

1 **Running title:** Spatial models for distance sampling
2 **Number of words:** ~4925
3 **Number of tables:** 0
4 **Number of figures:** 6
5 **Number of references:** 46

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. Such models can be used to investigate the relationships between distribution and environmental covariates as well as reliably estimate abundances and create maps of animal/plant distribution.
2. Density surface models consist of a spatial model of the abundance of a biological population which has been corrected for uncertain detection via distance sampling methods.
3. We review recent developments in the field and consider the likely directions of future research before focussing on a popular approach based on generalized additive models. In particular we consider spatial modelling techniques that may be advantageous to applied ecologists such as quantification of uncertainty in a two-stage model and smoothing in areas with complex boundaries.
4. The methods discussed are available in an R package developed by the authors (`dsm`) and are largely implemented in the popular Windows package Distance.

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

39 Introduction

40 When surveying biological populations it is increasingly common to record
41 spatially referenced data, for example: coordinates of observations, habitat
42 type, elevation or (if at sea) bathymetry. Spatial models allow for vast data-
43 bases of spatially-referenced data (e.g. OBIS-SEAMAP, Halpin *et al.*, 2009)
44 to be harnessed, enabling investigation of interactions between environmental
45 covariates and population densities. Mapping the spatial distribution of a
46 population can be extremely useful, especially when communicating results
47 to non-experts. Recent advances in both methodology and software have
48 made spatial modelling readily available to the non-specialist (e.g., Wood,
49 2006; Rue *et al.*, 2009). Here we use the term “spatial model” to refer to
50 any model that includes any spatially referenced covariates, not only those
51 models that include location as a covariate. This article is concerned with
52 combining spatial modelling techniques with distance sampling (Buckland
53 *et al.*, 2001, 2004).

54 Distance sampling extends plot sampling to the case where detection
55 is not certain. Observers move along lines or visit points and record the
56 distance from the line or point to the object of interest (y). These distances
57 are used to estimate the *detection function*, $g(y)$ (for example, Fig. 1), by
58 modelling the decrease in detectability with increasing distance from the
59 line or point (conventional distance sampling, CDS). The detection function
60 may also include covariates (multiple covariate distance sampling, MCDS;
61 Marques *et al.*, 2007) which affect the scale of the detection function. From
62 the fitted detection function, the average probability of detection can be

63 estimated by integrating out distance. The estimated average probability
 64 that an animal is detected given that it is in the area covered by the survey,
 65 \hat{p}_i , can then be used to estimate abundance as

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

66 where A is the area of the study region, a is the area covered by the survey
 67 (i.e., the sum of the areas of all of the strips/circles) and the summation
 68 takes place over the n observed clusters, each of size s_i (if individuals are
 69 observed, $s_i = 1 \forall i$) (Buckland *et al.*, 2001, Chapter 3). Often up to half
 70 the observations in a plot sampling data set are discarded to ensure the
 71 assumption of certain detection is met. In contrast, distance sampling uses
 72 observations that would have been discarded to model detection (although
 73 typically some detections are discarded beyond a given *truncation distance*
 74 during analysis).

75 Estimators such as eqn (1) rely on the design of the study to ensure that
 76 abundance estimates over the whole study area (scaling up from the covered
 77 region) are valid. This article focusses on *model-based* inference to extrapol-
 78 ate to a larger study area. Specifically, we consider the use of spatially explicit
 79 models to investigate the response of biological populations to biotic and abi-
 80 otic covariates that vary over the study region. A spatially-explicit model can
 81 explain the between-transect variation (which is often a large component of
 82 the variance in design-based estimates) and so using a model-based approach
 83 can lead to smaller variance in estimates of abundance. Model-based infer-
 84 ence also enables the use of data from opportunistic surveys, for example,

85 incidental data arising from “ecotourism” cruises (Williams *et al.*, 2006).

86 Our aims in creating a spatial model of a biological population are usu-
87 ally two-fold: (i) estimating overall abundance and (ii) investigating the re-
88 lationship between abundance and environmental covariates. As with any
89 predictions that are outside the range of the data, one should heed the usual
90 warnings regarding extrapolation. For example, if a model contains eleva-
91 tion as a covariate, predictions at high, unsampled elevations are unlikely to
92 be reliable. Frequently, maps of abundance or density are required and any
93 spurious predictions can be visually assessed, as well as by plotting a histo-
94 gram of the predicted values. A sensible definition of the region of interest
95 avoids prediction outside the range of the data.

96 In this article we review the current state of spatial modelling of detection-
97 corrected count data, illustrating some recent developments useful to applied
98 ecologists. The methods discussed have been available in Distance software
99 (Thomas *et al.*, 2010) for some time but the recent advances covered here
100 have been implemented in a new R package, `dsm` (Miller *et al.*, 2013) and are
101 to be incorporated into Distance.

102 Throughout this article a motivating data set is used to illustrate the
103 methods. These data are sightings of pantropical spotted dolphins (*Stenella*
104 *attenuata*) during April and May of 1996 in the Gulf of Mexico. Observers
105 aboard the NOAA vessel Oregon II recorded sightings and environmental co-
106 variates (see <http://seamap.env.duke.edu/dataset/25> for survey details).
107 A complete example analysis is provided in Appendix A. The data used in
108 the analysis are available in the `dsm` package and Distance.

109 The rest of the article reviews approaches for the spatial modelling of

110 distance sampling data before focussing on the density surface modelling ap-
111 proach of Hedley & Buckland (2004) to estimate abundance and uncertainty.
112 We then describe recent advances and provide practical advice regarding
113 model fitting, formulation and checking. Finally we discuss future directions
114 for research in spatially modelling detection-corrected count data.

115 **Approaches to spatial modelling of distance sampling** 116 **data**

117 Modelling of spatially referenced distance sampling data is equivalent to
118 modelling spatially-referenced count data, with the additional information
119 provided by collecting distances to account for imperfect detection. We re-
120 view recent efforts to model such data; some consist of two steps (correction
121 for imperfect detection, then spatial modelling), while others jointly estimate
122 the relevant parameters.

123 TWO-STAGE APPROACHES

124 The focus of this article is the “count model” of Hedley & Buckland (2004),
125 we will henceforth refer to this approach as *density surface modelling* (DSM).
126 Modelling proceeds in two steps: a detection function is fitted to the distance
127 data to obtain detection probabilities for clusters (flocks, pods, etc.) or in-
128 dividuals. Counts are allocated to corresponding segments (contiguous tran-
129 sect sections). A generalised additive model (GAM; e.g. Wood, 2006) is then
130 constructed with the per-segment counts as the response with either counts

131 or segment areas corrected for detectability (see *Density surface modelling*,
132 below). GAMs provide a flexible class of models that include generalized
133 linear models (GLMs; McCullagh & Nelder, 1989) but extend them with the
134 possible addition of splines to create smooth functions of covariates, ran-
135 dom effects terms or correlation structures. We cover advances using this
136 approach in *Recent developments*.

