Spatial Models for Line Transect Sampling

Sharon L. HEDLEY and Stephen T. BUCKLAND

This article develops methods for fitting spatial models to line transect data. These allow animal density to be related to topographical, environmental, habitat, and other spatial variables, helping wildlife managers to identify the factors that affect abundance. They also enable estimation of abundance for any subarea of interest within the surveyed region, and potentially yield estimates of abundance from sightings surveys for which the survey design could not be randomized, such as surveys conducted from platforms of opportunity. The methods are illustrated through analyses of data from a shipboard sightings survey of minke whales in the Antarctic.

Key Words: Distance sampling; Generalized additive model; Generalized linear model; Population size estimation; Sightings survey; Spatial distribution.

1. INTRODUCTION

Line transect sampling (Buckland et al. 2001) is one of the most widely used techniques for estimating the size of wildlife populations. Increasingly, wildlife managers wish to extract more than just an abundance estimate from their sightings surveys. They frequently need to relate animal density to spatial variables reflecting topography, habitat, and other factors that affect the animals' environment. This aids assessment of how to manage that environment and the animals within it. There is therefore a strong demand for spatial models for analyzing line transect data.

Spatial line transect models allow wildlife managers to estimate abundance for any subset of a survey region, by numerically integrating under the relevant section of the fitted density surface. In contrast, conventional line transect methods restrict estimation of abundance to a set of predefined survey blocks, defined at the design stage of a survey using stratified random or stratified systematic sampling. A spatial model for density allows total abundance to be proportioned between any set of subregions of interest, and is less

Sharon L. Hedley is Research Fellow, Research Unit for Wildlife Population Assessment, University of St. Andrews, The Observatory, Buchanan Gardens, St. Andrews KY16 9LZ, Scotland (E-mail: sharon@mcs.st-and.ac.uk). Stephen T. Buckland is Director, Centre for Reseach into Ecological and Environmental Modelling, University of St. Andrews, The Observatory, Buchanan Gardens, St. Andrews KY16 9LZ, Scotland (E-mail: steve@mcs.st-and.ac.uk).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 9, Number 2, Pages 181–199
DOI: 10.1198/1085711043578

susceptible to the problem of small sample sizes in subregions than are stratified schemes. However, if we select a poor spatial model, estimates by subregion may have large bias, especially close to the edge of the survey region.

Adopting a model-based approach for density estimation represents a departure from traditional line transect methodology, for which unbiased estimation relies on a survey design with randomized trackline locations. Although it is desirable to have a set of tracklines with sufficient spatial spread to provide representative coverage of the survey area, spatial modeling does not require that the tracklines are designed according to a formal survey sampling scheme. If the spatial coverage of the nonrandom data is adequate, spatial models can yield estimates of abundance from sightings data collected from "platforms of opportunity" (ferries, merchant navy vessels, oceanographic survey vessels, etc.). Because it is often prohibitively expensive to hire a ship, many marine surveys are conducted from these types of platforms, collecting large quantities of nonrandom data at a cost substantially lower than that of a properly designed survey.

The main disadvantage of a model-based analysis is that we risk bias in abundance estimation due to model mis-specification. A second disadvantage is that more sophisticated analysis methods are needed than for a design-based analysis.

Our motivating example is of minke whales (*Balaenoptera bonaerensis*) in the Southern Ocean. This species has a circumpolar distribution in Antarctic waters, and in the 1970s and 1980s was the main target of the Antarctic whaling fleets (Jefferson, Leatherwood, and Webber 1993). Obtaining independent assessment information was therefore important for the International Whaling Commission (IWC), and so in 1978–1979, a systematic survey program was established—the International Decade of Cetacean Research (IDCR)—to estimate abundance of this species. For the first seven years, the surveys used both line transect and mark-recapture methods, but the latter were abandoned due to uncertainty about the rate of shedding of marks. Line transect surveys have continued annually since then (now under the acronym SOWER: Southern Ocean Whale and Ecosystem Research), reflecting the continuing importance of the work despite the moratorium on commercial whaling. In particular, as part of a long-term goal of assessing the impact of climate change on Southern Ocean whale populations, there is interest in using these surveys to relate minke whale densities to features of their environment.

2. POINT PROCESS LINE TRANSECT FORMULATION

The conventional line transect estimator of density, $\hat{D}$, when animals occur singly is

$$\hat{D} = \frac{n}{2L\hat{\mu}_w}, \tag{2.1}$$

where n is the number of detections, L is the total transect length, and $\hat{\mu}_w$ is the estimated effective strip half-width. This is also the estimator of group density when animals occur in well-defined groups. The methodology described in this article is appropriate for estimation of animal density for species which occur as individuals, or for group density when animals occur in groups. For clarity, however, we shall consider only the former case in the general

description of the methodology, and then in Section 5 we note how to apply these methods to animals that are frequently encountered in groups.

Point process formulations of line transect estimators were considered by Stoyan (1982) and Högmander (1991, 1995). The conventional line transect estimator given by Equation (2.1) was reformulated by Stoyan (1982) in terms of the intensity of a stationary marked point process, where the points were the locations of the animals, and individual points were marked by some probability representing the detectability at each point, based on its sighting distance. Stationarity assumes that animals are uniformly distributed throughout the region of interest, but conventional line transect estimation does not require this assumption provided that the transect lines are placed at random with respect to the locations of the animals. Although this stationarity assumption simplifies the point process formulation substantially, it is not realistic, since animals (due to habitat, environmental, or prey preferences) typically occur in clustered aggregations. Therefore we attempt to address the problem of estimating density when this is modeled as a function of spatial location, rather than as the average intensity of a spatial point process.

2.1 DERIVING A LIKELIHOOD

Let the function $D(u, v)$ represent the density of animals at location (u, v), using a Cartesian coordinate system. Animals' locations are modeled as an inhomogeneous Poisson process, with rate $D(u, v)$.

An important consequence of modeling the animals' locations using the inhomogeneous Poisson process is that if $A_1, \ldots, A_k$ are arbitrary disjoint sets within the survey area, A, then the numbers of animal groups in each set, $N(A_j),\ j = 1, \ldots, k$, are independent Poisson variables, with the following expected values

$$\mathrm{E}\left[N(A_j)\right] = \int_{A_j} D(u, v). \tag{2.2}$$

Suppose the areas A_j are k nonoverlapping strips (of width $2w$ and length $b_j,\ j = 1, \ldots, k$, with $\sum_j b_j = L$) from a line transect survey, where w is the perpendicular truncation distance (assumed, for simplicity, to be symmetric about the transect line). Without loss of generality, we switch to a different, locally defined Cartesian coordinate system (x, y) for strip j, in which x is distance along the line, with $d_j \le x \le d_j + b_j$, and y is distance from the line, with $-w \le y \le w$. We set $d_j = \sum_{j'=1}^{j-1} b_{j'}$ for $j = 2, \ldots, k$, with $d_1 = 0$. Let $g(x, y)$ be the probability of detecting an animal, given that it is located at (x, y).

