

Model-Based Distance Sampling

S. T. BUCKLAND, C. S. OEDEKOVEN, and D. L. BORCHERS

Conventional distance sampling adopts a mixed approach, using model-based methods for the detection process, and design-based methods to estimate animal abundance in the study region, given estimated probabilities of detection. In recent years, there has been increasing interest in fully model-based methods. Model-based methods are less robust for estimating animal abundance than conventional methods, but offer several advantages: they allow the analyst to explore how animal density varies by habitat or topography; abundance can be estimated for any sub-region of interest; they provide tools for analysing data from designed distance sampling experiments, to assess treatment effects. We develop a common framework for model-based distance sampling, and show how the various model-based methods that have been proposed fit within this framework.

Key Words: Distance sampling; Line transect sampling; Model-based inference; Point transect sampling.

1. INTRODUCTION

Distance sampling is a suite of methods for estimating animal abundance (Buckland et al. 2001). Surveys are conducted on a set of plots, selected from a wider study region according to some randomized scheme (usually a random systematic sample or stratified random systematic sample). The two most commonly used methods are line transect sampling, for which the plots are long, narrow strips, and an observer travels along each strip centreline, recording the distance from the line of each animal detected; and point transect sampling, for which the plots are circles, and the observer searches for animals from the centre of each circular plot, recording the distance of each detected animal from the centre point.

Conventional distance sampling is a mix of design-based and model-based methods. Models are proposed for the detection function $g(y)$, which is the probability of detection of an animal, expressed as a function of its distance y from the line or point, and these are fitted to the distance data using maximum likelihood methods. This component of estimation is therefore model-based. However, the likelihood maximized is not the full likelihood, but a conditional likelihood: the likelihood of the distances y , conditional on the number n of animals detected.

S. T. Buckland (✉), C. S. Oedekoven and D. L. Borchers Centre for Research into Ecological and Environmental Modelling, The Observatory, Buchanan Gardens, University of St Andrews, St Andrews KY16 9LZ, Scotland, UK (E-mail: steve@st-andrews.ac.uk).

© 2015 The Author(s)
Journal of Agricultural, Biological, and Environmental Statistics
DOI: 10.1007/s13253-015-0220-7

Conventional distance sampling exploits design-based methods in two ways. First, we rely on the design to ensure that on average, animals are distributed uniformly over the sample plots. The distances y are then sufficient to estimate $g(y)$, and hence abundance on the plots, with the additional assumption that $g(0) = 1$; that is, an animal at distance zero (i.e. on the line or at the point) is detected with certainty. Given estimates of abundance on the plots, together with a randomized design, we can then extrapolate to the wider study area using design-based methods, to estimate total abundance.

Conventional distance sampling estimators have proven very effective for estimating abundance (Fewster and Buckland 2004). However, full model-based methods offer greater flexibility: they can be used to analyse designed experiments that use distance sampling methods, for example to test whether animal densities on treatment plots differ significantly from those on control plots; they allow animal density to be modelled as a function of spatial variables that reflect for example habitat or climate; abundance can be estimated for any sub-region of interest.

Several model-based methods have been proposed. Borchers et al. (2002) specified a binomial model for the number n of animals detected out of a population of size N , and multiplied the resulting likelihood by the line transect or point transect likelihood arising from the detection function from conventional distance sampling. Royle and Dorazio (2008) also adopted this approach for line transect sampling. Plot abundance and plot count models were developed by Royle et al. (2004), by Buckland et al. (2009) and by Oedekoven et al. (2013, 2014). These enabled data from designed distance sampling experiments to be analysed.

Stoyen (1982) and Höglmander (1991) were the first to consider point process models for line transect sampling, an approach developed further by Hedley and Buckland (2004), who specified a non-homogeneous Poisson process model for n to develop a spatial distance sampling model. Johnson et al. (2010) provided software for fitting such models, and Miller et al. (2013) provided software for a simpler approach suggested by Hedley and Buckland (2004) based on generalized additive models.

In this paper, we show how the various full model-based methods relate to each other, and how they fit within a more general framework. This framework allows distances of detected animals from the line or point to be grouped (interval) or exact, includes Poisson, binomial and multinomial models, and allows additional covariates (other than distance) to be included for modelling both detection probability and counts. For counts, the covariates are assumed to be recorded at the plot level or higher, whilst for detection probability, covariates may be at the individual animal level. In Sect. 2.1, we consider the model-based conventional distance sampling methods of Borchers et al. (2002) for exact and grouped data. In Sect. 2.2, we add covariates other than distance to the detection function model (model-based multiple-covariate distance sampling), and in Sect. 2.3, we consider model-based mark-recapture distance sampling. In Sect. 3.1, we develop plot count models, and we show that these are equivalent to the plot abundance models of Royle et al. (2004) in Sect. 3.2. We show useful generalizations incorporating random effects in Sect. 4, and present a case study in Sect. 5. In Sect. 6, we provide a brief summary of other examples, and discuss the pros and cons of a fully model-based approach to distance sampling, relative to the more usual hybrid approach.

2. NON-SPATIAL MODEL-BASED METHODS

2.1. MODEL-BASED CONVENTIONAL DISTANCE SAMPLING

In model-based distance sampling, we introduce a likelihood component for sample size (i.e. number of detected animals) n . The objective is to estimate mean animal density or animal abundance in the study area based on a sightings survey conducted along a sample of lines or at a sample of points, distributed according to a randomized design (typically a systematic random sample, possibly with stratification).

2.1.1. Exact Distance Data

Suppose that $0 \leq y \leq w$, where w is the half-width of the strip (line transect sampling) or the radius of the circle (point transect sampling). (In the case of line transect sampling, we fold the distances over, so that distances to the left of the line are pooled with distances to the right of the line.) Denote the full likelihood by $\mathcal{L}_{n,y}$. We assume that this can be expressed as the product of two likelihoods, one \mathcal{L}_n corresponding to sample size n , and the other \mathcal{L}_y corresponding to the distances y . Then we can write (Borchers and Burnham 2004)

$$\mathcal{L}_y = \prod_{i=1}^n f_y(y_i) = \prod_{i=1}^n \frac{g(y_i)\pi_y(y_i)}{P_a}, \quad (1)$$

where $f_y(y)$ is the probability density function of distance y , $g(y)$ is the probability that an animal at distance y from the line or point is detected, $\pi_y(y)$ is the distribution of distances of animals from the line or the point, irrespective of whether they are detected and P_a is the probability that an animal on the plot is detected, unconditional on its distance y . Thus we have

$$P_a = \int_0^w g(y)\pi_y(y) dy \quad (2)$$

which is the normalizing constant in (1) to ensure that $f_y(y)$ is a valid probability density function.

