate an abundance estimate for a region of interest, the environ-  
178 mental covariates included in the model must be available at each prediction  
179 point at the required resolution (using prediction grid cells that are smal-  
180 ler than the resolution of the spatially referenced data have no effect on  
181 abundance/density estimates). The areas of the prediction cells must also be  
182 included in the prediction data. Predictions are made for the each grid cell  
183 using covariate values associated with each cell and abundance estimates are  
184 produced for a given region by summing predicted values over corresponding  
185 grid cells.

## 186 VARIANCE ESTIMATION

187 Estimating the variance of abundances calculated using a DSM is not straight-  
188 forward: uncertainty from the estimated parameters of the detection function  
189 must be incorporated into the spatial model. A second consideration is that  
190 in a line transect survey, adjacent segments are likely to be correlated; failure

191 to account for this spatial autocorrelation will lead to artificially low variance  
192 estimates and hence misleadingly narrow confidence intervals.

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

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

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

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

210 The above procedure assumes that there is no correlation in space between  
211 segments however, if many animals are observed in a particular segment then  
212 we might expect there to be high numbers in the adjacent segments. A  
213 moving block bootstrap (MBB; Efron & Tibshirani, 1993, Section 8.6) can

214 account for some of this spatial autocorrelation in the variance estimation.  
215 The segments are grouped together into overlapping blocks, (so if the block  
216 size is 5, block one is segments 1,  $\dots$ , 5, block two is segments 2,  $\dots$ , 6, and so  
217 on). Then, at step (2) above, resamples are taken of the blocks (contiguous  
218 collections of segments) rather than individual segments within the transects.  
219 Using blocks should account for some of the autocorrelation between the  
220 segments, inflating the variances accordingly. However, because the block size  
221 dictates the maximum amount of spatial autocorrelation accounted for, this  
222 may not fully account for the autocorrelation. These bootstrap procedures  
223 can also be modified to take into account detection function uncertainty by  
224 generating new distances from the fitted detection function and then re-  
225 calculating the offset by fitting a detection function to the new distances.

226 DSM uncertainty can be visualised via a plot of per-cell coefficient of  
227 variation obtained by dividing the standard error for each cell by its predicted  
228 abundance.

## 229 Recent developments

### 230 *GAM uncertainty and variance propagation*

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

Williams *et al.* (2011) go a step further and incorporate the uncertainty in the estimation of the detection function into the variance of the spatial model, albeit only when only environmental covariates are in the DSM. Their procedure is as follows:

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

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

We consider that propagating the uncertainty in this manner is not only more computationally efficient but also preferable from a technical perspective. A moving block 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 confidence intervals produced via the above method are narrower than their bootstrap equivalents, while maintaining good coverage (results of a small simulation study are given in Appendix C).

Fig. 4 shows a map of the coefficient of variation for the model which includes both location and depth covariates. Variance has been calculated using the variance propagation method.

262 Recent work (Ramsay, 2002; Wang & Ranalli, 2007; Wood *et al.*, 2008; Scott-  
263 Hayward *et al.*, 2013; Miller & Wood) has highlighted the need to take care  
264 when smoothing over areas with complicated boundaries, e.g., those with  
265 rivers, peninsulae or islands. If two parts of the domain (either side of a  
266 river or inlet, say) are inappropriately linked by the model (the distance  
267 between the points is measured as a straight line, rather taking into account  
268 obstacles) then the boundary feature can be “smoothed across” leading to  
269 incorrect inference. Ensuring that a realistic spatial model has been fitted  
270 to the data is essential for valid inference. The soap film smoother of Wood  
271 *et al.* (2008) is particularly appealing as the model jointly estimates boundary  
272 conditions for a complex study area along with the interior smooth. This  
273 can be particularly helpful when uncertainty is estimated via a bootstrap as  
274 the model helps avoid large, unrealistic predictions which can plague other  
275 smoothers (Bravington & Hedley, 2009).

276 Even if the study area does not have a complicated boundary, edge effects  
277 can still be problematic. Miller *et al.* show that global smoothers which have  
278 unpenalized plane components tend to cause the fitted surface to increase  
279 unrealistically as predictions move further away from the locations of survey  
280 effort. They suggest the use of Duchon splines (a generalisation of thin plate  
281 regression splines) to alleviate the problem.