The statistical problem that we wish to address is to estimate the spatial density surface, $D(x, y)$, throughout the survey region, A, from the observed data—the locations of detections with respect to the transect lines. Therefore, we now consider only the *detected* animals from a line transect survey. Conventional line transect estimation requires estimation of a detection function, $g(y),\ \ 0 \le g(y) \le 1$, the probability of detecting an object, given that it is at a distance y from the line (Buckland et al. 2001). By analogy, we define $g(x, y)$ as the probability of detecting an animal, given that it is located at (x, y).

Assuming that animals are located according to an inhomogeneous Poisson process of rate $D(x, y),\ x, y \in A$, the expected number of animals in an area A_j was given in Equation (2.2). Given any rectangular strip of length l and width $2w$ centered lengthways along a transect line, and starting at $x = x_i$, the expected number of animals within it is

$$\int_{-w}^{w} \int_{x_i}^{x_i+l} D(x, y) \mathrm{d}x \mathrm{d}y \,. \tag{2.3}$$

The expected number of detections within the strip is a function of how many animals are actually present in the strip and of how detectable they are. Under the additional assumption that the detection process is independent of the density of animals, the detection locations represent an independent thinning of the inhomogeneous Poisson process given by all of the animals' locations (i.e., both those that are detected and those that remain undetected). It can be shown that the resultant process from a thinning of this type is also an inhomogeneous Poisson process (Cressie 1991, p. 625–626). Thus, the locations of detected animals within $\pm w$ of the trackline may be described by an inhomogeneous Poisson process with rate $D(x, y)g(x, y)$.

Writing $\mu(A)$ as $\int_A D(x, y)g(x, y)\mathrm{d}x\mathrm{d}y$ (where $\int_A$ denotes that the integral is taken over the entire survey region), then the joint density of the number of detections and their locations is

$$f((x_1, y_1), \ldots, (x_n, y_n), n) = \begin{cases} \exp\left[-\mu(A)\right] \,, & n = 0 \,, \\ \exp\left[-\mu(A)\right] \prod_{i=1}^{n} D(x_i, y_i)g(x_i, y_i)/n! \,, & n = 1, 2, \ldots \end{cases} \tag{2.4}$$

(e.g., Cressie 1991, p. 651). The parameter of interest is $D(x, y)$, not the product $D(x, y)$ $g(x, y)$. An ability to identify this parameter from observations on the thinned process is essential but not straightforward. The assumption that $g(x, 0) = 1$ (certain detection on the line) helps, but is not sufficient. The scales over which the two functions D and g operate tend to be very different. D varies along the line, but typically varies little within a strip in the y-direction, since strip width is typically narrow. By contrast, g varies largely in the y-direction rather than the x-direction. We exploit this to separate estimation of the two functions in the next subsection.

2.2 A LIKELIHOOD BASED ON INTERDETECTION DISTANCES

We define the "waiting distance" as the along-transect distance surveyed between two successive detections. Hence $l_i = x_i - x_{i-1}$ are the along-trackline distances between successive detections, with $i = 2, \ldots, n$. If we denote the x-value at the start of survey effort as $x_0 = 0$ and the x-value corresponding to the end of survey effort as x_{n+1}, then l_1 is the distance surveyed until the first detection and l_{n+1} is the distance surveyed after the last detection.

We make the following assumptions.

1. $D(x, y) \equiv D(x)$ independent of y;
2. $g(x, y) \equiv g(y)$ independent of x;
3. $g(0) = 1$;
4. $g(-y) = g(y)$ for $0 \leq y \leq w$; and
5. x and y are independent.

Noting that l_{n+1} is a right-censored observation, we can now write the joint density of observations as

$$f(l_1, \ldots, l_{n+1}, y_1, \ldots, y_n) = \left[\prod_{i=1}^{n} f_{l|x}(l_i|x_{i-1})\right] \times \mathrm{P}(L_{n+1} > l_{n+1}) \times \left[\prod_{i=1}^{n} f_y(y_i)\right]. \tag{2.5}$$

Conventional line transect theory yields $f_y(y_i) = g(y_i)/2\mu_w$, $-w \leq y_i \leq w$, where $\mu_w = \int_0^w g(y)\mathrm{d}y$. (Usually, the distribution is folded, so that $0 \leq y \leq w$; see Buckland et al. 2001, p. 60.)

We therefore need $f_{l|x}(l_i|x_{i-1})$ for $i = 1, \ldots, n+1$. The cumulative distribution function (CDF) of L_i is

$$\mathrm{F}_{l|x}(l_i|x_{i-1}) = \mathrm{P}(L_i \leq l_i|x_{i-1}) = 1 - \mathrm{P}(L_i > l_i|x_{i-1}), \qquad i = 1, \ldots, n+1. \tag{2.6}$$

Because $L_i > l_i$ if and only if there were no detections in the strip of half-width w, then

$$\begin{aligned} \mathrm{F}_{l|x}(l_i|x_{i-1}) &= 1 - \exp\left\{-\int_{-w}^{w}\int_{x_{i-1}}^{x_{i-1}+l_i} D(x)g(y)\mathrm{d}x\mathrm{d}y\right\} \\ &= 1 - \exp\left\{-2\mu_w \int_{x_{i-1}}^{x_{i-1}+l_i} D(x)\mathrm{d}x\right\}. \end{aligned} \tag{2.7}$$

Differentiating, we obtain the conditional probability density function of the waiting distance l_i as

$$f_{l|x}(l_i|x_{i-1}) = 2\mu_w D(x_{i-1} + l_i)\ \exp\left[-2\mu_w \int_{x_{i-1}}^{x_{i-1}+l_i} D(x)\mathrm{d}x\right], \quad i = 1, \ldots, n \tag{2.8}$$

Note that uncertainty in n is handled implicitly in this formulation. If a sequence of l_i is generated according to an inhomogeneous Poisson process, n is determined as the largest i for which $\sum_{i'=1}^{i} l_{i'} \leq L$, the total line length.

Suppose we have a vector of spatial parameters $\boldsymbol{\theta}$ for the density surface $D(x)$ and parameters $\boldsymbol{\beta}$ for the detection function $g(y)$. Then if we write x_i for $x_{i-1} + l_i$ and with appropriate substitution into Equation (2.5), we obtain the marginal likelihood $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{l}, \mathbf{y})$:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{l}, \mathbf{y}) = \left[\prod_{i=1}^{n} D(x_i)\right] \exp\left[-2\mu_w \sum_{i=1}^{n+1}\int_{x_{i-1}}^{x_i} D(x)\mathrm{d}x\right] \left[\prod_{i=1}^{n} g(y_i)\right], \tag{2.9}$$

where $\mathbf{l}$ is the vector of observed waiting distances $l_i = x_i - x_{i-1}$, $i = 1, \ldots, n+1$, $\mathbf{y}$ is the vector of observed perpendicular distances y_i, $i = 1, \ldots, n$, and n is the number of detections.

Variation in detectability with x can be modeled by introducing a vector of covariates $\mathbf{z}$, so that the detection function is $g(y, \mathbf{z})$ (e.g., Marques 2001).