Given random placement of plots, then $\pi_y(y) = 1/w$, independent of y , for line transect sampling, and $\pi_y(y) = 2y/(w^2)$ for point transect sampling (Borchers and Burnham 2004).

We need a model for sample size n . A natural model is the binomial distribution:

$$\mathcal{L}_n = \binom{N}{n} (\gamma_c P_a)^n (1 - \gamma_c P_a)^{N-n}, \quad (3)$$

where N is the number of animals in the study region and γ_c is the probability that an animal within the study region is on one of the surveyed plots.

The full likelihood is thus

$$\mathcal{L}_{n,y} = \mathcal{L}_n \times \mathcal{L}_y = \binom{N}{n} (\gamma_c)^n (1 - \gamma_c P_a)^{N-n} \prod_{i=1}^n g(y_i)\pi_y(y_i). \quad (4)$$

This formulation is given by [Borchers et al. \(2002\)](#). For the case of line transect sampling, [Royle and Dorazio \(2008\)](#) give the same formulation, and refer to it as individual-based modelling, because each detected animal i has its own detection distance y_i . We can proceed to inference adopting for example maximum likelihood or Bayesian methods.

If instead we adopt a Poisson model for n , we can have a spatial distance sampling model. Adopting a non-homogeneous Poisson process model, we can write the likelihood as ([Hedley and Buckland 2004](#))

$$\mathcal{L}_{n,l} = \exp[-\mu_A] \prod_{i=1}^n D(l_i)g(y(l_i))/n! \quad (5)$$

for $n = 1, 2, \dots$, where l_i is the location of detected animal i , $y(l_i)$ is its distance from the transect, $g(y(l_i))$ is the corresponding probability of detection, $D(l_i)$ is the density of animals at location l_i and $\mu(A) = \int_A D(l)g(y(l)) dl$, where the integral is over the entire survey region A .

2.1.2. Grouped Distance Data

If the distances y are grouped into intervals defined by cutpoints $c_0 = 0, c_1, \dots, c_J = w$, then we can still define \mathcal{L}_n as above, but we replace \mathcal{L}_y by ([Borchers and Burnham 2004](#))

$$\mathcal{L}_m = \left(\frac{n!}{\prod_{j=1}^J m_j!} \right) \prod_{j=1}^J f_j^{m_j}, \quad (6)$$

where m_j is the number of detections in distance interval j , with $\sum_{j=1}^J m_j = n$, and

$$f_j = \int_{c_{j-1}}^{c_j} f(y) dy = \frac{\int_{c_{j-1}}^{c_j} g(y) \pi_y(y) dy}{P_a}. \quad (7)$$

The full likelihood is now $\mathcal{L}_{n,m} = \mathcal{L}_n \times \mathcal{L}_m$.

2.2. MODEL-BASED MULTIPLE-COVARIATE DISTANCE SAMPLING

2.2.1. Exact Distance Data

If our model for the detection function $g(y)$ includes a scale parameter σ , we may model the scale parameter as a function of covariates. Adopting the approach of [Marques and Buckland \(2003\)](#), we write for observation i

$$\sigma(\mathbf{z}_i) = \exp(\beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \dots), \quad (8)$$

where $\mathbf{z}_i = (z_{1i}, z_{2i}, \dots)'$ are the covariate values associated with observation i , $i = 1, \dots, n$.

The full likelihood now consists of three components: the likelihood for the count model (3), the likelihood for the distribution of covariates \mathbf{z} that are part of the detection model and the likelihood for the observed distances y given covariates \mathbf{z} : $\mathcal{L}_n \times \mathcal{L}_z \times \mathcal{L}_{y|\mathbf{z}}$. Given random line or point placement, we can assume that the joint distribution $\pi_{y,\mathbf{z}}(y, \mathbf{z}) = \pi_{y|\mathbf{z}}(y|\mathbf{z})\pi_z(\mathbf{z}) = \pi_y(y)\pi_z(\mathbf{z})$; that is that the distribution of distances y for all animals on sample plots (whether detected or not) is independent of that of covariates \mathbf{z} . This allows us to factorize the joint likelihood $\mathcal{L}_{z,y} = \mathcal{L}_z \times \mathcal{L}_{y|\mathbf{z}}$ as two components, the first of which is a function of \mathbf{z} alone, and the second a function of y alone (Borchers and Burnham 2004). We now have

$$\mathcal{L}_n = \binom{N}{n} (\gamma_c P_a)^n (1 - \gamma_c P_a)^{N-n}, \quad (9)$$

$$\mathcal{L}_z = \prod_{i=1}^n \frac{P_a(\mathbf{z}_i)\pi_z(\mathbf{z}_i)}{P_a} \quad (10)$$

and

$$\mathcal{L}_{y|\mathbf{z}} = \prod_{i=1}^n f(y_i|\mathbf{z}_i) = \prod_{i=1}^n \frac{g(y_i, \mathbf{z}_i)\pi_y(y_i)}{P_a(\mathbf{z}_i)}, \quad (11)$$

where

$$P_a(\mathbf{z}_i) = \int_0^w g(y, \mathbf{z}_i)\pi_y(y) dy \quad (12)$$

and

$$P_a = \int_{\mathbf{z}} P_a(\mathbf{z})\pi_z(\mathbf{z}) d\mathbf{z}. \quad (13)$$

Thus we have the full likelihood

$$\mathcal{L}_{n,z,y} = \mathcal{L}_n \times \mathcal{L}_z \times \mathcal{L}_{y|\mathbf{z}} = \binom{N}{n} (\gamma_c)^n (1 - \gamma_c P_a)^{N-n} \prod_{i=1}^n \pi_z(\mathbf{z}_i) g(y_i, \mathbf{z}_i) \pi_y(y_i). \quad (14)$$