137 Niemi & Fernández (2010) proposed a Bayesian point process approach
138 to spatial abundance modelling. The density of the objects is described by
139 an intensity function, which included spatially-referenced covariates. Model
140 fitting proceeded in two stages: first the detection function was fitted, then
141 the spatial model (via MCMC) assuming the detection function parameters
142 were known, so detection function uncertainty was not incorporated in the
143 spatial model. A marked point process (Cox & Isham, 1980, Section 5.5)
144 could be used to incorporate cluster size information.

145 Ver Hoef *et al.* (2013) modeled seal populations in the Bering Sea using
146 a Bayesian spatial model, using a detection function to account for uncer-
147 tain detection and incorporating additional information from a (frequent-
148 ist) model of seal haul-outs on ice. The detection function and haul-out
149 model corrected the observed density estimates which were modelled using a
150 Bayesian hierarchical model for the spatial component. The Bayesian hier-
151 archical model was itself was split into two parts (*i*) a presence/absence part
152 to allow modelling of the large number of zeros in the data and (*ii*) a density
153 part also used to account for spatial autocorrelation. The analysis shows
154 that that when extra information is available (such as telemetry data for the
155 haul-out process) additional insight can be derived.

156 We note that there are many approaches to modelling spatially referenced
157 count data (Oppel *et al.*, 2011, provides an overview of such methods for
158 marine bird modelling). Also worthy of note is the approach of Barry &
159 Welsh (2002) who used a two-stage approach to model presence/absence
160 then spatial pattern (via two GAMs) to account for zero-inflation.

161 ONE-STAGE APPROACHES

162 Rather than fitting two separate models, some authors have combined the
163 detection function and spatial model fitted (mostly via hierarchical Bayesian
164 methods). The first of these was Royle *et al.* (2004), who estimated the
165 parameters of a specified detection function, formulating an unconditional
166 likelihood per-point/line as a function of the unobserved transect abund-
167 ances. These unobserved abundances were treated as random effects, integ-
168 rated out to give a per-transect likelihood as a function of detection function
169 and random effects parameters (linear functions of the environmental covari-
170 ates). Due to the multinomial nature of the per-transect likelihood proposed,
171 distance data must be allocated to bins (e.g. 0-5m, 5-15m, etc). Chelgren
172 *et al.* (2011) proposed replacing the multinomial per-transect likelihood with
173 a binomial distribution multiplied by a detection function. The binomial
174 term collapses the multinomial bins into a single bin and gives the number
175 of animals detected in the transect, thus allowing the use of exact distances.

176 The work of Schmidt *et al.* (2011) took a similar approach to Royle &
177 Dorazio (2008), building a presence/absence-type model for clusters, aug-
178 menting the data with unobserved clusters. The authors then used a Poisson
179 distribution to model cluster size (using a random effect to incorporate over-

180 dispersion), combining these parts gave a model of individual abundance.
181 Conn *et al.* (2012) also used a hierarchical Bayesian model but in terms of
182 abundance rather than density using a super-population/data augmentation
183 approach (as in Link & Barker, 2009). In their formulation, the whole pop-
184 ulation within the study region is modelled, not just those animals observed
185 during the survey.

186 Moore & Barlow (2011) adopted a hierarchical Bayesian state-space model,
187 separating the problem into observation and process components. The pro-
188 cess component described the underlying population density as it changes
189 over time and space (though the authors only included strata as a spatial
190 component). The observation part of the model then linked the process
191 model to the data via the detection function.

192 Johnson *et al.* (2010) proposed a point process-based model for distance
193 sampling data. They first assumed that the locations of all individuals in
194 the survey area (not just those observed) form a realisation of a Poisson pro-
195 cess. Parameters of the intensity function were then estimated via standard
196 maximum likelihood methods for point processes (Baddeley & Turner, 2000).
197 All parameters were estimated jointly so uncertainty from both the spatial
198 pattern and the detection function was incorporated into variance estimates
199 of the abundance. This also ensured that correlations between the detection
200 function and underlying point process are estimated correctly (and do not
201 falsely inflate or deflate variance estimates). A post-hoc correction factor was
202 used to address overdispersion unmodelled by spatial covariates (i.e. counts
203 that do not follow a Poisson mean-variance relationship).

205 Generally very little information is lost by taking a two-stage approach. This
 206 is because transects are typically very narrow compared with the width of the
 207 study area so, provided no significant density variation takes place “across”
 208 the of the lines or within the point, there is no information in the distances
 209 about the spatial distribution of animals (this is an assumption of two-stage
 210 approaches).

211 Two-stage approaches are effectively divide and conquer techniques: con-
 212 centrating on the detection function first, and then given the detection func-
 213 tion, fitting the spatial model. One-stage models are more more difficult to
 214 both estimate and check as both steps occur at once; models are potentially
 215 simpler from the perspective of the user and perhaps more mathematically
 216 elegant.

217 Two-stage models have the disadvantage that to accurately quantify model
 218 uncertainty one must appropriately combine uncertainty from the detection
 219 function and spatial models. This can be challenging; however the alternative
 220 of ignoring uncertainty from the detection process (e.g. Niemi & Fernández,
 221 2010) can produce confidence or credible intervals for abundance estimates
 222 that have coverage below the nominal level. More information regarding how
 223 variance estimation is addressed for DSMs is given in *Recent developments*.

224 Density surface modelling

225 This section focuses on modelling the density/abundance estimation stage of
 226 the DSM approach introduced previously. Both line and point transects can

227 be used, but if lines are used then they are split into contiguous *segments*
 228 (indexed by j), which are of length l_j . Segments should be small enough such
 229 that neither density of objects nor covariate values vary appreciably within
 230 a segment (usually making the segments approximately square is sufficient;
 231 $2w \times 2w$, where w is the truncation distance). The area of each segment enters
 232 the model as (or as part of) an offset: the area of segment j is $A_j = 2wl_j$
 233 and for point j is $A_j = \pi w^2$.

234 Count or estimated abundance (per segment or point) is then modelled
 235 as a sum of smooth functions of covariates (z_{jk} with k indexing the covari-
 236 ates, e.g., location, sea surface temperature, weather conditions; measured at
 237 the segment/point level) using a generalized additive model. Smooth func-
 238 tions are modelled as splines, providing flexible unidimensional (and higher-
 239 dimensional) curves (and surfaces, etc) that describe the relationship between
 240 the covariates and response. Wood (2006) and Ruppert *et al.* (2003) provide
 241 more in-depth introductions to smoothing and generalized additive models.