3. SPATIAL LINE TRANSECT MODELS WHEN DIRECT LIKELIHOOD MAXIMIZATION IS INFEASIBLE

Although direct likelihood maximization allows us to estimate the parameters of some functional forms, convergence difficulties are increasingly likely if more complex functions are required. In this section, we describe two alternative approaches that enable spatial variation in density to be modeled using the suite of potentially complex and flexible functions available in standard generalized linear modeling (GLM) or generalized additive modeling (GAM) software. The first uses waiting distance data and is a simplification of the likelihood-based approach described in the previous section; the second is a simple count model, which we have found to be a practical option for some datasets.

3.1 A FORMULATION FOR WAITING DISTANCES

If animal detections occurred randomly along the transect lines (i.e., according to a homogeneous Poisson process), then the distances between detections would be distributed exponentially, with rate equal to the reciprocal of the average density of detections. In the case of clustered data, where the rate of encounters varies with spatial location, an inhomogeneous Poisson process is likely to provide a better representation of the underlying distribution of detection locations, but the distances between detections would no longer be exponentially distributed. A GLM/GAM may be formulated for a situation that is intermediate between a fully spatially inhomogeneous Poisson process and a homogeneous Poisson process. By assuming that the density of animals and the expected encounter rate are constant in the interval between two successive detection locations, but may change at each location, then each waiting distance can be modeled using an exponential distribution, with rate proportional to the reciprocal of the fitted density in that interval. The GAM formulation for waiting distances is then

$$g[\mathrm{E}(l_i)] = \theta_0 + \sum_k f_k(z_{ik}), \quad i = 1, \ldots, n, \tag{3.1}$$

where the link function g is a monotonic differentiable function (the logarithmic link may often be suitable, as it ensures positive values of the mean response), z_{ik} is the value of the kth covariate for the ith observation, $f_k(.)$ is the one-dimensional smooth function for covariate k, and n is the number of observations. Letting $\hat{l}_i$ denote the fitted values of $\mathrm{E}(l_i)$, then given the estimated effective strip half-width, $\hat{\mu}_w$, the estimated animal density in the interval l_i between detections $i-1$ and i is $[2\hat{\mu}_w \hat{l}_i]^{-1}$.

Although the underlying density of animals might be expected to be approximately constant between two closely spaced detections, it would not be so for successive detections that are far apart. In the latter case in particular, animal density would be expected to vary along the trackline in the interval between detections. However, the GLM/GAM framework outlined above only allows for a constant value of this density. We therefore propose that an iterative procedure be implemented that can accommodate variation in density, in the spatial covariates and in the expected encounter rate between detections. This procedure is such that each observed waiting distance is transformed to the corresponding distance (with respect to the CDF) had the underlying (inhomogeneous) Poisson process been homogeneous with rate equal to the estimated rate at the location of the next detection. The steps of this iterative procedure are as follows:

1. Fit a GLM or GAM to the waiting distance data, as if the effort associated with waiting distance l_i was all located at x_i rather than in the interval $(x_{i-1}, x_i]$ for $i = 1, \ldots, n$ (where $x_i = x_{i-1} + l_i$). From the fitted model, obtain the estimated density along the line, $\hat{D}(x)$. This estimated density is biased because it ignores the fact that density varies along the line.
2. Adjust each waiting distance, l_i, as follows. From Equation (2.7), the cumulative distribution function of l_i given x_{i-1} is estimated by

$$\hat{\mathrm{F}}_{l|x}(l_i|x_{i-1}) = 1 - \exp\left\{-2\hat{\mu}_w \int_{x_{i-1}}^{x_{i-1}+l_i} \hat{D}(x)\mathrm{d}x\right\} .$$

   If, between x_{i-1} and x_i, density were constant along the line, with estimated value $\hat{D}(x_i)$, then we would have

$$\hat{\mathrm{F}}_{l|x}(l_i|x_{i-1}) = 1 - \exp\left\{-2\hat{\mu}_w l_i \hat{D}(x_i)\right\} .$$

   By equating these expressions, with l_i' substituted for l_i in the second expression, we obtain an adjusted waiting distance, where the adjustment is for variable density along the line:

$$l_i' = \frac{\int_{x_{i-1}}^{x_{i-1}+l_i} \hat{D}(x)\mathrm{d}x}{\hat{D}(x_i)}, \quad i = 1, \ldots, n\,. \tag{3.2}$$

3. The GLM or GAM is now fitted to the adjusted waiting distances, to obtain the estimated density along the line, $\hat{D}(x)$. This should have lower bias than the estimate from Step 1, because we replace the observed waiting distances along the lines by waiting distances that have been adjusted to the estimated density at the locations of detections. Each adjusted waiting distance is therefore associated with a single location in space, allowing simple fitting of the GLM or GAM. However, $\hat{D}(x)$ was obtained by fitting to unadjusted waiting distances. We therefore recalculate the adjusted waiting distances, using the updated estimate of density along the line, and the procedure is iterated until convergence.

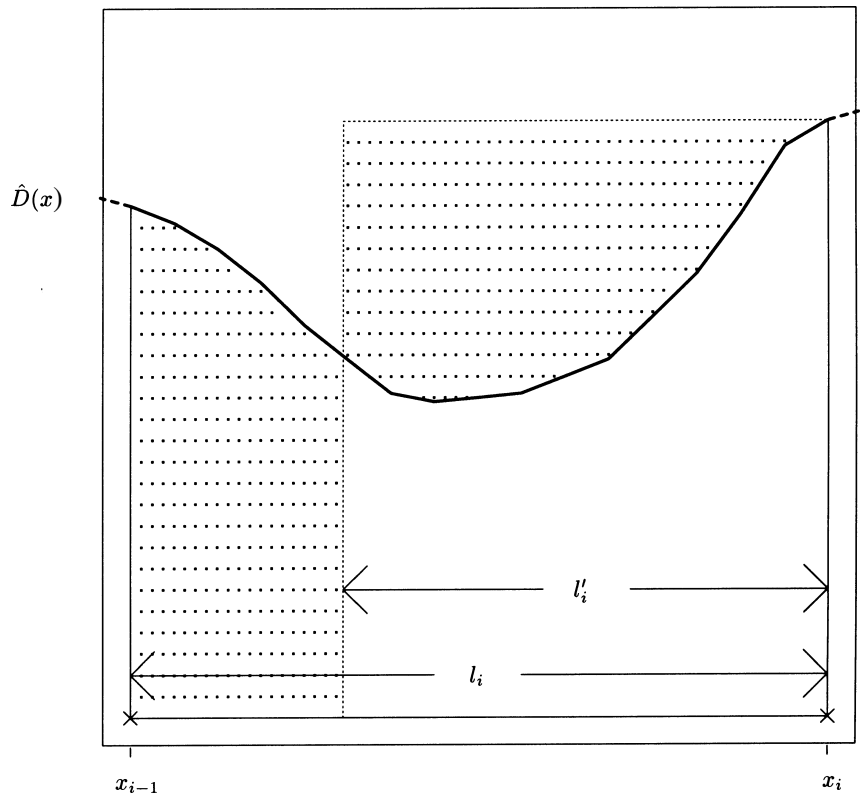

Figure 1. Density along the trackline, $\hat{D}(x)$, with detections at x_{i-1} and x_i. The distance between detections is shown as l_i; the adjusted waiting distance is shown as l_i'. l_i' is such that the area of the rectangle of width l_i' and height $\hat{D}(x_i)$ equals the area under $\hat{D}(x)$ between x_{i-1} and x_i (i.e., so that the dotted regions are equal in area).