In general, the integral of (13) will be a multiple integral. Inference is now more problematic because $\pi_z(\mathbf{z})$, unlike $\pi_y(y)$, is unknown, and so we need to specify a model for it. Where we have just a single covariate, and one for which we can specify a suitable model, then this approach may be useful. However, with multiple covariates, and the need to specify a model for their joint distribution, this approach is unappealing. In this circumstance, we can instead simply maximize the conditional likelihood $\mathcal{L}_{y|\mathbf{z}}$. If we need to estimate abundance, then we can use an estimate of $P_a(\mathbf{z}_i)$ and hence estimate the inclusion probability $\gamma_c P_a(\mathbf{z}_i)$ for detection i (Borchers and Burnham 2004). If this probability were known, then the Horvitz–Thompson estimator of population size would be

$$\hat{N} = \sum_{i=1}^n \frac{1}{\gamma_c P_a(\mathbf{z}_i)}. \quad (15)$$

We obtain an estimate $\hat{P}_a(\mathbf{z}_i)$ by maximizing the likelihood component $\mathcal{L}_{y|z}$, giving the following Horvitz–Thompson-like estimator:

$$\hat{N} = \sum_{i=1}^n \frac{1}{\gamma_c \hat{P}_a(\mathbf{z}_i)}. \quad (16)$$

If animals occur in groups (termed ‘clusters’ in the distance sampling literature), and the size of the i th detected group is s_i (which may or may not be one of the covariate values in \mathbf{z}_i), then the Horvitz–Thompson-like estimator is

$$\hat{N} = \sum_{i=1}^n \frac{s_i}{\gamma_c \hat{P}_a(\mathbf{z}_i)}. \quad (17)$$

2.2.2. Grouped Distance Data

When covariates are only recorded at plot level or higher (e.g. stratum), we will consider the analysis of grouped distance data under plot count models. If there are covariates recorded at the level of the individual animal, then we may replace (7) by

$$f_{ij} = \int_{c_{j-1}}^{c_j} f(y|\mathbf{z}_i) dy = \frac{\int_{c_{j-1}}^{c_j} g(y|\mathbf{z}_i) \pi_y(y) dy}{P_a(\mathbf{z}_i)}, \quad (18)$$

and replace (6) by

$$\mathcal{L}_m = \prod_{j=1}^J \prod_{i=1}^{m_j} f_{ij}. \quad (19)$$

2.3. MODEL-BASED MARK-RECAPTURE DISTANCE SAMPLING

In mark-recapture distance sampling, two observers survey the same plots, recording which individuals were seen by both and which by only one or the other. This allows us to remove the assumption that all animals on the line or at the point are detected. We can extend multiple-covariate distance sampling full-likelihood methods to mark-recapture distance sampling simply by including an additional likelihood component. We denote the *capture history* of an animal by the vector $\boldsymbol{\omega}$, comprising two elements, each of which is zero or one. For an animal detected by observer 1 but not observer 2, $\boldsymbol{\omega} = (1, 0)$, while for an animal detected by observer 2 but not observer 1, $\boldsymbol{\omega} = (0, 1)$. For an animal detected by both observers, $\boldsymbol{\omega} = (1, 1)$. Note that the capture history $\boldsymbol{\omega} = (0, 0)$ is not observed, and so does not appear in the mark-recapture component of the likelihood. We can now write

$$\mathcal{L}_\omega = \prod_{i=1}^n \Pr(\omega_i | \text{detected}) = \prod_{i=1}^n \frac{\Pr(\omega_i)}{p \cdot (y_i, \mathbf{z}_i)}, \quad (20)$$

where

$$\begin{aligned} \Pr(\omega_i = (1, 0)) &= p_1(y, \mathbf{z}) (1 - p_{2|1}(y, \mathbf{z})), \\ \Pr(\omega_i = (0, 1)) &= p_2(y, \mathbf{z}) (1 - p_{1|2}(y, \mathbf{z})), \\ \Pr(\omega_i = (1, 1)) &= p_1(y, \mathbf{z}) p_{2|1}(y, \mathbf{z}) = p_2(y, \mathbf{z}) p_{1|2}(y, \mathbf{z}). \end{aligned}$$

Here, $p \cdot (y_i, \mathbf{z}_i)$ is the probability that an animal at distance y_i from the line or point and with covariates \mathbf{z}_i is detected by at least one observer, and is equivalent to $g(y_i, \mathbf{z}_i)$ of the previous section. Note however that $p \cdot (0, \mathbf{z}_i)$ is not assumed to be one. Further, $p_1(y, \mathbf{z})$ is the unconditional probability that observer 1 detects an animal at distance y and covariates \mathbf{z} , while $p_{1|2}(y, \mathbf{z})$ is the probability of detection, conditional on the animal having been detected by observer 2, with equivalent expressions for observer 2. Models for these probabilities are proposed by [Laake and Borchers \(2004\)](#), [Borchers et al. \(2006\)](#) and [Buckland et al. \(2010\)](#).

The full likelihood is given by $\mathcal{L}_{n,\omega,z,y} = \mathcal{L}_n \times \mathcal{L}_\omega \times \mathcal{L}_z \times \mathcal{L}_{y|z}$. Again, we usually avoid the problem of specifying a suitable model for $\pi_z(\mathbf{z})$ by maximizing the conditional likelihood $\mathcal{L}_\omega \times \mathcal{L}_{y|z}$ and estimating abundance using a Horvitz–Thompson-like estimator.

3. PLOT-BASED MODELS

For designed distance sampling experiments, we wish to test for treatment effects on counts, while accounting for variation in detectability. For this, we need plot-based models.

3.1. PLOT COUNT MODELS

3.1.1. Exact Distance Data

So far, we have ignored the spatial information in our data. Denote the unknown number of animals on plot k ($k = 1, \dots, K$) by N_k , and the number of animals detected on plot k by n_k , where $\sum_{k=1}^K n_k = n$. Here we define plot k to mean the strip of half-width w and length L_k centred on line k (line transect sampling) or the circle of radius w centred on point k (point transect sampling). We denote the area of plot k by a_k , so that $a_k = 2wL_k$ (line transect sampling) or $a_k = \pi w^2$ (point transect sampling).