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

246 COUNT AS RESPONSE

247 The model for the count per segment is:

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

248 where the f_k s are smooth functions of the covariates and β_0 is an intercept
 249 term. Multiplying the segment area (A_j) by the probability of detection (\hat{p}_j)
 250 gives the *effective area* for segment j . If there are no covariates other than
 251 distance in the detection function then the probability of detection is constant
 252 for all segments (i.e., $\hat{p}_j = \hat{p}, \forall j$). The distribution of n_j can be modelled
 253 as an overdispersed Poisson, negative binomial, or Tweedie distribution (see
 254 *Recent developments*).

255 Fig. 2 shows the raw observations of the dolphin data, along with the
 256 transect lines, overlaid on the depth data. A half-normal detection function
 257 was fitted to the distances and is shown in Fig. 1. Fig. 3 shows a DSM fitted
 258 to the dolphin data. The top panel shows predictions from a model where
 259 depth was the only covariate, the bottom panel shows predictions where
 260 a (bivariate) smooth of spatial location was also included. Comparing the
 261 models using GCV score, the latter had a considerably lower score (39.12 vs
 262 48.46) and so would be selected as our preferred model.

263 As well as simply calculating abundance estimates, relationships between
 264 covariates and abundance can be illustrated via plots of marginal smooths.
 265 The effect of depth on abundance (on the scale of the link function) for the
 266 dolphin data can be seen in Fig. 4.

267 An alternative to modelling counts is to use the per-segment/circle abund-
 268 ance using distance sampling estimates as the response. In this case we
 269 replace n_j by:

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

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

272 The following model is then fitted:

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

273 where \hat{N}_j , as with n_j , is assumed to follow an overdispersed Poisson, negative
 274 binomial, or Tweedie distribution (see *Recent developments*, below). Note
 275 that the offset (A_j) is now the area of segment/point rather than effective
 276 area of the segment/point. Although \hat{N}_j can always be modelled instead of
 277 n_j , it seems preferable to use n_j when possible, as one is then modelling actual
 278 (integer) counts as the response rather than estimates. Note that although
 279 \hat{N}_j may take non-integer values, this does not present an estimation problem
 280 for the response distributions covered here.

281 *DSM with covariates at the observation level*

282 The above models consider the case where the covariates are measured at
 283 the segment/point level. Often covariates (z_{ij} , for individual/cluster i and
 284 segment/point j) are collected on the level of observations; for example sex
 285 or cluster size of the observed object or identity of the observer. In this
 286 case the probability of detection is a function of the object (individual or
 287 cluster) level covariates $\hat{p}(z_i)$. Object level covariates can be incorporated
 288 into the model by adopting the following estimator of the per-segment/point
 289 abundance:

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

290 Density can be modelled rather than abundance by not including an offset,
 291 but instead dividing the count (or estimated abundance) by the area of the
 292 segment/point (and weighting observations by the segment/point areas). We
 293 concentrate on abundance here; see Hedley & Buckland (2004) for further
 294 details on modelling density.

295 PREDICTION

296 A DSM can be used to predict abundance over a larger/different area than
 297 was originally surveyed. In that case the investigator must create a series
 298 of prediction cells over the prediction region. For each cell the covariates
 299 included in the DSM must be available; the area of each cell is also required.
 300 Having made predictions for each cell, these can be plotted as an abundance
 301 map (as in Fig. 3) or, summing over cells, an overall estimate of abundance
 302 can be calculated. It is worth noting that using prediction grid cells that are
 303 smaller than the resolution of the spatially referenced data has no effect on
 304 abundance/density estimates.

305 VARIANCE ESTIMATION

306 Estimating the variance of abundances calculated using a DSM is not straight-
 307 forward: uncertainty from the estimated parameters of the detection function
 308 must be incorporated into the spatial model. A second consideration is that
 309 in a line transect survey, abundances in adjacent segments are likely to be

310 correlated; failure to account for this spatial autocorrelation will lead to ar-
 311 tificially low variance estimates and hence misleadingly narrow confidence
 312 intervals.

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

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

318 1. Resample (with replacement) the per-segment/point residuals, store
 319 the values in \mathbf{r}_b .

320 2. Refit the model but with the response set to $\hat{\boldsymbol{\eta}} + \mathbf{r}_b$ (where $\hat{\boldsymbol{\eta}}$ are the
 321 fitted values from the original model).

322 3. Take the predicted values for the new model and store them.

323 From the predicted values stored in the last step the variance originating in
 324 the spatial part of the model can be calculated. The total variance of the
 325 abundance estimate (over the whole region of interest or sub-areas) can then
 326 be found by combining the variance estimate from the bootstrap procedure
 327 with the variance of the probability of detection from the detection function
 328 model using the delta method (which assumes that the two components of
 329 the variance are independent; Ver Hoef, 2012).

330 The above procedure assumes that there is no correlation in space between
 331 segments, which are usually contiguous along transects. If many animals are
 332 observed in a particular segment then we might expect there to be high num-
 333 bers in the adjacent segments. A moving block bootstrap (MBB; Efron &

334 Tibshirani, 1993, Section 8.6) can account for some of this spatial autocor-
335 relation in the variance estimation. The segments are grouped together into
336 overlapping blocks (so if the block size is 5, block one is segments 1, . . . , 5,
337 block two is segments 2, . . . , 6, and so on). Then, at step (2) above, resamples
338 are taken at the block level (rather than individual segments within a tran-
339 sect). Using MMB will account for correlation between the segments at scales
340 smaller than the block size, inflating the variances accordingly. Block size can
341 be selected by plotting an autocorrelogram of the residuals from the DSM.
342 Block size dictates the maximum amount of spatial autocorrelation accoun-
343 ted for, this may not fully account for the autocorrelation (testing sensitivity
344 to block size by trying several different sizes can be time consuming).

345 Both bootstrap procedures can also be modified to take into account de-
346 tection function uncertainty by simulating distances from the fitted detection
347 function and then re-calculating the offset by fitting a detection function to
348 the simulated distances.

349 Estimation of uncertainty for a given prediction region can be found by
350 calculating the appropriate quantiles of the resulting abundance estimates
351 (outlier removal may be required before quantile calculation). DSM uncer-
352 tainty can be visualised via a plot of per-cell coefficient of variation obtained
353 by dividing the standard error for each cell by its predicted abundance (as
354 in Fig. 5).