283 The Tweedie distribution offers a very flexible alternative to the quasi-Poisson  
 284 and negative binomial distributions as a response distribution when model-  
 285 ling count data (Candy, 2004). Through the parameter  $\lambda$ , many common  
 286 distributions arise; varying  $\lambda$  between 1 (Poisson) and 2 (gamma) leads to  
 287 a random variable which is a sum of  $M$  gamma variables where  $M$  is Pois-  
 288 son distributed (Jørgensen, 1987). The distribution does not change appre-  
 289 ciably when  $\lambda$  is changed by less than 0.1 therefore, a simple line search  
 290 over the possible values of  $\lambda$  is usually reasonable. Mark Bravington (pers.  
 291 comm.) suggested plotting the square root of the absolute value of the re-  
 292 siduals against fitted values; a “flat” plot (points forming a horizontal line)  
 293 give an indication of a “good” value for  $\lambda$ . We additionally suggest using the  
 294 metrics described in the next section for model selection.

## 295 Practical advice

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

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

305 groups. Bias due to group size can be assessed by regressing evaluations of  
306 the fitted detection function onto the logarithm of group size, then comparing  
307 the expected and observed values of the group size at zero distance, if there  
308 is a large difference then it may be necessary to include size as a covariate in  
309 the detection function see (see Buckland *et al.*, 2001, Section 4.8.2.4). The  
310 bottom right panel of Fig. 1 shows a such a plot with the regression line  
311 overlaid.

312 For the spatial model, smooth terms can also be selected using (approx-  
313 imate)  $p$ -values. A useful technique for covariate selection is to have an  
314 additional penalty to each term in the GAM which allows smooth terms to  
315 be removed from the model during fitting (this is illustrated in the example  
316 analysis Wood, 2006, Section 4.1.6). Smoothness selection is performed by  
317 generalized cross validation (GCV) score, UnBiased Risk Estimator (UBRE)  
318 or REstricted Maximum Likelihood (REML) score, these scores can be used  
319 for model selection (although it should be noted that REML cannot be used  
320 to compare models with different fixed effects, see Wood (2011)). Percent-  
321 age deviance explained is suitable replacement for the usual (adjusted)  $R^2$ .  
322 We highly recommend the use of standard GAM diagnostic plots. Wood  
323 (2006), Chapter 5, provides practical information on GAM model selection  
324 and fitting.

325 In the analysis of the dolphin data, we included a smooth of location.  
326 This not only nearly doubles the percentage deviance explained (27.3% to  
327 52.7%), it also allows us to account for spatial autocorrelation (in a primitive  
328 way). One can see this when comparing the two plots in Fig. 2 and the  
329 plot of the depth in Fig. 1, the plot of the smooth of depth alone looks very

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

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

## 349 **Direct modelling of the spatial point process**

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



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

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

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

374 Both of the above Poisson process models do not account for group size,  
375 but both state that this could be included by considering a marked point  
376 process (Cox & Isham, 1980, Section 5.5). Both methods offer direct mod-  
377 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.

A final example of direct modelling of density is given in Royle *et al.* (2004). In their article, the authors formulate an unconditional likelihood per-point/line, which is a function of the unobserved point/transect abundances. These unobserved abundances are treated as (Poisson or negative binomial) random effects, which are then integrated out to give a per-point/line likelihood which is a function only of detection function parameters and parameters of the random effects (linear functions of the environmental covariates). Due to the multinomial nature of the per-point/line likelihood proposed distance data must be binned, resulting in a loss of information. Although an arbitrarily large number of bins could be used as an approximation, this is both inelegant and computationally intensive.

## 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.

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

408       The field is quickly evolving to allow modelling of more complex data  
409       building on the basic ideas of density surface modelling. We expect to see  
410       large advances in two areas: temporal inferences and the handling of spa-  
411       tial correlation. These should become more mainstream as modern spatio-  
412       temporal modelling techniques are adopted. Petersen *et al.* (2011) provided  
413       a very basic framework for temporal modelling; their model included smooth  
414       terms both before and after the construction of an offshore windfarm. Spa-  
415       tial autocorrelation can be accounted for via approaches that explicitly intro-  
416       duce correlations such as generalized estimating equations (GEEs; Hardin &  
417       Hilbe, 2003) or via mechanisms such as that of Skaug (2006), which allowed  
418       observations to cluster according to one of several states (e.g. “feeding” or  
419       “transit”) taking into account short-term agglomerations (“hot spots”). These  
420       advances should assist both modellers and wildlife managers to make optimal  
421       conservation decisions.