As depicted in Figure 1, the iterative procedure equates the area under the predicted density surface between successive detections (at x_{i-1} and x_i, say) to the area of the rectangle of width l_i' and height $\hat{D}(x_i)$. We note that the procedure as described does not use the effort data recorded following the last detection. One solution is to modify the waiting distance data such that the distance to the first detection is redefined to be the sum of l_1 plus the right-censored distance, l_{n+1}.

3.2 A FORMULATION BASED ON COUNTS

Representing a special case of line transect surveys, strip transect surveys provide a good method for estimating the abundance of some wildlife species, for example, feeding or resting seabirds. Transect lines of total length L are covered within a survey area A, and it is assumed that all animals out to a perpendicular width w on either side of the lines are detected with certainty. Any detections made beyond w are excluded from the analysis. The

strip transect formulation provides a starting point for formulating a spatial line transect model for count data.

Suppose that a set of transects from a strip transect survey have been divided into T small contiguous sampling units or "segments" each of (approximately) equal length, the length of each segment being such that the geographic location does not change appreciably within a segment. Let the length of the ith segment be denoted by l_i and the number of animals detected within it by $n_i,\ i = 1, \ldots, T$. Now suppose that for each segment, a set of K spatial covariates is available, and let z_{ik} denote the value of the kth spatial covariate in the ith segment. The expected values of the n_i may be related to the spatial covariates using a GLM or GAM formulation, for example,

$$\mathrm{E}(n_i) = \exp\left[\ln(2l_i w) + \beta_0 + \sum_k \beta_k z_{ik}\right], \quad i = 1, \ldots, T\,. \tag{3.3}$$

The logarithm of the area of each segment, $\ln(2l_i w)$, enters the linear predictor as an offset, and the $\beta_k,\ k = 0, \ldots, K$, are parameters to be estimated. The error distribution (or the variance-mean relationship) must also be specified, and should be such that overdispersion, if present, is accounted for.

To extend this formulation so that it may be used to model line transect data, we first consider the case where the estimated probability of detection, $\hat{p}_{ij}$, of the jth animal in the ith segment is equal for all animals in that segment, such that $\hat{p}_{ij} = \hat{p}_i,\ j = 1, \ldots, n_i$. Inclusion of $\hat{p}_i$ in the offset term yields a formulation suitable for spatial modeling of line transect data conditional on the $\hat{p}_i$, for example, a GLM of the following general form:

$$\mathrm{E}(n_i) = \exp\left[\ln(2l_i w\hat{p}_i) + \beta_0 + \sum_k \beta_k z_{ik}\right], \quad i = 1, \ldots, T\,. \tag{3.4}$$

No additional modeling difficulties arise, but the component of variance due to uncertainty in the estimation of the p_i must be included in variance estimates of density and abundance.

In both cases, the response variable in Equations (3.3) and (3.4) is n_i, the number of detected animals in segment i. By definition, the probability of detecting an animal within w in the strip transect example is unity, and is therefore the same for all animals. In the above line transect case (Equation (3.4)), attention was restricted to the case where the probability of detection of animals was allowed to vary between segment locations, but every animal in a given segment was assumed to have the same probability of detection, regardless of the properties (e.g., behavior, size) of the animal. In some cases, individual covariates might be available for modeling probability of detection. From Equation (3.4), it can be seen that the estimated number of animals (both detected and undetected), $\hat{N}_i$ say, in the ith segment is $n_i/\hat{p}_i$. This type of estimator resembles the Horvitz–Thompson estimator (Horvitz and Thompson 1952). When estimated detection probability varies for each animal, then the $\hat{N}_i$ are estimated by

$$\hat{N}_i = \sum_{j=1}^{n_i} \frac{1}{\hat{p}_{ij}}, \quad i = 1, \ldots, T\,. \tag{3.5}$$

These estimates then form the response for a spatial model, with the offset in Equation (3.4) being modified appropriately.

3.3 VARIANCE ESTIMATION

Having fitted a spatial model, interest lies not only in the ability of the model to predict density and abundance throughout some large survey region, and possibly within small geographic areas within this region, but also in its ability to quantify the precision of these estimates. We consider that a resampling technique is most appropriate for this purpose, preferring it to an analytical method because it offers greater flexibility for incorporating different sources of variability, and also because it provides a common approach for all three modeling approaches described in this article.

Implementation of the nonparametric bootstrap is relatively straightforward for these models. As in conventional distance sampling, transect lines or days of survey effort can be assumed to be independently and identically distributed, and these units should be resampled with replacement. Both components of the modeling (the detection function estimation and the spatial density estimation) are then carried out on the new resampled datasets. For each resample, density and abundance throughout the survey region can be predicted just as for the model fit from the original data. Variance in abundance, for example, is simply the sample variance of the abundance estimates from the predictions using the resampled data. The number of resamples should be at least 100, and preferably in the range 400–1,000 for reliable estimation (Buckland 1984). Model selection uncertainty could be incorporated in both the detection function estimation and in the spatial modeling. The disadvantage of the nonparametric bootstrap is that it almost inevitably leads to poorer spatial coverage in the resampled datasets than in the original data. This is less problematic if there are a large number of sampling units.

The parametric bootstrap provides an alternative approach. Consider first the spatial modeling component of the estimation. The fitted model can be used to estimate density at "every" point (i.e., some set of closely spaced points) along the transect lines, and a rejection sampling technique, for example, provides a means to generate new data points at various points along those lines. Conditioning on the original total effort, the number of points to generate in each resample is a random Poisson variable, with mean equal to the number of sightings in the original data. The spatial model is then fitted to each resampled dataset. As for the nonparametric bootstrap, model selection uncertainty could optionally be incorporated by reselecting a model for each resample, rather than conditioning on the original model fit (Buckland, Burnham, and Augustin 1997). The sample variance of the predicted abundance estimates from the resamples provides one component of the variance; the uncertainty in estimated strip width, or detection probabilities, must be incorporated also. Estimation of the latter component was described by Buckland et al. (2001). Assuming independence between the two components, they may be combined using the delta method (Seber 1982) to obtain an estimate of the overall variance.