Recent developments

GAM uncertainty and variance propagation

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

Williams *et al.* (2011) go a step further and incorporate the uncertainty in the estimation of the detection function into the variance of the spatial model, albeit only when segment level covariates are in the DSM. Their procedure is to fit the density surface model with an additional random effect term that characterises the uncertainty in the estimation of the detection function (via the derivatives of the probability of detection, \hat{p} , with respect to their parameters). Variance estimates of the abundance calculated using standard GAM theory will include uncertainty from the estimation of the detection function. A more complete mathematical explanation of this result is given in Appendix B.

We consider that propagating the uncertainty in this manner to be preferable to the MBB because it is more computationally efficient meaning investigators can easily and quickly estimate variances of complex models. The confidence intervals produced via variance propagation appear comparable (if not narrower) than their bootstrap equivalents, while maintaining good coverage (results of a small simulation study are given in Appendix C).

379 Fig. 5 shows a map of the coefficient of variation for the model which
380 includes both location and depth covariates. Variance has been calculated
381 using the variance propagation method.

382 EDGE EFFECTS

383 Previous work (Ramsay, 2002; Wang & Ranalli, 2007; Wood *et al.*, 2008;
384 Scott-Hayward *et al.*, 2013; Miller & Wood, submitted) has highlighted the
385 need to take care when smoothing over areas with complicated boundaries,
386 e.g., those with rivers, peninsulae or islands. If two parts of the study area
387 (either side of a river or inlet, say) are inappropriately linked by the model
388 (i.e. if the distance between the points is measured as a straight line, rather
389 taking into account obstacles) then the boundary feature (river, etc) can
390 be “smoothed across” so positive abundances are predicted in areas where
391 animals could not possibly occur. Ensuring that a realistic spatial model has
392 been fitted to the data is essential for valid inference. The soap film smoother
393 of Wood *et al.* (2008) is an appealing solution: a bivariate smooth function
394 of location that can be included in any GAM but that allows for boundary
395 conditions to be estimated and obeyed for a complex study area. Such an
396 approach can be helpful when uncertainty is estimated via a bootstrap as
397 edge effects can also cause large, unrealistic predictions which can plague
398 other smoothers (Bravington & Hedley, 2009).

399 Even if the study area does not have a complicated boundary, edge effects
400 can still be problematic. Miller (2012) notes that some smoothers have plane
401 components that tend to cause the fitted surface to increase unrealistically as
402 predictions are made further away from the locations of survey effort. This

403 problem can be alleviated by the using a different type of smoother (e.g. a
404 generalisation of thin plate regression splines called *Duchon splines*).

405 TWEEDIE DISTRIBUTION

406 The Tweedie distribution offers a flexible alternative to the quasi-Poisson and
407 negative binomial distributions as a response distribution when modelling
408 count data (Candy, 2004). In particular it is useful when there are a high
409 proportion of zeros in the data (Shono, 2008; Peel *et al.*, 2012) and avoids
410 multiple-stage modelling of zero-inflated data (as in Barry & Welsh, 2002).

411 The distribution has three parameters parameters: a mean, dispersion
412 and a third power parameter, which leads to additional flexibility. The dis-
413 tribution does not change appreciably when the power parameter is changed
414 by less than 0.1 and therefore a simple line search over the possible values
415 for the power parameter is usually a reasonable approach to estimating the
416 parameter. Mark Bravington (pers. comm.) suggested plotting the square
417 root of the absolute value of the residuals against fitted values; a “flatter”
418 plot (points forming a horizontal line) give an indication of a “good” value.
419 We additionally suggest using the metrics described in the next section for
420 model selection.

421 Appendix D gives further details about the Tweedie distribution (includ-
422 ing its PDF and further references).

423 Practical advice

424 A flow diagram of the modelling process for creating a DSM is shown in Fig.
425 6. The diagram shows which methods are compatible with each other and
426 what the options are for modelling a particular data set.

427 In our experience, it is sensible to obtain a detection function that fits
428 the data as well as possible and only begin spatial modelling after a satisfact-
429 ory detection function has been obtained. Model selection for the detection
430 function can be performed using AIC and model checking using goodness-of-
431 fit tests given in Burnham *et al.* (2004, Section 11.11). If animals occur in
432 clusters rather than individually, bias can be incurred due to the higher visib-
433 ility of larger clusters. It may then be necessary to include size as a covariate
434 in the detection function (see Buckland *et al.*, 2001, Section 4.8.2.4). For
435 some species cluster size may change according to location, Ferguson *et al.*
436 (2006) use two GAMs (one to model observed clusters and one to model the
437 cluster size) to deal with spatially-varying cluster size amongst delphinids,
438 though the authors do not present the variance of the resulting predictions.

439 Smooth terms can be selected using (approximate) p -values (Wood, 2006,
440 Section 4.8.5). An additional useful technique for covariate selection is to
441 use an extra penalty for each term in the GAM allowing smooth terms to
442 be removed from the model during fitting (illustrated in Appendix A; Wood,
443 2011). Smoothness selection is performed by generalized cross validation
444 (GCV) score, unbiased risk estimator (UBRE) or restricted maximum likeli-
445 hood (REML) score. When model covariates are effectively functions of one
446 another (e.g. depth could be written as a function of location) GCV and

447 UBRE can suffer from optimisation problems (Wood, 2006, Section 4.5.3)
448 which can lead to unstable models (Wood, 2011). REML provides a fitting
449 criteria with a more pronounced optima which avoids some problems with
450 parameter estimation, though caution should always be taken when deal-
451 ing with highly correlated covariates. A significant drawback of REML is
452 that scores cannot be used to compare models with different linear terms or
453 offsets (Wood, 2011), though the p -value and additional penalty techniques
454 described above can be used to select model terms. We highly recommend
455 the use of standard GAM diagnostic plots; Wood (2006) provides further
456 practical information on GAM model selection and fitting.

457 In the analysis of the dolphin data we included a smooth of location that
458 nearly doubles the percentage deviance explained (27.3% to 52.7%). One can
459 see this when comparing the two plots in Fig. 3 and the plot of the depth
460 (Fig. 2), the plot of the model containing only a smooth of depth looks very
461 similar to the raw plot of the depth data. Using a smooth of location can be
462 a primitive way to account for spatial autocorrelation and/or as a proxy for
463 other spatially varying covariates that are unavailable.

464 A more sophisticated way to account for spatial autocorrelation between
465 segments (within transects) is to use a an autocorrelation structure within
466 the DSM (e.g. autoregressive models). Appendix A shows an example using
467 generalized additive mixed model (GAMMs; Wood, 2006, Section 6.6, see
468 Appendix A for an example) to construct an autoregressive (lag 1) correla-
469 tion structure. This gives a significant reduction in variance, tightening the
470 confidence interval around the abundance estimate.