422       Riding on the back of the advances of Royle *et al.* (2004), Niemi & Fernán-  
423       dez (2010) and Johnson *et al.* (2010), direct modelling of the process should  
424       be possible via use of integrated nested Laplace approximation (INLA; Rue  
425       *et al.* (2009)). Such an advance would make computation both fast and allow

426 for a flexible modelling.

427     Density surface modelling allows wildlife managers to make best use of  
428 the available spatial data to understand patterns of abundance, and hence  
429 make better conservation decisions (e.g., about reserve placement). The re-  
430 cent advances mentioned here increase the reliability of the outputs from a  
431 modelling exercise, and hence the efficacy of these decisions. Density surface  
432 modelling from survey data is a very active area of current research, and we  
433 look forward to further improvements and extensions in the near future.

## 434 **Acknowledgments**

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

## 440 References

- 441 Baddeley, A. & Turner, R. (2000) Practical maximum pseudolikelihood for spatial  
442 point patterns. *Australian & New Zealand Journal of Statistics*, **42**, 283–322.
- 443 Bravington, M. & Hedley, S.L. (2009) Antarctic minke whale abundance estimates  
444 from the second and third circumpolar IDCR/SOWER surveys using the  
445 SPLINTR model.  
446 URL [http://www.iwcoffice.org/\\_documents/sci\\_com/sc61docs/](http://www.iwcoffice.org/_documents/sci_com/sc61docs/SC-61-IA14.pdf)  
447 [SC-61-IA14.pdf](http://www.iwcoffice.org/_documents/sci_com/sc61docs/SC-61-IA14.pdf)
- 448 Buckland, S.T., Anderson, D., Burnham, K.P., Laake, J.L., Borchers, D.L. &  
449 Thomas, L. (2001) *Introduction to Distance Sampling*. Oxford University Press.
- 450 Buckland, S.T., Anderson, D., Burnham, K.P., Laake, J.L., Borchers, D.L. &  
451 Thomas, L. (2004) *Advanced Distance Sampling*. Oxford University Press.
- 452 Candy, S. (2004) Modelling catch and effort data using generalised linear models,  
453 the Tweedie distribution, random vessel effects and random stratum-by-year  
454 effects. *Ccamlr Science*, **11**, 59–80.  
455 URL [http://www.ccamlr.org/ccamlr\\_science/Vol-11-2004/04candy.pdf](http://www.ccamlr.org/ccamlr_science/Vol-11-2004/04candy.pdf)
- 456 Cox, D.R. & Isham, V. (1980) *Point Processes*. Monographs on Applied Probability  
457 and Statistics. Chapman and Hall. ISBN 9780412219108.  
458 URL <http://www.worldcat.org/title/point-processes/oclc/7088477>
- 459 Efron, B. & Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Chapman  
460 & Hall/CRC. ISBN 9780412042317.  
461 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=gbs_api)  
462 [introduction+to+the+bootstrap&hl=&cd=1&source=gbs\\_api](http://books.google.com/books?id=gLlpIUxRntoC&dq=an+introduction+to+the+bootstrap&hl=&cd=1&source=gbs_api)
- 463 Halpin, P., Read, A., Fujioka, E., Best, B., Donnelly, B., Hazen, L., Kot, C.,  
464 Urian, K., LaBrecque, E., Dimatteo, A., Cleary, J., Good, C., Crowder, L. &  
465 Hyrenbach, K.D. (2009) OBIS-SEAMAP: The World Data Center for Marine  
466 Mammal, Sea Bird, and Sea Turtle Distributions. *Oceanography*, **22**, 104–115.  
467 URL [http://www.tos.org/oceanography/archive/22-2\\_halpin.html](http://www.tos.org/oceanography/archive/22-2_halpin.html)
- 468 Hardin, J. & Hilbe, J. (2003) *Generalized Estimating Equations*. Chapman and  
469 Hall/CRC, London, UK.
- 470 Hedley, S.L. & Buckland, S.T. (2004) Spatial models for line transect sampling.  
471 *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 181–199.
- 472 Johnson, D.S., Laake, J.L. & Ver Hoef, J.M. (2010) A model-based approach for  
473 making ecological inference from distance sampling data. *Biometrics*, **66**, 310–  
474 318.