We consider two models for the plot counts n_k : the multinomial, which involves extending the binomial approach of Sect. 2.1, and the Poisson.

Two new issues arise when we move to plot count models. The first is that we need to decide whether we are interested in inference about total abundance N , as in Sect. 2.1, or whether we only wish to compare densities among the plots, as occurs in a designed distance sampling experiment. In the second case, we wish to restrict inference to the plots, for example to compare a treatment with a control. The second new issue is that we may have covariates at the plot level or higher, but we may also have individual covariates

(i.e. covariates whose values are recorded for each individual detection). If the detection function depends only on plot-level covariates, we can condition on the covariates, and there is no advantage in terms of estimating probability of detection to adding a component to the likelihood for these covariates. However, if the detection function also (or instead) depends on individual covariates, then the detections on a given plot are biased towards those taking covariate values that increase the probability of detection. Thus we need to include a component in the likelihood corresponding to the distribution of individual covariates.

In Sect. 2.1, for model-based conventional distance sampling, we wrote the full likelihood as $\mathcal{L}_{n,y} = \mathcal{L}_n \times \mathcal{L}_y$. We can again take \mathcal{L}_y as given in (1). In place of \mathcal{L}_n , we write $\mathcal{L}_{\{n_k\}}$ where $\{n_k\}$ is the set of plot counts $n_k, k = 1, \dots, K$. Extending the binomial likelihood of (3), we obtain the multinomial model:

$$\mathcal{L}_{\{n_k\}} = \frac{N!}{\prod_{k=1}^K n_k!(N-n)!} \left[1 - \sum_{k=1}^K \alpha_k P_k \right]^{N-n} \prod_{k=1}^K (\alpha_k P_k)^{n_k}, \quad (21)$$

where α_k is the probability that an animal is located on plot k , and P_k is the probability that an animal is detected, given that it is on plot k . (When $K = 1$, (21) reduces to (3).) Under a uniform density model, α_k is simply the area of plot k divided by the total study area. To model how density varies through the study area, we can express α_k as a function of plot covariates \mathbf{x}_k : $\alpha_k \equiv \pi_x(\mathbf{x}_k)$. For model-based conventional distance sampling (with no covariates other than distance in the detection function), $P_k = P_a = \int_0^w g(y)\pi_y(y) dy$, the same for every plot. If the detection function is a function of plot covariates \mathbf{x}_k but not of individual covariates (other than distance y), then

$$P_k = \int_0^w g(y, \mathbf{x}_k)\pi_y(y) dy, \quad (22)$$

and if it is also a function of individual covariates \mathbf{z} (i.e. model-based multiple-covariate distance sampling), then we must specify a model for the probability density function $\pi_z(\mathbf{z})$ of \mathbf{z} in the population, and take the expectation over this density:

$$P_k = \int_0^w \int_{\mathbf{z}} g(y, \mathbf{z}, \mathbf{x}_k)\pi_z(\mathbf{z})\pi_y(y) d\mathbf{z} dy. \quad (23)$$

In the latter case, to complete the full likelihood, instead of taking \mathcal{L}_y , we would multiply $\mathcal{L}_{\{n_k\}}$ by $\mathcal{L}_z \times \mathcal{L}_{y|z}$, where \mathcal{L}_z is from (10) and $\mathcal{L}_{y|z}$ is from (11).

If we wish to restrict inference to the plots, as occurs in designed experiments using distance sampling, then we can replace (21) by

$$\mathcal{L}_{\{n_k\}} = \frac{N_c!}{\prod_{k=1}^K n_k!(N_c-n)!} \left[1 - \sum_{k=1}^K \alpha_k P_k \right]^{N_c-n} \prod_{k=1}^K (\alpha_k P_k)^{n_k}, \quad (24)$$

where $N_c = \sum_{k=1}^K N_k$ is total abundance on the plots.

Poisson models offer a simpler alternative when inference is restricted to the plots:

$$\mathcal{L}_{\{n_k\}} = \prod_{k=1}^K \frac{\lambda_k^{n_k} \exp[-\lambda_k]}{n_k!}, \quad (25)$$

where for model-based conventional distance sampling,

$$\lambda_k = E(n_k) = E(N_k) \times P_a = \exp \left(\sum_{q=1}^Q x_{qk} \beta_q + \log_e(a_k P_a) \right) \quad (26)$$

so that the vector \mathbf{x}_k , with q th element x_{qk} , represents covariates recorded at the plot level.

Equation (26) defines a generalized linear model with log link function and an offset term of $\log_e(a_k P_a)$. The complication is that P_a is an unknown parameter. This suggests a two-stage approach: maximize \mathcal{L}_y , to give an estimate of P_a , and then substitute our estimate \hat{P}_a into the offset, and maximize $\mathcal{L}_{\{n_k\}}$ using standard generalized linear modelling software. [Melville and Welsh \(2014\)](#) adopted this strategy. The method fails to take account of the uncertainty in \hat{P}_a at the second stage. One way to propagate the uncertainty from stage 1 into stage 2 is to use a bootstrap ([Buckland et al. 2009](#)). [Williams et al. \(2011\)](#) described a less computer-intensive and more stable method, but given its complexity, and the fact that the two-stage approach does not in general give maximum likelihood estimates of the parameters, a full likelihood approach, in which P_a is just treated as another parameter to estimate, seems preferable:

$$\mathcal{L}_{\{n_k\},y} = \mathcal{L}_{\{n_k\}} \times \mathcal{L}_y = \prod_{k=1}^K \frac{\lambda_k^{n_k} \exp[-\lambda_k]}{n_k!} \times \prod_{i=1}^n \frac{g(y_i) \pi_y(y_i)}{P_a}. \quad (27)$$

If we have individual covariates \mathbf{z} in the detection function, then the full likelihood is $\mathcal{L}_{\{n_k\},z,y} = \mathcal{L}_{\{n_k\}} \times \mathcal{L}_z \times \mathcal{L}_{y|z}$ where \mathcal{L}_z is given by (10) and $\mathcal{L}_{y|z}$ by (11). Further, P_k , the probability of detection on plot k , now varies by plot, so that our model for plot counts becomes

$$\lambda_k = E(n_k) = E(N_k) \times P_k = \exp \left(\sum_{q=1}^Q x_{qk} \beta_q + \log_e(a_k P_k) \right). \quad (28)$$

When the detection function depends on plot-level covariates but not on individual covariates, then we do not need to specify a distribution for these covariates; instead, we simply form the likelihood conditional on the covariate values, as for the multinomial model.