471 In the analysis presented here, spatial location has been transformed from

latitude and longitude to kilometres north and east of the centre of the survey region at $(27.01^\circ, -88.3^\circ)$. This is because the bivariate smoother used (the thin plate spline; Wood, 2003) is isotropic: there is only one parameter controlling the smoothness in both directions. Moving one degree in latitude is not the same as moving one degree in longitude and so using kilometres from the centre of the study region makes the covariates isotropic. Using metric units rather than non-standard units of measure such as degrees or feet throughout makes analysis much easier.

A smooth of an environment-level covariate such as depth can be very useful for assessing the relationships between abundance and the covariate (as in Fig. 4). Caution should be employed when interpreting smooth relationships and abundance estimates, especially if there are gaps over the range of covariate values. Large counts may occur at large values of depth but if no further observations occur at such a large value, then investigators should be skeptical of any relationship.

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. Spatial models can be particularly useful when it comes to prediction: making predictions for some subset of the study area relies on stratification in design-based methods and as such can be rather limited. Our models also allow inference from a sample of sightings to a population in a study area without depending upon a random sample design,

495 and therefore data collected from "platforms of opportunity" (Williams *et al.*,
496 2006) can be used (although a well designed survey is always preferable).

497 Unbiased estimates are dependent upon either (i) distribution of sampling
498 effort being random throughout the study area (for design-based inference)
499 or (ii) model correctness (for model-based inference). It is easier to have
500 confidence in the former rather than in the latter because our models are
501 always wrong. Nevertheless model-based inference will play an increasing
502 role in population assessment as the availability of spatially-referenced data
503 increases.

504 The field is quickly evolving to allow modelling of more complex data
505 building on the basic ideas of density surface modelling. We expect to see
506 large advances in temporal inferences and the handling of zero-inflated data
507 and spatial correlation. These should become more mainstream as modern
508 spatio-temporal modelling techniques are adopted. Petersen *et al.* (2011)
509 provided a very basic framework for temporal modelling; their model included
510 "before" and "after" smooth terms to quantify the impact of the construction
511 of an offshore windfarm. Zero-inflation in count data may be problematic
512 and two-stage approaches such as Barry & Welsh (2002) as well as more flex-
513 ible response distributions made possible by Rigby & Stasinopoulos (2005)
514 have yet to be exploited by those using distance sampling data. Spatial
515 autocorrelation can be accounted for via approaches that explicitly intro-
516 duce correlations such as generalized estimating equations (GEEs; Hardin &
517 Hilbe, 2003) or generalized additive mixed models or via mechanisms such
518 as that of Skaug (2006), which allow observations to cluster according to one
519 of several states (such as high vs low density patches, possibly in response to

temporary agglomerations of prey, although the mechanism is unimportant). These advances should assist both modellers and wildlife managers to make optimal conservation decisions.

Advances in Bayesian computation (INLA; Rue *et al.*, 2009), make one-step, Bayesian, density surface models computationally feasible (as INLA is an alternative to MCMC). We anticipate that such a direct modelling technique will dominate future developments in the field.

Density surface modelling allows wildlife managers to make best use of the available spatial data to understand patterns of abundance, and hence make better conservation decisions (e.g., about reserve or development placement). The recent advances mentioned here increase the reliability of the outputs from a modelling exercise, and hence the efficacy of these decisions. Density surface modelling from survey data is an active area of research, and we look forward to further improvements and extensions in the near future.

Acknowledgments

We wish to thank Paul Conn, another anonymous reviewer, and the associate editor for their helpful comments. DLM wishes to thank Mark Bravington and Sharon Hedley for their detailed discussions and for providing code for their variance propagation method. Funding for the implementation of the recent advances into the `dsm` package and Distance software came from the US Navy, Chief of Naval Operations (Code N45), grant number N00244-10-1-0057.

542 References

- 543 Baddeley, A. & Turner, R. (2000) Practical maximum pseudolikelihood for spatial
544 point patterns. *Australian & New Zealand Journal of Statistics*, **42**, 283–322.
- 545 Barry, S.C. & Welsh, A.H. (2002) Generalized additive modelling and zero inflated
546 count data. *Ecological Modelling*, **157**, 179–188.
- 547 Bravington, M.V. & Hedley, S.L. (2009) Antarctic minke whale abundance estim-
548 ates from the second and third circumpolar IDCR/SOWER surveys using the
549 SPLINTR model.
- 550 Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. &
551 Thomas, L. (2001) *Introduction to Distance Sampling*. Oxford University Press.
- 552 Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. &
553 Thomas, L. (2004) *Advanced Distance Sampling*. Oxford University Press.
- 554 Burnham, K.P., Buckland, S.T., Laake, J.L., Borchers, D.L., Marques, T.A.,
555 Bishop, J.R. & Thomas, L. (2004) Further topics in distance sampling. *Ad-
556 vanced Distance Sampling* (eds. S.T. Buckland, D.R. anderson, K.P. Burnham,
557 J.L. Laake, D.L. Borchers & L. Thomas). Oxford University Press.
- 558 Candy, S. (2004) Modelling catch and effort data using generalised linear models,
559 the Tweedie distribution, random vessel effects and random stratum-by-year
560 effects. *CCAMLR Science*, **11**, 59–80.
- 561 Chelgren, N.D., Samora, B., Adams, M.J. & McCreary, B. (2011) Using spati-
562 otemporal models and distance sampling to map the space use and abundance
563 of newly metamorphosed western toads (*Anaxyrus boreas*). *Herpetological Con-
564 servation and Biology*, **6**, 175–190.
- 565 Conn, P.B., Laake, J.L. & Johnson, D.S. (2012) A hierarchical modeling framework
566 for multiple observer transect surveys. *PLoS ONE*, **7**, e42294.
- 567 Cox, D.R. & Isham, V. (1980) *Point Processes*. Monographs on Applied Probability
568 and Statistics. Chapman and Hall. ISBN 9780412219108.
- 569 Efron, B. & Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Chapman &
570 Hall/CRC. ISBN 9780412042317.
- 571 Ferguson, M.C., Barlow, J., Fiedler, P., Reilly, S.B. & Gerrodette, T. (2006) Spatial
572 models of delphinid (family Delphinidae) encounter rate and group size in the
573 eastern tropical Pacific Ocean. *Ecological Modelling*, **193**, 645–662.