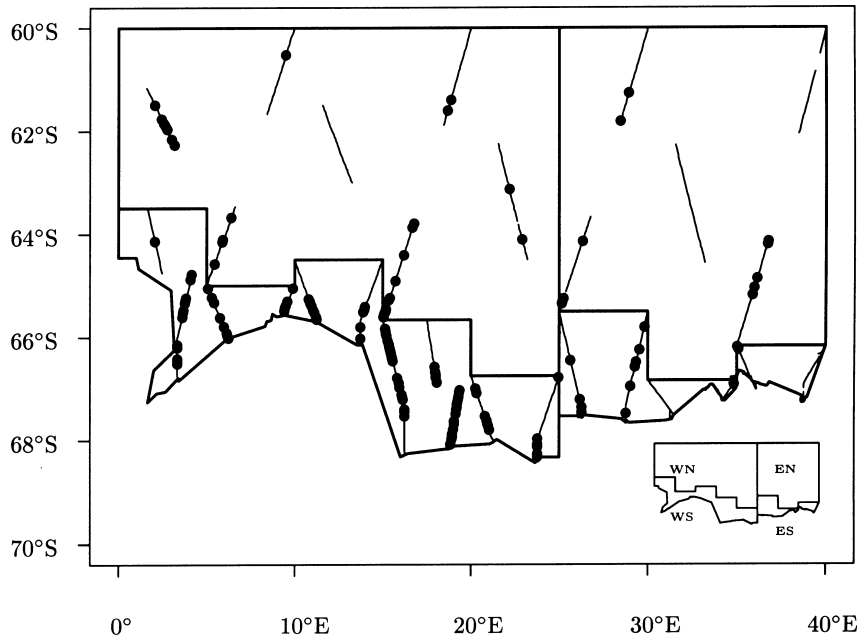

Figure 2. Realized survey effort in IO mode and sightings of minke whale pods during the 1992–1993 IWC/IDCR Antarctic Survey. The southern survey boundary is defined by the extent of sea ice from the Antarctic continent. Subplot shows the division of the region into four strata: WN, WS, EN, and ES.

4. APPLICATION: SPATIAL DISTRIBUTION OF ANTARCTIC MINKE WHALES

The data used in this example are sightings of minke whale pods (with typically 1–3 animals in a pod) from the IWC IDCR line transect survey conducted in the Southern Ocean in the austral summer of 1992–1993. The Southern Ocean is divided into six IWC "Management Areas," each covering a longitudinal width of 60°, with latitudinal coverage from 60° S to the ice edge. The 1992–1993 survey covered most of IWC Area III, from 0° to 40° E (the part of the Southern Ocean to the south of southern Africa). Two research vessels (the *Shonan Maru* (SM1) and the *Shonan Maru No. 2* (SM2)) undertook the survey. These vessels are virtually identical, and on this survey each operated three observation platforms: the barrel (or crow's nest) accommodating two observers; the Independent Observer Platform (IOP) accommodating one observer; and the front bridge, accommodating up to three observers in addition to three researchers. Observers rotated positions every hour, and rest periods were scheduled. Because minke whale densities are generally higher closer to the ice edge, the survey region was divided into four geographic strata with relatively more survey effort dedicated to the two southern strata, that is, those adjacent to the ice edge (Figure 2). The vessels were each assigned one southern and one northern stratum to complete.

Transects were placed in a zig-zag design, which was neither strictly randomized nor systematic with random start point. The SM1's start point was predetermined to be at 0°

Table 1. Stratum Estimates of $\hat{p}$, the Detection Probability of Minke Pods Within a Strip of Half-Width 1.5 n.miles, and of $\widehat{\mu_w}$, the Effective Strip Half-Width. Estimates and their coefficients of variation (% CV) were calculated using Distance software (Thomas et al. 2003).

Pooled strata	$\hat{p}$	*% CV*	$\widehat{\mu_w}$ *(n.miles)*	*% CV*
WN and EN	0.742	7.61	1.112	7.61
WS and ES	0.360	15.42	0.540	15.42

longitude, 60° S (the northwestern corner of the survey area), and the SM2's start point had some degree of "environmental randomization"—its latitudinal point being determined by the ice edge location (at longitude 0°). The survey was conducted in two alternating survey modes: closing mode and independent observer (IO) mode. A description of these modes (and comprehensive details of all IDCR/SOWER surveys up to 1997–1998) was provided by Branch and Butterworth (2001). Note that in the example analysis presented here, we use data from IO mode only.

Sightings data were pooled across the northern strata and southern strata for estimating effective strip half-widths, and in both cases, the perpendicular distance data were truncated at $w = 1.5$ nautical miles (n.miles). Resulting estimates are shown in Table 1. These estimates were used to apply the three methods described in this article. The detection function was assumed to depend only on the perpendicular distance y from the line, i.e., $g(x, y) \equiv g(y)$. Three spatial covariates (latitude, longitude, and distance from the ice edge) were available for inclusion in the models.

For direct likelihood maximization, it was assumed that, within the surveyed strip, minke whale pod density varied only along the transect lines, that is, $D(x, y) = D(x)$, independently of y, for $|y| \leq w$. This assumption is probably reasonable because perpendicular detection distances (of up to about three n.miles) were relatively small compared with both the scale at which the spatial trend in animal density varied, and with the lengths of the transect lines (several tens of miles). $D(x)$ was modeled as a linear function of the three spatial covariates.

For the two GLM/GAM-based models, GAMs with logarithmic links were fitted using all three covariates. Latitude and longitude were included as smoothing spline terms each with four degrees of freedom; the effect of distance from the ice edge was modeled linearly (Figures 3 and 4). This strategy was adopted to compare the GAM-based models on an equal footing, providing valid comparisons between them, rather than provide "best fits" to the data in each case. For the waiting distance model, a gamma error distribution was assumed, while for the count model, transects were divided into segments of approximately three n.miles and then the estimated number of minke pods in each segment was modeled assuming a variance-mean relationship corresponding to an overdispersed Poisson distribution.

Pod abundance estimates for the three approaches, together with estimates from a conventional line transect analysis, are given in Table 2. Coefficients of variation for the spatial methods were obtained using the parametric bootstrap procedure described in Section 3.3, conditioning on the original model.

The spatial methods all yielded higher estimates of pod abundance than the stratified method, and the precision of the estimates was greater. The spatial models gave even larger

improvements in precision for estimates of abundance by stratum. This occurs because spatial models use data from all strata to estimate the density surface, and hence abundance, in each; stratum estimates are therefore estimated with higher precision, but also with positive correlation, so that the gain in precision in the estimate of total abundance is less. The improved precision of the spatial methods in this example is substantial, an important consideration for surveys in which ship time is expensive. A further advantage of the spatial methods is that the density surface is estimated. The three models fitted here yield the surfaces shown in Figures 5–7.

Considering the markedly different model formulations and the modest sample sizes, Figures 6 and 7 are very similar. They both show high densities in the south western stratum