Note that if we adopt a spatial non-homogeneous Poisson process distance sampling model, we can write

$$\lambda_k = \int_k D(l) g(y(l)) dl, \quad (29)$$

where the integral is over plot k (compare with μ_A of (5)). In this case, for plot k , $\pi_y(y)$ is the integral of $D(l)$ across the sections of plot at distance y from the line or point (two parallel incremental strips for line transect sampling, and an incremental annulus for point transect sampling), divided by the integral across the whole plot. In practice, if plots are

small, we are likely to approximate this by assuming that $\pi_y(y)$ is uniform (line transects) or triangular (point transects). If a fully spatial model is preferred, it would be sensible to record location l_i of detection i , and not just its distance y_i from the line or point. A point process likelihood of the form given by [Hedley and Buckland \(2004\)](#) might then be used.

We can specify models for λ_k of the form of (26), where the covariates \mathbf{x}_k (assumed to be at the plot level or higher) might define the design in the case of a designed distance sampling experiment, or might be spatial covariates for a spatial model, or might simply be any explanatory variables that are potentially useful for modelling animal density. Generally, we would define $x_{1k} = 1$ for all k , so that β_1 is an intercept term. We can also replace linear terms by smooth terms to give greater flexibility (e.g. [Hedley and Buckland 2004](#)). Instead of maximizing the two likelihood components separately, we can maximize the full likelihood, or use Bayesian methods to draw inference on all unknown parameters. Thus the parameter P_k in the offset, which for the two-stage approach was estimated in stage 1 then treated as known in stage 2, is now a function of the detection function parameters (below), and estimated along with all other parameters in a single step. [Oedekoven et al. \(2014\)](#) proposed the above approach, with the inclusion of a random effect for location in the model for λ_k (see below).

If counts are summed across repeat visits to a plot, the offset term is multiplied by the effort, where effort is defined to be the number of repeat visits; and if counts are summed across replicate plots, plot size a_k is the combined size of the plots whose counts have been combined.

The product $a_k P_k$ is the effective area surveyed on plot k . The P_k are defined exactly as for the multinomial models.

Note that neither population size N nor plot abundances N_k appear as parameters in the Poisson likelihood. For designed distance sampling experiments, we only wish to compare densities in the plots, and have no interest in estimating N in a wider study area. However, with this approach, we can still draw inference on abundance. For spatial distance sampling models, we can predict density throughout the study area, and so can use numerical integration under the fitted density surface to estimate abundance either for the full study area or for any subset of it. We can also estimate plot abundance N_k by $\hat{N}_k = \hat{\lambda}_k / \hat{P}_k$. By contrast, the multinomial model allows direct inference for both population size N and plot abundances $N_k = N\pi_x(\mathbf{x}_k)$, and, as with the Poisson model, the effect of the covariates \mathbf{x} on abundance or density can be investigated through the parameters of the model for $\pi_x(\mathbf{x})$.

3.1.2. Grouped Distance Data

For the case without covariates, let m_{jk} be the number of detections in distance interval j on plot k , with $\sum_{j=1}^u m_{jk} = n_k$. Adopting a Poisson model for these counts, and given $E(n_k) = \lambda_k$, then $E(m_{jk}) = \lambda_k f_j$ for $j = 1, \dots, J$, where f_j is given in (7).

We can now write the full likelihood as

$$\prod_{k=1}^K \prod_{j=1}^J \frac{(\lambda_k f_j)^{m_{jk}} \exp(-\lambda_k f_j)}{m_{jk}!}. \quad (30)$$

For detection function covariates \mathbf{z} recorded at the plot level, or at the stratum level if the design is stratified (but not at the individual level), we can define

$$f_{jk} = \int_{c_{j-1}}^{c_j} f_{y|z}(y|\mathbf{z}_k) dy = \frac{\int_{c_{j-1}}^{c_j} g(y, \mathbf{z}_k) \pi_y(y) dy}{P_a(\mathbf{z}_k)} \quad (31)$$

giving the full likelihood

$$\prod_{k=1}^K \prod_{j=1}^J \frac{(\lambda_k f_{jk})^{m_{jk}} \exp(-\lambda_k f_{jk})}{m_{jk}!}. \quad (32)$$

Note that when using the Poisson model for the expected abundances for this approach, we do not have separate components for $\mathcal{L}_{\{n_k\}}$ and \mathcal{L}_m . Oedekoven et al. (2014) adopted a different strategy which is essentially the grouped data equivalent of (27), and which does have separate components for $\mathcal{L}_{\{n_k\}}$ (using a generalized version of (25)) and \mathcal{L}_m (using (6)).

3.2. PLOT ABUNDANCE MODELS

Royle et al. (2004) adopted what appears to be a different strategy for grouped distance data. Again m_{jk} is the count of detected animals in distance interval j on plot k , with $\sum_{j=1}^J m_{jk} = n_k$. We define the proportion P_j of plot abundance that was observed within distance interval j :

$$P_j = \int_{c_{j-1}}^{c_j} g(y) \pi_y(y) dy, \quad (33)$$

where $g(y)$ and $\pi_y(y)$ represent the detection function and the distribution of distances from the line or point in the population as before. The sum of proportions P_j over all J distance intervals gives the average detection probability P_a , i.e. $\sum_{j=1}^J P_j = P_a$, where $P_a = \int_0^w g(y) \pi_y(y) dy$ (2).

The P_j represent the proportion of plot abundance N_k that was both located in distance interval j and detected, while the f_j represent the proportion of detected animals n_k that were located in distance interval j . Hence we have the relationship $f_j = P_j/P_a$.