- 574 Halpin, P., Read, A., Fujioka, E., Best, B., Donnelly, B., Hazen, L., Kot, C., Urian,
575 K., LaBrecque, E., Dimatteo, A., Cleary, J., Good, C., Crowder, L. & Hyrenbach,
576 K.D. (2009) OBIS-SEAMAP: The world data Center for marine mammal, sea
577 bird, and sea turtle distributions. *Oceanography*, **22**, 104–115.
- 578 Hardin, J. & Hilbe, J. (2003) Generalized Estimating Equations. Chapman and
579 Hall/CRC, London, UK.
- 580 Hedley, S.L. & Buckland, S.T. (2004) Spatial models for line transect sampling.
581 *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 181–199.
- 582 Johnson, D.S., Laake, J.L. & Ver Hoef, J.M. (2010) A model-based approach for
583 making ecological inference from distance sampling data. *Biometrics*, **66**, 310–
584 318.
- 585 Link, W.A. & Barker, R.J. (2009) *Bayesian Inference: with ecological applications*.
586 Academic Press, London, UK.
- 587 Marques, T.A., Thomas, L., Fancy, S. & Buckland, S.T. (2007) Improving estimates
588 of bird density using multiple-covariate distance sampling. *The Auk*, **124**, 1229–
589 1243.
- 590 McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*. Chapman &
591 Hall/CRC.
- 592 Miller, D.L. (2012) *On smooth models for complex domains and distances*. Ph.D.
593 thesis, University of Bath.
- 594 Miller, D.L., Rexstad, E.A., Burt, M.L., Bravington, M.V. & Hedley, S.L. (2013)
595 *dsm: Density surface modelling of distance sampling data*.
596 URL <http://github.com/dill/dsm>
- 597 Miller, D.L. & Wood, S.N. (submitted) Finite area smoothing with generalized
598 distance splines. pp. 1–27.
- 599 Moore, J.E. & Barlow, J. (2011) Bayesian state-space model of fin whale abundance
600 trends from a 1991-2008 time series of line-transect surveys in the California
601 Current. *Journal of Applied Ecology*, **48**, 1195–1205.
- 602 Niemi, A. & Fernández, C. (2010) Bayesian spatial point process modeling of line
603 transect data. *Journal of Agricultural, Biological, and Environmental Statistics*,
604 **15**, 327–345.
- 605 Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O’Connell, A., Miller, P. &
606 Louzao, M. (2011) Comparison of five modelling techniques to predict the spatial
607 distribution and abundance of seabirds. *Biological Conservation*.

- 608 Peel, D., Bravington, M.V., Kelly, N., Wood, S.N. & Knuckey, I. (2012) A Model-
609 Based Approach to Designing a Fishery-Independent Survey. *Journal of Agri-
610 cultural, Biological, and Environmental Statistics*, **18**, 1–21.
- 611 Petersen, I.K., MacKenzie, M.L., Rexstad, E.A., Wisz, M.S. & Fox, A.D. (2011)
612 Comparing pre- and post-construction distributions of long-tailed ducks *Clan-
613 gula hyemalis* in and around the Nysted offshore wind farm, Denmark: a quasi-
614 designed experiment accounting for imperfect detection, local surface features
615 and autocorrelation. Technical report 2011-1, Centre for Research into Environ-
616 mental and Ecological Modelling.
- 617 Ramsay, T. (2002) Spline smoothing over difficult regions. *Journal of the Royal
618 Statistical Society. Series B, Statistical Methodology*, **64**, 307–319.
- 619 Rigby, R. & Stasinopoulos, D. (2005) Generalized additive models for location, scale
620 and shape. *Journal of the Royal Statistical Society-Series C Applied Statistics*,
621 **54**, 507–554.
- 622 Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical Modeling and Inference in Ecology*.
623 Academic Press, London, UK.
- 624 Royle, J., Dawson, D. & Bates, S. (2004) Modeling abundance effects in distance
625 sampling. *Ecology*, **85**, 1591–1597.
- 626 Rue, H., Martino, S. & Chopin, N. (2009) Approximate Bayesian inference for
627 latent Gaussian models by using integrated nested Laplace approximations. *J.
628 R. Statist. Soc. B*, **71**, 319–392.
- 629 Ruppert, D., Wand, M. & Carroll, R.J. (2003) *Semiparametric Regression*. Cam-
630 bridge Series on Statistical and Probabilistic Mathematics. Cambridge University
631 Press.
- 632 Schmidt, J.H., Rattenbury, K.L., Lawler, J.P. & Maccluskie, M.C. (2011) Using
633 distance sampling and hierarchical models to improve estimates of Dall’s sheep
634 abundance. *The Journal of Wildlife Management*, **76**, 317–327.
- 635 Scott-Hayward, L.A.S., MacKenzie, M.L., Donovan, C.R., Walker, C.G. & Ashe,
636 E. (2013) Complex region spatial smoother (CReSS). *Journal of Computational
637 and Graphical Statistics*.
- 638 Shono, H. (2008) Application of the Tweedie distribution to zero-catch data in
639 CPUE analysis. *Fisheries Research*, **93**, 154–162.
- 640 Skaug, H.J. (2006) Markov modulated Poisson processes for clustered line transect
641 data. *Environmental and Ecological Statistics*, **13**, 199–211.

- 642 Thomas, L., Buckland, S.T., Rexstad, E.A., Laake, J.L., Strindberg, S., Hedley,
643 S.L., Bishop, J.R., Marques, T.A. & Burnham, K.P. (2010) Distance software:
644 design and analysis of distance sampling surveys for estimating population size.
645 *Journal of Applied Ecology*, **47**, 5–14.
- 646 Ver Hoef, J.M. (2012) Who invented the delta method? *The American Statistician*,
647 **66**, 124–127.
- 648 Ver Hoef, J.M., Cameron, M.F., Boveng, P.L., London, J.M. & Moreland, E.E.
649 (2013) A spatial hierarchical model for abundance of three ice-associated seal
650 species in the eastern Bering Sea. *Statistical Methodology*, pp. 1–44.
- 651 Wang, H. & Ranalli, M. (2007) Low-rank smoothing splines on complicated do-
652 mains. *Biometrics*, **63**, 209–217.
- 653 Williams, R., Hedley, S.L., Branch, T.A., Bravington, M.V., Zerbini, A.N. & Find-
654 lay, K.P. (2011) Chilean blue whales as a case study to illustrate methods to
655 estimate abundance and evaluate conservation status of rare species. *Conserva-
656 tion Biology*, **25**, 526–535.
- 657 Williams, R., Hedley, S.L. & Hammond, P. (2006) Modeling distribution and
658 abundance of Antarctic baleen whales using ships of opportunity. *Ecology and
659 Society*, **11**, 1.
- 660 Wood, S.N. (2003) Thin plate regression splines. *Journal of the Royal Statistical
661 Society. Series B, Statistical Methodology*, **65**, 95–114.
- 662 Wood, S.N. (2006) *Generalized Additive Models: An introduction with R*. Chapman
663 & Hall/CRC.
- 664 Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal like-
665 lihood estimation of semiparametric generalized linear models. *Journal of the
666 Royal Statistical Society. Series B, Statistical Methodology*, **73**, 3–36.
- 667 Wood, S.N., Bravington, M.V. & Hedley, S.L. (2008) Soap film smoothing. *Journal
668 of the Royal Statistical Society. Series B, Statistical Methodology*, **70**, 931–955.