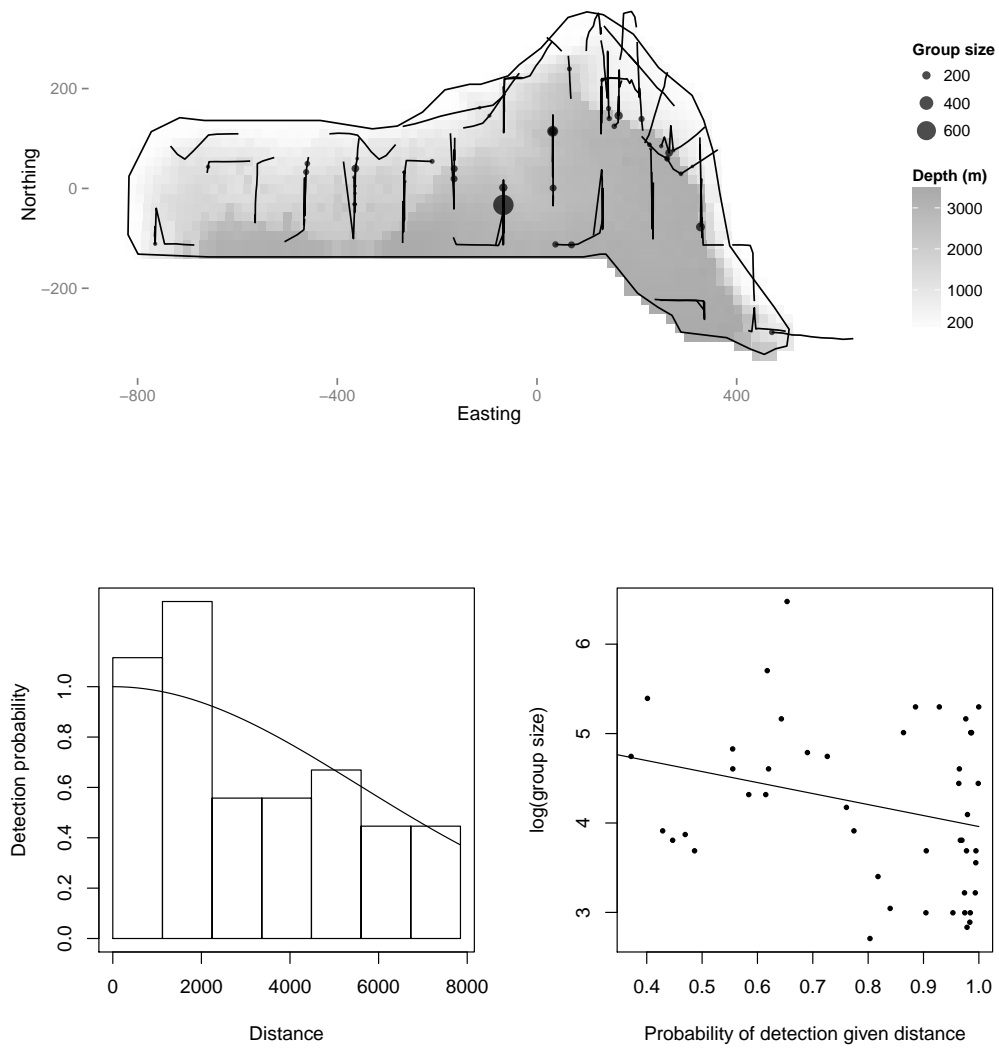
- 475 Jørgensen, B. (1987) Exponential dispersion models. *Journal of the Royal Statist-*  
476 *ical Society. Series B, Statistical Methodology*, **49**, 127–162.
- 477 Marques, T.A., Thomas, L., Fancy, S. & Buckland, S.T. (2007) Improving estimates  
478 of bird density using multiple-covariate distance sampling. *The Auk*, **124**, 1229–  
479 1243.
- 480 Miller, D.L., Jones, E. & Matthiopoulos, J. (????) Reliable spatial smoothing  
481 without edge effects. pp. 1–8.
- 482 Miller, D.L. & Wood, S.N. (????) Finite area smoothing with generalized distance  
483 splines. pp. 1–27.
- 484 Niemi, A. & Fernández, C. (2010) Bayesian Spatial Point Process Modeling of Line  
485 Transect Data. *Journal of Agricultural, Biological, and Environmental Statistics*,  
486 **15**, 327–345.
- 487 Petersen, I.K., MacKenzie, M.L., Rexstad, E.A., Wisz, M.S. & Fox, A.D. (2011)  
488 Comparing pre- and post-construction distributions of long-tailed ducks *Clan-*  
489 *gula hyemalis* in and around the Nysted offshore wind farm, Denmark: a quasi-  
490 designed experiment accounting for imperfect detection, local surface features  
491 and autocorrelation. 2011-1.
- 492 Ramsay, T. (2002) Spline smoothing over difficult regions. *Journal of the Royal*  
493 *Statistical Society. Series B, Statistical Methodology*, **64**, 307–319.  
494 URL [http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00339/](http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00339/full)  
495 [full](http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00339/full)
- 496 Royle, J., Dawson, D. & Bates, S. (2004) Modeling abundance effects in distance  
497 sampling. *Ecology*, **85**, 1591–1597.  
498 URL <http://www.esajournals.org/doi/pdf/10.1890/03-3127>
- 499 Rue, H., Martino, S. & Chopin, N. (2009) Approximate Bayesian inference for  
500 latent Gaussian models by using integrated nested Laplace approximations. *J.*  
501 *R. Statist. Soc. B*, **71**, 319–392.
- 502 Scott-Hayward, L.A.S., MacKenzie, M.L., Donovan, C.R., Walker, C.G. & Ashe, E.  
503 (2013) Complex Region Spatial Smoother (CReSS). *Journal of Computational*  
504 *and Graphical Statistics*.
- 505 Seber, G.A.F. (1982) *The Estimation of Animal Abundance and Related Paramet-*  
506 *ers*. Blackburn Pr. ISBN 9781930665552.  
507 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)  
508 [10&source=gbp\\_api](http://books.google.com/books?id=bnGaPQAACAAJ&dq=seber&cd=10&source=gbp_api)
- 509 Skaug, H.J. (2006) Markov modulated Poisson processes for clustered line transect  
510 data. *Environmental and Ecological Statistics*, **13**, 199–211.