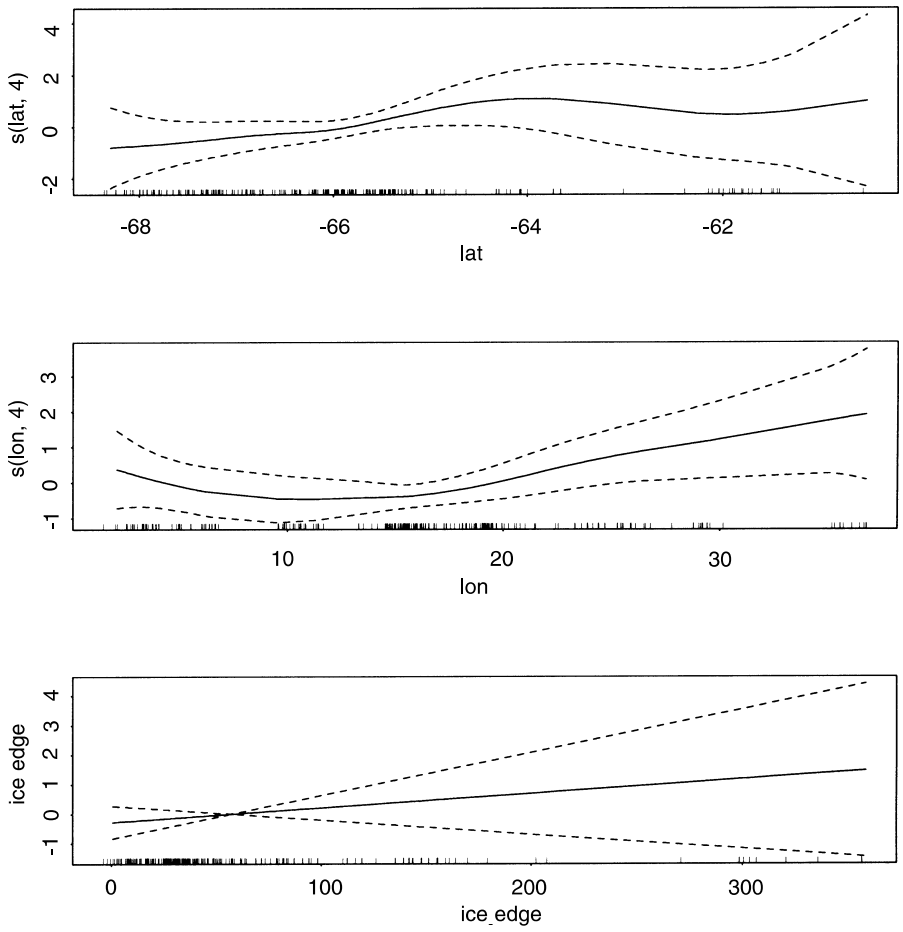

Figure 3. Estimated conditional dependence of waiting distance on latitude and longitude (smooths with 4 df) and distance from the ice edge (linear). Estimates (solid) and confidence intervals (dashed), with covariate values as a rug plot along the bottom of each plot are shown. A high value of the smooth corresponds to long waiting distances and hence low density.

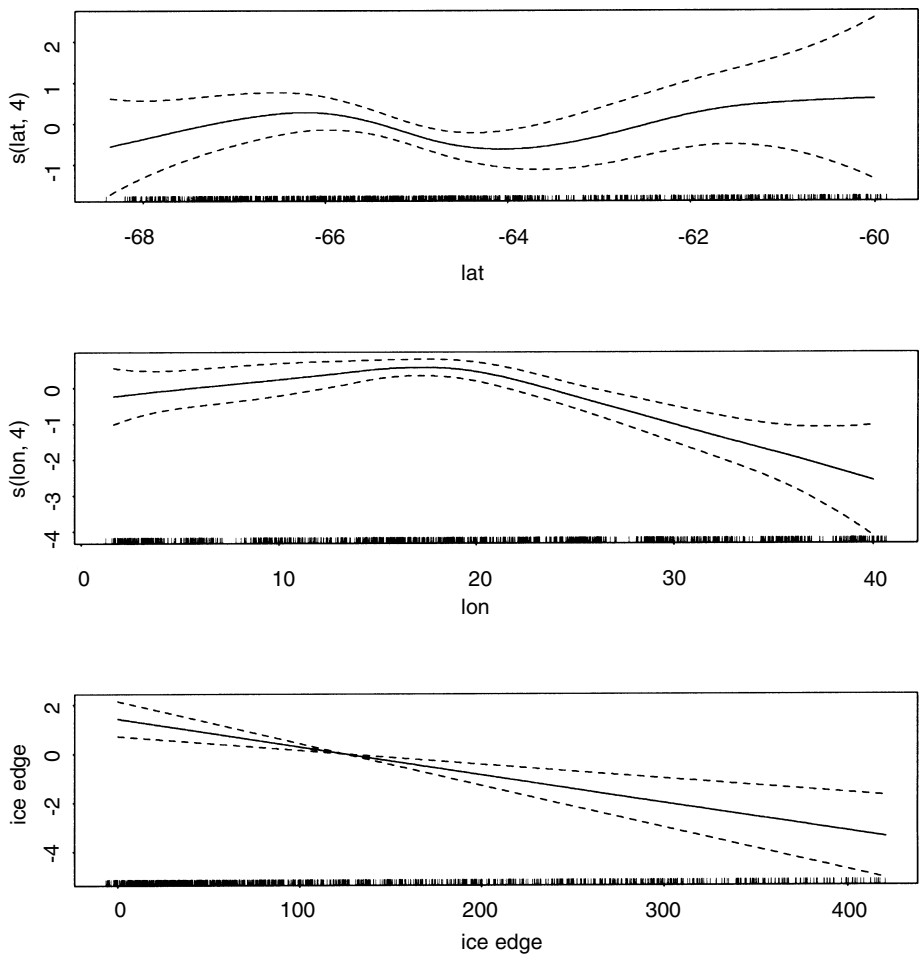

Figure 4. Estimated conditional dependence of density on latitude and longitude (smooths with 4 df) and distance from the ice edge (linear), using the count model. Estimates (solid) and confidence intervals (dashed), with covariate values as a rug plot along the bottom of each plot are shown.

of the survey area, with increasing densities at increasing latitudes. The highest densities do not extend into the eastern sector of the survey region. Although it is known that minke whale densities are generally higher close to the ice edge, the predicted distribution patterns are quite plausible. It is not possible to interpret them definitively from the available data, but the east-west variation in densities in the southern strata could be caused by any number of factors related to cetacean habitat (e.g., temperature, salinity, ice type and concentration, krill distribution and abundance, and so on). Within the western sector, the waiting distance model suggests a distribution shifted slightly to the east relative to the count model. This would suggest a lack of robustness if we required abundance estimates by subregions appreciably smaller than the ice-edge strata. Figure 5 also shows high density along the ice-edge, but the model clearly oversmooths the data. A consequence of using direct maximization of the likelihood, for which convergence is problematic, is that we have fitted too simple a model.

Table 2. Comparison of Estimates of Abundance of Minke Whale Pods ($\hat{N}$) from a Conventional Stratified Line Transect Analysis, Direct Maximization of the Likelihood, and Two GAM-based Approaches, One Modeling the Waiting Distances Between Detections, and the Other Modeling the Estimated Number of Minke Whale Pods in Segments of Length 3 n.miles. Coefficients of variation (% CVs) were estimated using the parametric bootstrap.