Figures

Fig. 1 Estimated detection function for pantropical dolphin clusters overlaid onto the scaled histogram of observed distances. Distances are recorded in metres.

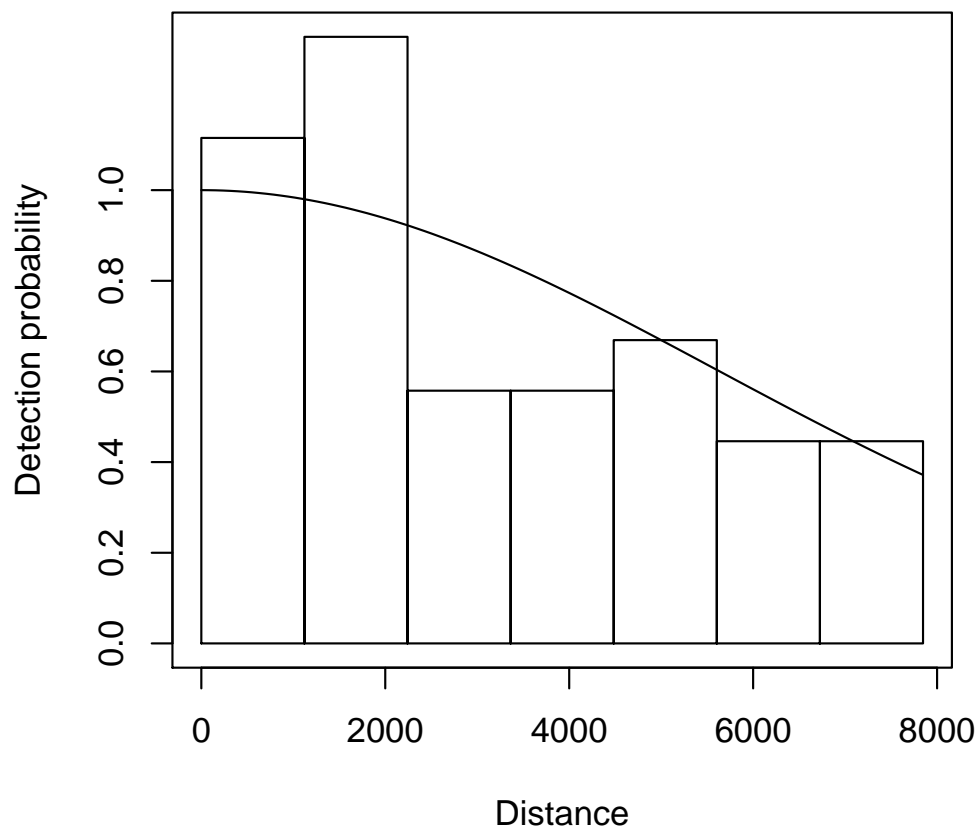


Fig. 2 The region, transect centrelines and location of detected pantropical dolphin clusters, where size of circle corresponds to the cluster size, overlaid onto depth data.

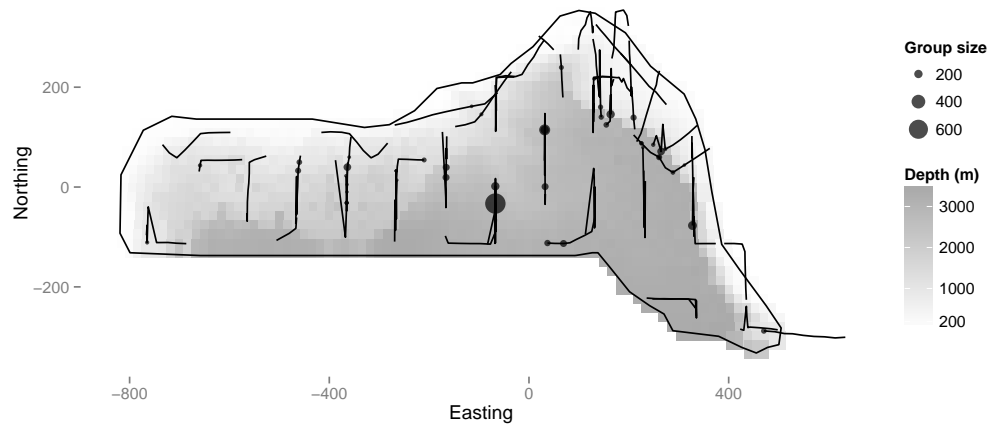


Fig. 3 Predicted abundance of dolphins from the DSM using only depth as an explanatory variable (top) and the model using both depth and location (bottom).