As we do not observe the true abundances on the plot, we set $E(N_k) = \kappa_k$ and model these using a log-linear Poisson model:

$$\kappa_k = \exp \left(\sum_{q=1}^Q x_{qk} \beta_q \right). \quad (34)$$

The observed counts in distance interval j are then modelled as a Poisson random variable, $m_{jk} \sim \text{Poisson}(\kappa_k \times P_j)$. The likelihood for this model is

$$\mathcal{L}_{\{n_k\},m} = \prod_{k=1}^K \prod_{j=1}^J \frac{(\kappa_k P_j)^{m_{jk}} \exp(-\kappa_k P_j)}{m_{jk}!}. \quad (35)$$

By noting that $\lambda_k = \kappa_k P_a$, we see that the above model for κ_k is equivalent to our model for λ_k , provided plot sizes are all the same, and arbitrarily set as $a_k = 1$:

$$\lambda_k = \kappa_k P_a = \exp\left(\sum_{q=1}^Q x_{qk} \beta_q + \log(P_a)\right). \quad (36)$$

Thus the Poisson rate corresponding to count m_{jk} is $\lambda_k f_j = \lambda_k P_j / P_a = \kappa_k P_j$, so that the likelihood of (35) is equivalent to that of (30). Here, we use (30), as it allows plot area a_k to vary.

One of the approaches of [Hedley and Buckland \(2004\)](#) combines a plot abundance model with a two-stage modelling strategy: a Horvitz–Thompson-like estimator is used to estimate abundance N_k on plot k , and these estimates \hat{N}_k are taken as the responses for a spatial model. When individual covariates are recorded, this offers a simpler, if conceptually less appealing, approach to the use of (23) in a plot count model.

4. ADDING RANDOM EFFECTS

We might wish to add random effects to a distance sampling model for several reasons. For example, if there are multiple lines or points within a plot or site, then a site random effect allows for spatial correlation in the observations; similarly, if there are repeat counts at any given location, we might allow for temporal correlation using random effects ([Oedekoven et al. 2013, 2014](#)). A further reason to consider random effects is if there is heterogeneity in the detection probabilities that is not modelled by the available covariates ([Oedekoven et al. 2015](#)).

4.1. RANDOM EFFECTS IN THE COUNT MODEL

If there are repeat visits to plots, for some purposes, we can just pool data from the repeat visits, but sometimes we may wish to model the separate plot counts. We can then define plot to be a random effect, which allows for correlation between repeat counts on a given plot. Our model for expected plot count (26) is readily extended:

$$\lambda_{kt} = \exp\left(\sum_{q=1}^Q x_{qkt} \beta_q + b_k + \log_e(a_k P_{kt})\right), \quad (37)$$

where the subscript t indicates visit t to plot k . (If there are no time-varying covariates, we can drop the t subscript from this expression.) Typically, the random effects are assumed to be normally distributed:

$$b_k \sim N(0, \sigma_b^2). \quad (38)$$

The likelihood for the count model now includes a normal density for the random effects (Oedekoven et al. 2014) and is given by

$$\mathcal{L}_{\{n_{kt}\}} = \prod_{k=1}^K \int_{-\infty}^{\infty} \left\{ \prod_{t=1}^{T_k} \frac{\lambda_{kt}^{n_{kt}} \exp[-\lambda_{kt}]}{n_{kt}!} \right\} \times \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left[-\frac{b_k^2}{2\sigma_b^2}\right] db_k, \quad (39)$$

where T_k represents the total number of visits to plot k .

If a sampling location comprises more than one line or point, again for some purposes, we can just pool the data for the location. However, we can model the separate plot counts by introducing a location random effect, to allow for correlation across plots within a single location:

$$\lambda_{kl} = \exp\left(\sum_{q=1}^Q x_{qkl}\beta_q + b_l + \log_e(a_{kl}P_{kl})\right), \quad (40)$$

where l indicates location, and $b_l \sim N(0, \sigma_l^2)$. The likelihood for the counts is now

$$\mathcal{L}_{\{n_{kl}\}} = \prod_{l=1}^L \int_{-\infty}^{\infty} \left\{ \prod_{k=1}^K \frac{\lambda_{kl}^{n_{kl}} \exp[-\lambda_{kl}]}{n_{kl}!} \right\} \times \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\left[-\frac{b_l^2}{2\sigma_l^2}\right] db_l, \quad (41)$$

where L is the total number of locations.

We could also (or instead) define a coefficient β_q to be a location random effect, which would mean that the effect of the corresponding covariate x_q varies by location.

4.2. INDIVIDUAL RANDOM EFFECTS

Random effects can also be included in the model for the detection function, to model any heterogeneity not accounted for by any covariates \mathbf{z} included in the model (Oedekoven et al. 2015). For example if we have a multiple-covariate distance sampling model, the model for the scale parameter σ , given in (8), may be extended as

$$\sigma(\mathbf{z}_i) = \exp\left(t_i + \sum_{q=1}^Q \beta_q z_{qi}\right), \quad (42)$$

where $t_i \sim N(0, \sigma_t^2)$. The corresponding likelihood for line transect sampling is given by

$$\mathcal{L}_{y|\mathbf{z}}(\boldsymbol{\beta}, \sigma_t|\mathbf{z}) = \prod_{i=1}^n \frac{\int_{-\infty}^{\infty} g(y_i|\mathbf{z}, t_i) N(t_i, 0, \sigma_t) dt_i}{\int_{-\infty}^{\infty} \int_0^w g(u|\mathbf{z}, t_i) du N(t_i, 0, \sigma_t) dt_i} \quad (43)$$

where $N(t_i, 0, \sigma_t) = \exp\left[-0.5\left(\frac{t_i}{\sigma_t}\right)^2\right] \left(\sqrt{2\pi}\sigma_t\right)^{-1}$ (Oedekoven et al. 2015). Again, we could have coefficient β_q as a random effect, so that the effect of the corresponding z_q varies by individual detection.

The general framework for modelling measurement error of [Borchers et al. \(2010\)](#) may be regarded as modelling individual random effects. In the above formulation, the response y is observed and the random effect unobserved, whereas for measurement error models, the response y is not observed; instead, we observe a version of it contaminated by measurement error, say v . We can then write

$$f_{v|z}(v|\mathbf{z}) = \int_0^\infty \pi_{\text{err}}(v|y, \mathbf{z}) f_{y|z}(y|\mathbf{z}) \, dy, \quad (44)$$

where $\pi_{\text{err}}(v|y, \mathbf{z})$ is an error model, specified as a probability density function of v given y and \mathbf{z} , and $f_{y|z}(y|\mathbf{z})$ is the probability density function of the true distances y , conditional on covariates \mathbf{z} .