	Stratified Analysis		*Likelihood Maximization*		*GAM Waiting distances*		*GAM Counts*	
Stratum	$\hat{N}$	*% CV*	$\hat{N}$	*% CV*	$\hat{N}$	*% CV*	$\hat{N}$	*% CV*
WN	4,810	40.1	4,970	13.6	5,810	20.8	4,920	16.3
EN	1,460	49.5	1,440	14.5	800	31.1	760	25.0
WS	7,410	25.1	7,740	16.4	7,820	17.1	8,130	17.1
ES	640	44.3	1,210	17.0	910	22.9	990	21.5
All	14,320	23.0	15,360	15.9	15,340	16.1	14,800	15.9

5. DISCUSSION

Our count model is rather simplistic and subjectively divides the line into segments. However, it seems to work well in practice. The approach is particularly appropriate for strip transect counts of seals hauled out on ice. In this case, animals, where they occur, may be numerous and close together, so it is not practical to attempt to estimate distances along the line between successive animals. Instead, photographs are taken, and the number of animals in each photograph, after deleting any overlap, is counted later. Thus, the data

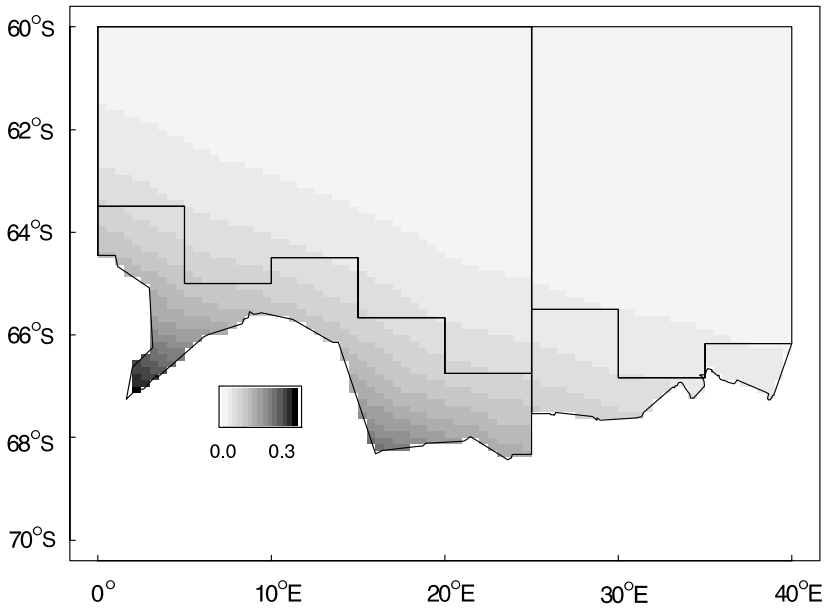

Figure 5. Density of minke whale pods in the surveyed region, predicted by maximizing the likelihood of observed waiting distances, conditional on the estimates of effective strip half-widths given in Table 1.

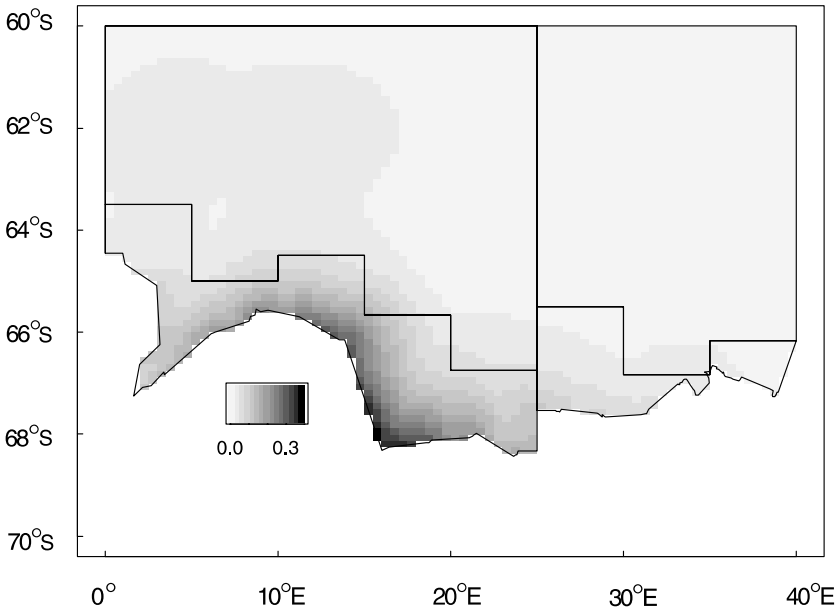

Figure 6. Density of minke whale pods in the surveyed region, predicted from a waiting distances model. Explanatory variables were smoothing splines of latitude and longitude (each with 4 df), and a linear term for distance from the ice edge boundary.

are recorded as counts by line segment, with segment length being determined by the field of view of the camera, and the count method is therefore the natural way to fit a spatial model to these data. The count model can also be useful when the choice of segment length is not so clearly defined. In our experience, we have found the methods to be rather insensitive to choice of segment length, provided that the segments are sufficiently small that expected density is unlikely to vary much within a segment. For habitat-distribution studies, segment size should reflect the scale of the environment, so that habitat varies little within segments. The theoretical optimum segment length is zero, because this avoids bias from density varying within segments, but of course this is of no help to the practitioner. Therefore, in the example, we chose the segment length to be approximately of the same length as the strip width ($2w$), so that the segments were approximately square.

We used the parametric bootstrap to estimate variances in our estimates. For both the waiting distances model and the count model, there was evidence in our example of spatial autocorrelation, after having fitted spatial trends (p value = 0.04 for the waiting distances model, and 0.02 for the count model, using Monte Carlo tests). This may be true autocorrelation, caused perhaps by social behavior of the animals, or simply variation due to covariates that have not been measured. The two cases are indistinguishable given our data, and in either case, some bias in the variance estimates can be anticipated. The appropriate course of action to resolve the problem differs for the two cases. If we simply have the wrong covariates, the ideal would be to measure more relevant covariates, and fit a more appropriate model. This may result in reduced bias in the abundance estimates and

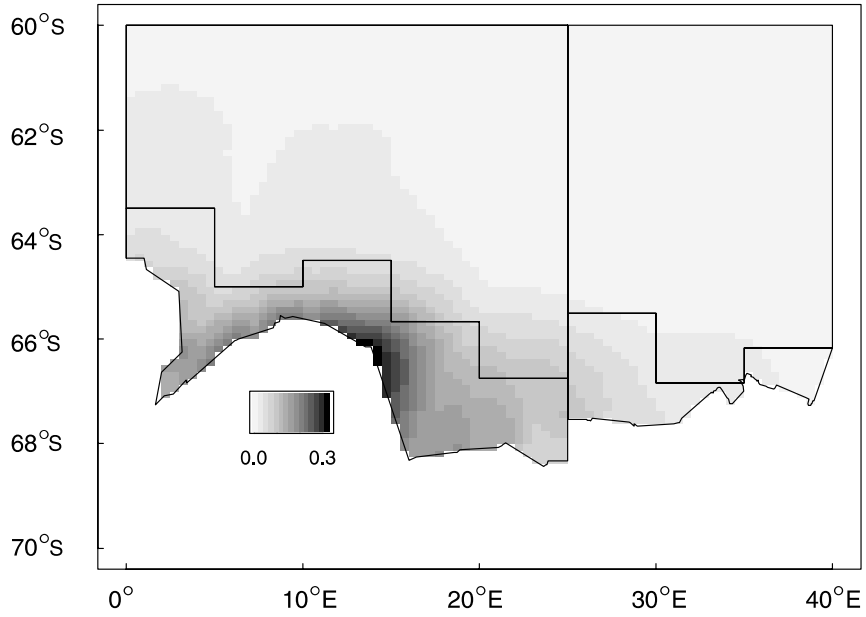

Figure 7. Density of minke whale pods in the surveyed region, predicted from a count model. Explanatory variables were smoothing splines of latitude and longitude (each with 4 df), and a linear term for distance from the ice edge boundary.

a decrease in the variance estimates. If there is true autocorrelation, then we might seek to model it, for example by using the approach of Gotway and Stroup (1997), who presented a framework for combining generalized linear models and quasi-likelihood with geostatistical methods, defining a general variance-covariance matrix that accounts for autocorrelation. This method might be extended to a GAM framework. The approach would undoubtedly yield larger variance estimates, in part because we would no longer be drawing inference on the population that happened to be present at the time of the survey, but on a hyper-population, from which that population was drawn (Augustin, Mugglestone, and Buckland 1998). For example, if animal density was high at a particular location simply because a loose cluster of animals was present on the day of the survey, then the density surface corresponding to the hyper-population should smooth out this cluster, whereas if we simply wanted the most precise estimate of total population size on the day of the survey, we would want to model the higher density at this location using for example flexible GAMs.