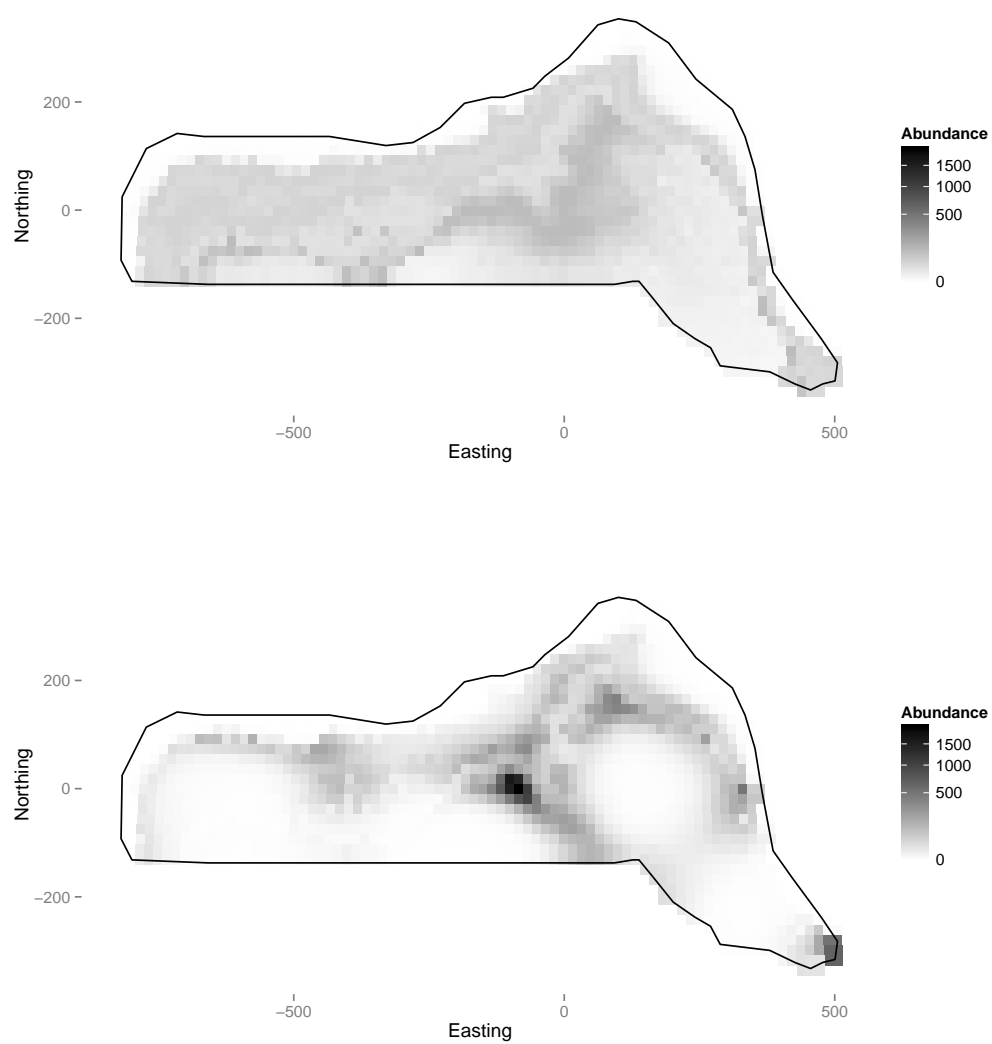


Fig. 4 Plot of the effect on the response of depth (from the model with both depth and location smooths), note that it is possible to draw a straight line between 750m and 3000m within the confidence band (between the dashed lines), so the wiggles in the smooth may not be indicative of any relationship. What is clear is that there the estimated number of dolphins increases up to about 500m. The rug ticks at the bottom of the plot indicate we have good coverage of the range of depth values in the survey area. Note that the y axis in such plots is on the scale of the link function (log in this case), so care should be taken in their interpretation.

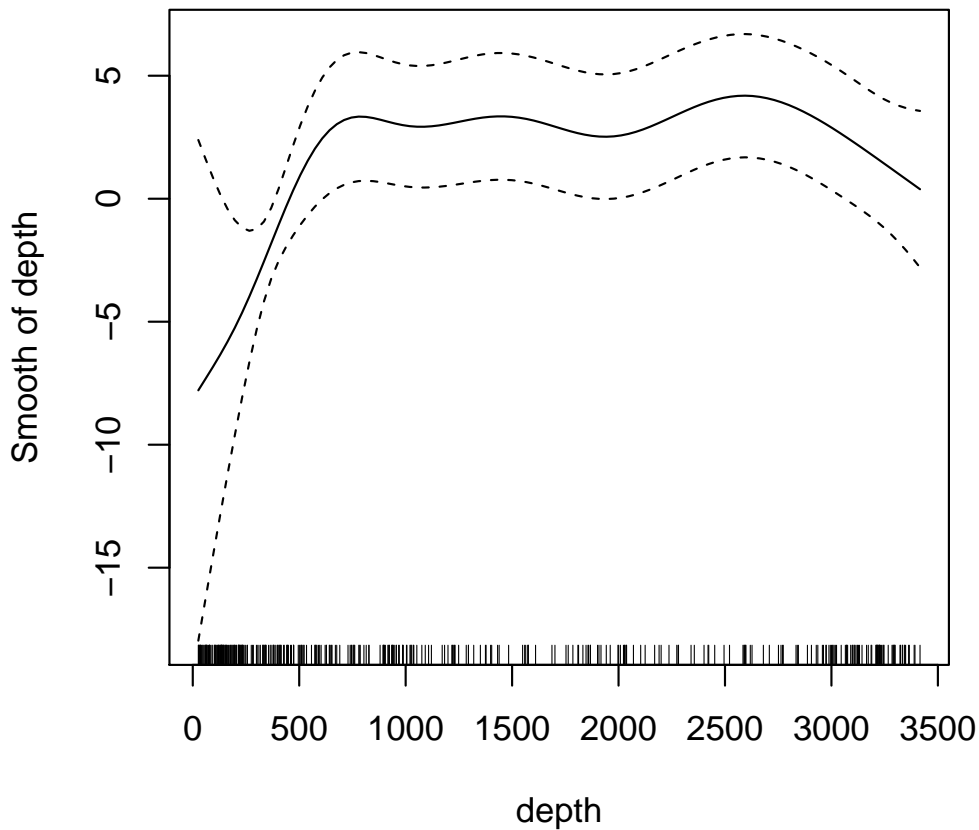


Fig. 5 Map of the coefficients of variation for the model with smooths of both depth and location. Uncertainty was estimated using the variance propagation method of Williams *et al.* (2011). As might be expected, there is high uncertainty where there is low sampling effort (Fig. 2).

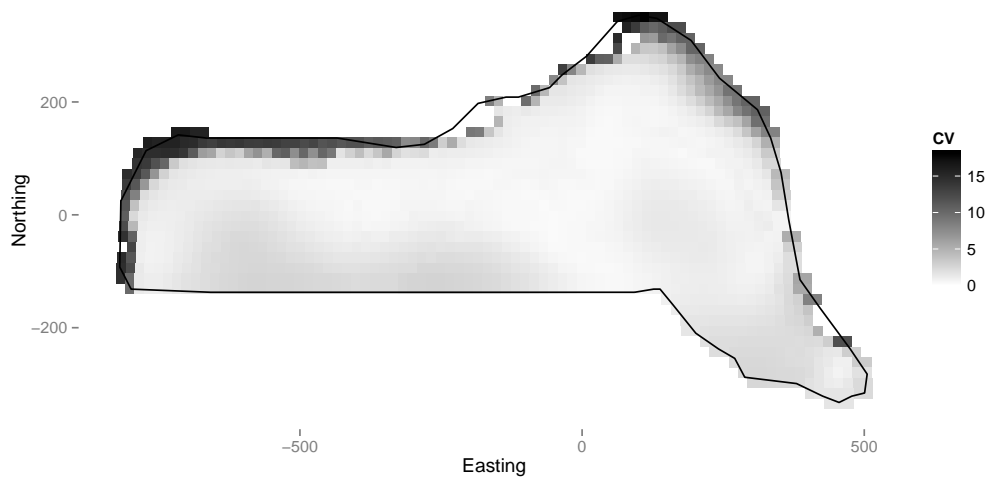


Fig. 6 Flow diagram showing the modelling process for creating a density surface model.