We need additional information to estimate the parameters of the measurement error model. [Borchers et al. \(2010\)](#) show how double-observer survey data (Sect. 2.3) may be used, if further assumptions are made. They also consider the case that we are able to take m further measurements, for which we can record both true distances y and distances with error v (together with any covariates \mathbf{z}). In this case, if we index the observations from the main survey by $i = 1, \dots, n$, and the additional observations by $i = n + 1, \dots, n + m$, then the likelihood $\mathcal{L}_{y|z}$ of (11) is replaced by

$$\mathcal{L}_{v|z} \times \mathcal{L}_{\text{err}} = \prod_{i=1}^n f_{v|z}(v_i|\mathbf{z}_i) \times \prod_{i=n+1}^{n+m} \pi_{\text{err}}(v_i|y_i, \mathbf{z}_i). \quad (45)$$

5. CASE STUDY

To illustrate plot-based methods, we consider a point transect experiment to assess whether conservation buffers along field margins increased densities of northern bobwhite quail coveys in the United States ([Oedekoven et al. 2014](#)). A matched pairs design was adopted, with two points within each site, one in a conservation buffer ($type=1$), and the other at the edge of a nearby field with the same crop but no buffer ($type=2$). We analysed data from 183 sites located in four states (MO: Missouri, MS: Mississippi, NC: North Carolina, TN: Tennessee)—a reduced dataset compared to [Oedekoven et al.](#)'s study. Repeat surveys were conducted in three years (2006–2008), resulting in 1023 covey detections from 1051 visits to points. Distance data were assumed exact and were truncated at 500 m.

We adopted a maximum likelihood approach where we used the likelihood $\mathcal{L}_{y, \{n_{ki}\}} = \mathcal{L}_{y|z} \times \mathcal{L}_{\{n_{ki}\}}$ (and so omitting \mathcal{L}_z from the full likelihood). For $\mathcal{L}_{y|z}$ we assumed a multiple-covariate distance sampling likelihood (11) with a hazard-rate detection function as the hazard-rate provided a much better fit to the observed distances compared to the half-normal. The count likelihood $\mathcal{L}_{\{n_{ki}\}}$ was similar to (41), but extended to take account of the design with multiple points at each site, and repeat visits to the points. We included the covariate $type$ in the detection model and covariates $type$, $state$ and $Julian\ date$ (centred around its mean) in the count model. The same data were analysed using the Bayesian approach of [Oedekoven et al. \(2014\)](#). The first 9,999 of the 100,000 iterations were considered as burn-in and excluded from the summary statistics.

Table 1. Parameter estimates and standard errors for count model parameters obtained with the maximum likelihood approach (MLE and SE), together with the corresponding posterior means and standard deviations (SD) from the Bayesian approach.

Parameter	Maximum likelihood		Bayesian posterior	
	MLE	SE	Mean	SD
Random effect std dev.	0.71	0.06	0.73	0.07
Intercept	-13.15	0.13	-13.15	0.14
Type	0.68	0.10	0.67	0.12
State MS	-0.43	0.17	-0.44	0.18
State NC	-1.46	0.18	-1.47	0.19
State TN	-1.30	0.19	-1.31	0.20
Julian date	-0.023	0.0023	-0.023	0.0039

Maximum likelihood estimates of count model parameters and their standard errors were similar to the corresponding posterior means and standard deviations from the Bayesian approach (Table 1). The parameter of interest was the *type* covariate. The maximum likelihood estimate for this parameter was 0.68 (SE = 0.10) indicating a 97 % increase in covey densities where conservation buffers were present relative to where they were not. This is very close to the Bayesian estimate: the posterior mean for the *type* covariate was 0.67 (SD = 0.12), corresponding to an estimated 95 % increase. (Oedekoven et al. 2014, analysed the full dataset using the Bayesian approach, and estimated that densities were 85 % higher where buffers were present.)

6. DISCUSSION

6.1. OTHER EXAMPLES

Our case study illustrates some of the advantages in adopting a model-based approach. Here, we briefly summarize further examples that illustrate different aspects of the general approach outlined above. We select examples for which there was a clear advantage in adopting a model-based approach.

Buckland et al. (2009) analysed data from a point transect before–after control–impact experiment to assess whether prescribed fire treatments in ponderosa pine forests in the southwestern United States affected densities of two species of warbler. A plot count model was adopted assuming exact distance (Sect. 3.1), with an offset that was a function of detectability. A two-stage method was implemented in which detectability was estimated in stage 1, and counts were modelled in stage 2, conditional on estimated detectability. Uncertainty in estimated detectability was propagated to stage 2 using a bootstrap. A significant interaction was detected between treatment (control or burning) and year, indicating strong evidence of reduced warbler densities in the year after burning, with moderate evidence of continuing lower densities a year later.

Oedekoven et al. (2013) considered a point transect experiment, to assess whether conservation buffers along field margins increased densities of indigo buntings in the United

States. Many of the sampled sites were in common between the bunting survey and the bobwhite survey of our case study, although the bunting survey was carried out in the breeding season, while the bobwhite survey was conducted in the fall. Further, detected buntings were assigned to distance intervals, while exact distances were measured for the bobwhites. For the buntings, a plot abundance model for interval distance data was adopted, which is equivalent to a plot count model (Sect. 3.2). Thus the approach of Sect. 3.1 was adopted, and maximum likelihood methods were used to fit the model. The detection function was stratified by *state* and the count model included covariates together with a random effect for site. The treatment effect was highly significant, with densities 35 % higher on buffered fields than on control fields.

In this paper, we only superficially address full spatial models for distance sampling. Key papers in this area are Högmander (1991), Hedley and Buckland (2004) and Johnson et al. (2010). Miller et al. (2013) explored how the density of pantropical spotted dolphins varies through a study region in the Gulf of Mexico. Shipboard line transect surveys were carried out, during which sightings and environmental covariates were recorded. In an online appendix, they provide a worked example of how to fit a density surface, based on the methods of Hedley and Buckland (2004). In the case of the spotted dolphins, they found that densities were very variable, and they were able to identify ‘hotspots’.