- 511 Thomas, L., Buckland, S.T., Rexstad, E.A., Laake, J.L., Strindberg, S., Hedley,  
512 S.L., Bishop, J.R., Marques, T.A. & Burnham, K.P. (2010) Distance software:  
513 design and analysis of distance sampling surveys for estimating population size.  
514 *Journal of Applied Ecology*, **47**, 5–14.
- 515 Wang, H. & Ranalli, M. (2007) Low-rank smoothing splines on complicated do-  
516 mains. *Biometrics*, **63**, 209–217.
- 517 Williams, R., Hedley, S.L., Branch, T.A., Bravington, M.V., Zerbini, A.N. & Find-  
518 lay, K.P. (2011) Chilean blue whales as a case study to illustrate methods to  
519 estimate abundance and evaluate conservation status of rare species. *Conserva-  
520 tion Biology*, **25**, 526–535.
- 521 Williams, R., Hedley, S.L. & Hammond, P. (2006) Modeling distribution and  
522 abundance of Antarctic baleen whales using ships of opportunity. *Ecology and  
523 Society*, **11**, 1.
- 524 Wood, S.N. (2003) Thin plate regression splines. *Journal of the Royal Statistical  
525 Society. Series B, Statistical Methodology*, **65**, 95–114.
- 526 Wood, S.N. (2006) *Generalized Additive Models: An introduction with R*. Chapman  
527 & Hall/CRC.
- 528 Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal  
529 likelihood estimation of semiparametric generalized linear models. *Journal of  
530 the Royal Statistical Society. Series B, Statistical Methodology*, **73**, 3–36.  
531 URL [http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2010.  
532 00749.x/full](http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2010.00749.x/full)
- 533 Wood, S.N., Bravington, M.V. & Hedley, S.L. (2008) Soap film smoothing. *Journal  
534 of the Royal Statistical Society. Series B, Statistical Methodology*, **70**, 931–955.