Ferguson and Bester (2002) reviewed the literature on spatial autocorrelation, and its relevance to modeling transect counts of Antarctic pack ice seals, concluding that methods that ignored autocorrelation overestimated the correlation between seal densities and environmental variables. Recent advances in the development of spatial generalized linear mixed models (Diggle, Tawn, and Moyeed 1998; Zhang 2002; Christensen and Waagepetersen 2002) are of interest; it might prove possible to extend these approaches for use with autocorrelated line transect data. Similarly, generalized additive mixed models (Lin and Zhang 1999) might also be extended for use with line transect data.

Animals often occur in groups, as with the minke whales of our example. In this case, the methods described here provide estimated density surfaces for groups. To convert these to density surfaces for animals, we need a spatial model for mean group size. Group density is then multiplied by predicted mean group size at any given location to estimate animal density at that location. A spatial model for group size can readily be fitted, because size is observed at n discrete locations, where n is the number of groups detected. GLMs or GAMs may be used to model these observed sizes as a function of spatial covariates. In the case of the count model with individual covariates, for which group density is estimated by Equation (3.5), a minor modification allows us to estimate animal density more directly:

$$\hat{N}_i = \sum_{j=1}^{n_i} \frac{s_{ij}}{\hat{p}_{ij}}, \quad i = 1, \ldots T, \tag{5.1}$$

where s_{ij} is the size of the jth detected group in the ith segment. Size bias (the bias arising from the higher probability of detection of larger groups and hence over-sampling of those groups) is accounted for in Equation (5.1). If size bias is suspected, then the estimated probability of detection of each group should be similarly included in a spatial model for group size, weighting each observation by the reciprocal of its estimated detection probability.

We have illustrated our methods using explanatory variables that are not ecologically meaningful. Instead, our example uses locational covariates (latitude, longitude, and distance from the ice edge) which only serve as proxies for other covariates which might genuinely be expected to influence habitat selection, such as food resource, sea surface temperature, and upwellings. The availability of spatial line transect models will encourage researchers to identify and measure variables more directly relevant to the species of interest.

Conventional line transect sampling draws model-based inference within the surveyed strips, by modeling the detection function $g(y)$. However, it relies on design-based methods to extend that inference to the wider survey region. This approach is robust for estimating overall abundance, but cannot estimate a density surface, or relate abundance to spatial covariates. If the main goal is simply abundance estimation, conventional methods would normally be preferred, because model selection and model fitting are not straightforward for spatial line transect models. Model selection in GAMs is an area of active research (e.g., Hastie and Tibshirani 1990; Wood 2000). If a poor model is selected, substantial bias in abundance estimation might arise. If this is suspected, we can fit more than one plausible model and carry out model averaging (e.g., Buckland et al. 1997).

ACKNOWLEDGMENTS

The authors thank David Borchers for many useful discussions relating to this work. The International Whaling Commission provided funding for Sharon Hedley's contribution to this article. The comments and suggestions of two anonymous referees and an Associate Editor are gratefully acknowledged.

[Received March 2002. Revised June 2002.]

REFERENCES

Augustin, N. H., Mugglestone, M. A., and Buckland, S. T. (1998), "The Role of Simulation in Modelling Spatially Correlated Data," *Environmetrics*, 9, 175–196.

Branch, T. A., and Butterworth, D. S. (2001), "Southern Hemisphere Minke Whales: Standardised Abundance Estimates from the 1978/79 to 1997/98 IDCR-SOWER Surveys," *Journal of Cetacean Research and Management*, 3, 143–174.

Buckland, S. T. (1984), "Monte Carlo Confidence Intervals," *Biometrics*, 40, 811–817.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., and Thomas, L. (2001), *Introduction to Distance Sampling*, Oxford: Oxford University Press.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618.

Christensen, O. F., and Waagepetersen, R. (2002), "Bayesian Prediction of Spatial Count Data Using Generalized Linear Mixed Models," *Biometrics*, 58, 280–286.

Cressie, N. A. C. (1991), *Statistics for Spatial Data*, New York: Wiley.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-Based Geostatistics," *Applied Statistics*, 47, 299–350.

Ferguson, J. W. H., and Bester, M. N. (2002), "The Treatment of Spatial Autocorrelation in Biological Surveys: The Case of Line Transect Surveys," *Antarctic Science*, 14, 115–122.

Gotway, C. A., and Stroup, W. W. (1997), "A Generalized Linear Model Approach to Spatial Data Analysis and Prediction," *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 157–178.

Hastie, T. J., and Tibshirani, R. J. (1990), *Monographs on Statistics and Applied Probability. 43. Generalized Additive Models*, London: Chapman and Hall.

Högmander, H. (1991), "A Random Field Approach to Transect Counts of Wildlife Populations," *Biometrical Journal*, 33, 1013–1023.

——— (1995), *Methods of Spatial Statistics in Monitoring Wildlife Populations*, Jyväskylä: University of Jyväskylä.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Jefferson, T. A., Leatherwood, S., and Webber, M. A. (1993), *FAO Species Identification Guide. Marine Mammals of the World*, Rome: FAO.

Lin, X. H., and Zhang, D. W. (1999), "Inference in Generalized Additive Mixed Models by Using Smoothing Splines," *Journal of the Royal Statistical Society*, Series B, 61, 381–400.

Marques, F. F. C. (2001), "Estimating Wildlife Distribution and Abundance from Line Transect Surveys Conducted from Platforms of Opportunity," unpublished Ph.D. thesis, University of St. Andrews, Scotland.

Seber, G. A. F. (1982), *The Estimation of Animal Abundance and Related Parameters* (2nd ed.), London: Edward Arnold.

Stoyan, D. (1982), "A Remark on the Line Transect Method," *Biometrical Journal*, 24, 191–195.

Thomas, L., Laake, J. L., Strindberg, S., Marques, F. F. C., Buckland, S. T., Borchers, D. L., Anderson, D. R., Burnham, K. P., Hedley, S. L., Pollard, J. H., and Bishop, J. R. B. (2003), *Distance 4.1*, Research Unit for Wildlife Population Assessment, University of St. Andrews, U.K.; available from http://www.ruwpa.st-and.ac.uk/distance.

Wood, S. N. (2000), "Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties," *Journal of the Royal Statistical Society*, Series B, 62, 413–428.

Zhang, H. (2002), "On Estimation and Prediction for Spatial Generalized Linear Mixed Models," *Biometrics*, 58, 129–136.