6.2. GENERAL DISCUSSION

Standard distance sampling methods are hybrid methods in that they use a model-based approach for modelling the detection function, but a design-based approach both to ensure that animal locations are independent of line or point locations and to extrapolate density on the surveyed plots to the wider study region. When the sole objective of a distance sampling survey is to estimate abundance (or equivalently, mean density) in the study region, this strategy is remarkably robust, if somewhat unsatisfactory conceptually. Fully model-based methods allow a more coherent approach, at the risk of sensitivity to choice of model. More importantly, model-based methods exploit the data more effectively, to answer questions other than how many animals occupy the study region. For example, designed distance sampling experiments are starting to come into use, and model-based methods allow formal assessment of treatment effects. Further, spatial distance-sampling models allow animal density to be modelled as a function of habitat, topography and/or environment, and by numerically integrating under the corresponding section of the fitted density surface, animal abundance may be estimated for any sub-region of interest. Spatial models also provide a means to address non-uniform distribution in the vicinity of samplers as a result of responsive movement or non-random sampler placement, by adding distance from sampler as a covariate in the density model. Marques et al. (2010, 2013) address non-uniform distribution, but do not incorporate a full density model. Instead, Marques et al. (2010) model the density surface around the points for point transect sampling along features, while for line transect sampling, Marques et al. (2013) model the distribution of animals as a function of distance from the line.

We expect to see rapid increase in the use of fully model-based distance sampling methods, given their potential for widening the applicability of distance sampling.

ACKNOWLEDGEMENTS

CSO was part-funded by EPSRC/NERC Grant EP/1000917/1. The national CP33 monitoring program was coordinated and delivered by the Department of Wildlife, Fisheries, and Aquaculture and the Forest and Wildlife Research Center, Mississippi State University. The national CP33 monitoring program was funded by the Multistate Conservation Grant Program (Grants MS M-1-T, MS M-2-R), a program supported with funds from the Wildlife and Sport Fish Restoration Program and jointly managed by the Association of Fish and Wildlife Agencies, U.S. Fish and Wildlife Service, USDA-Farm Service Agency and USDA-Natural Resources Conservation Service-Conservation Effects Assessment Project. We thank two anonymous reviewers, whose comments have led to a much improved paper.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

[Received February 2015. Accepted August 2015.]

REFERENCES

- Borchers DL, Buckland ST, Zucchini W (2002) Estimating Animal Abundance: Closed Populations. Springer Verlag, London.
- Borchers DL, Burnham KP (2004) General formulation for distance sampling. In: Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (eds) Advanced Distance Sampling. Oxford University Press, Oxford, pp 6–30
- Borchers D, Laake J, Southwell C, Paxton C (2006) Accommodating unmodeled heterogeneity in double-observer distance sampling surveys. *Biometrics* 62:372–378
- Borchers D, Marques T, Gunnlaugsson T, Jupp P (2010) Estimating distance sampling detection functions when distances are measured with errors. *Journal of Agricultural, Biological, and Environmental Statistics* 15:346–361
- Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (2001) Introduction to Distance Sampling: Estimating Abundance of Biological Populations. Oxford University Press, Oxford.
- Buckland ST, Laake JL, Borchers DL (2010) Double-observer line transect methods: levels of independence. *Biometrics* 66:169–177
- Buckland ST, Russell RE, Dickson BG, Saab VA, Gorman DN, Block WM (2009) Analysing designed experiments in distance sampling. *Journal of Agricultural, Biological, and Environmental Statistics* 14:432–442
- Fewster RM, Buckland ST (2004) Assessment of distance sampling estimators. In: Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (eds) Advanced Distance Sampling. Oxford University Press, Oxford, pp 281–306
- Hedley SL, Buckland ST (2004) Spatial models for line transect sampling. *Journal of Agricultural, Biological, and Environmental Statistics* 9:181–199
- Högmander H (1991) A random fields approach to transect counts of wildlife populations. *Biometrical Journal* 33:1013–1023
- Johnson DS, Laake JL, Ver Hoef JM (2010) A model-based approach for making ecological inference from distance sampling data. *Biometrics* 66:310–318
- Laake JL, Borchers DL (2004) Methods for incomplete detection at distance zero. In: Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (eds) Advanced Distance Sampling. Oxford University Press, Oxford, pp 108–189
- Marques FFC, Buckland ST (2003) Incorporating covariates into standard line transect analyses. *Biometrics* 59:924–935

- Marques TA, Buckland ST, Borchers DL, Tosh D, McDonald RA (2010) Point transect sampling along linear features. *Biometrics* 66:1247–1255
- Marques TA, Buckland ST, Bispo R, Howland B (2013) Accounting for animal density gradients using independent information in distance sampling surveys. *Statistical Methods and Applications* 22:67–80
- Melville GJ, Welsh AH (2014) Model-based prediction in ecological surveys including those with incomplete detection. *Australian and New Zealand Journal of Statistics* 56:257–281
- Miller DL, Burt ML, Rexstad EA, Thomas L (2013) Spatial models for distance sampling data: recent developments and future directions. *Methods in Ecology and Evolution* 4:1001–1010
- Oedekoven CS, Buckland ST, Mackenzie ML, Evans KO, Burger LW (2013) Improving distance sampling: accounting for covariates and non-independency between sampled sites. *Journal of Applied Ecology* 50:786–793
- Oedekoven CS, Buckland ST, Mackenzie ML, King R, Evans KO, Burger LW (2014) Bayesian methods for hierarchical distance sampling models. *Journal of Agricultural, Biological, and Environmental Statistics* 19:219–239
- Oedekoven CS, Laake JL, Skaug HJ (2015) Distance sampling with a random scale detection function. *Environmental and Ecological Statistics*. doi:[10.1007/s10651-015-0316-9](https://doi.org/10.1007/s10651-015-0316-9)
- Royle JA, Dawson DK, Bates S (2004) Modeling abundance effects in distance sampling. *Ecology* 85:1591–1597
- Royle JA, Dorazio RM (2008) *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press, San Diego.
- Stoyen D (1982) A remark on the line transect method. *Biometrical Journal* 24:191–195
- Williams R, Hedley SL, Branch TA, Bravington MA, Zerbini AN, Findlay KP (2011) Chilean blue whales as a case study to illustrate methods to estimate abundance and evaluate conservation status of rare species. *Conservation Biology* 25:526–535