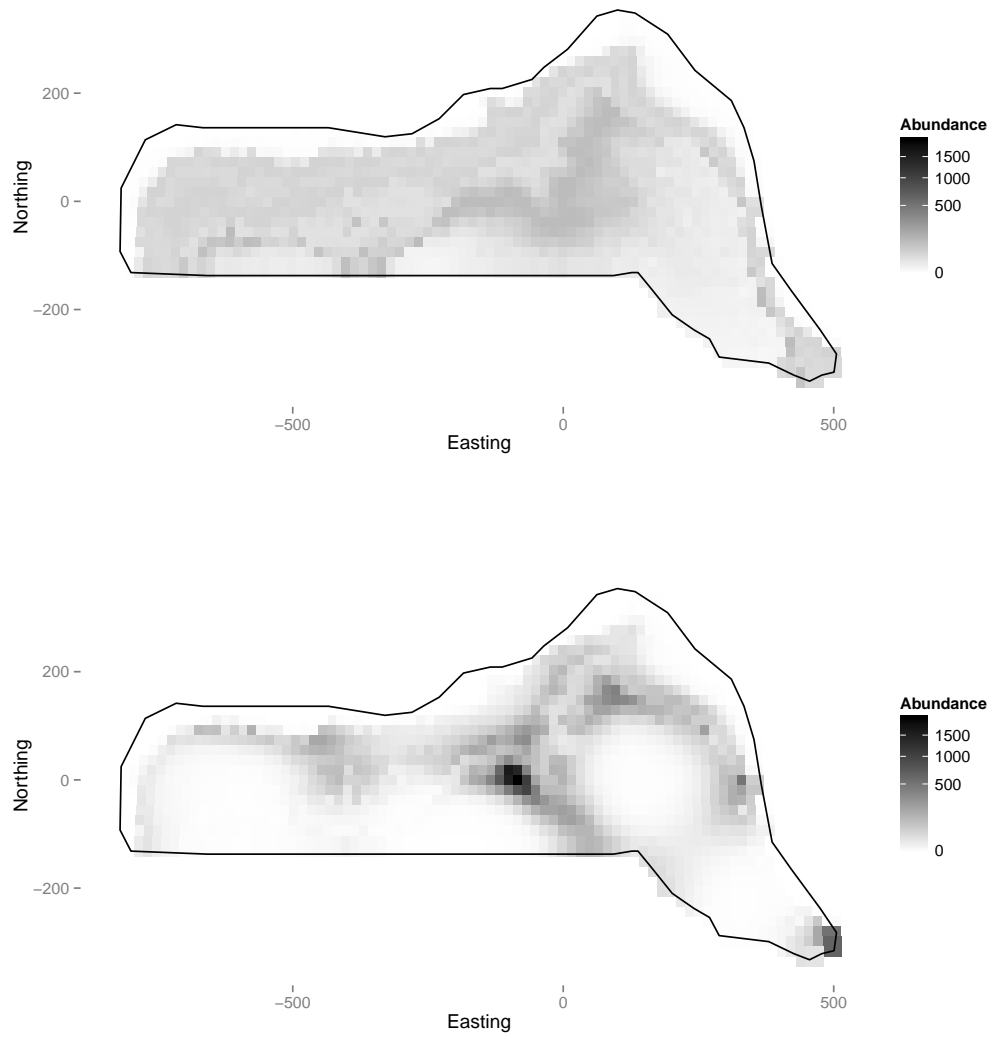




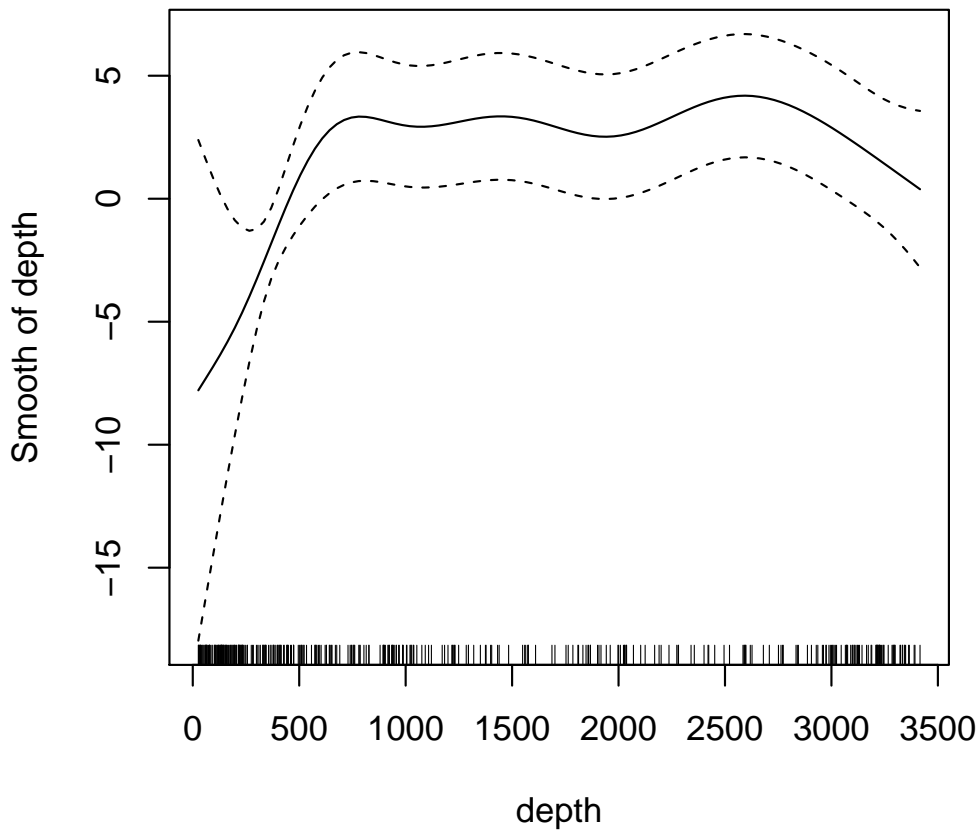
**Fig. 1** Top: the survey area, transect centrelines and observations with size of circle corresponding to the group size overlaid onto depth data; bottom left, histogram of observed distances with fitted detection function; bottom right, plot of evaluations of the fitted detection function at given distances versus the logarithm of group size with linear trend showing the relation between probability of detection (given distance) and group size.



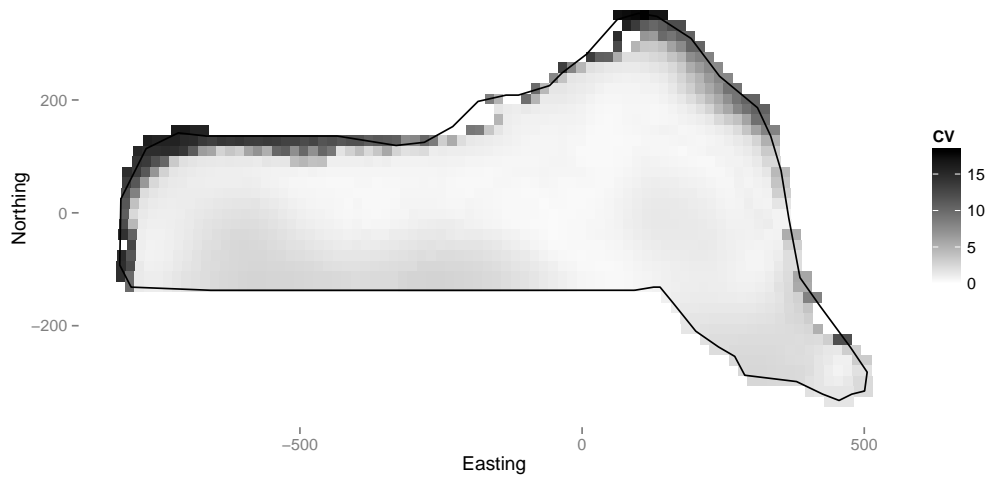
**Fig. 2** Predictions for the dolphin data. Top: Predictions from the model using only depth as an explanatory variable, bottom: the model using both depth and location.



**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.

