Spatial modelling from line transect data

SHARON L. HEDLEY, STEPHEN T. BUCKLAND AND DAVID L. BORCHERS

Research Unit for Wildlife Population Assessment, Mathematical Institute, North Haugh, St Andrews, Fife KY16 9SS, UK
Contact e-mail: sharon@mcs.st-and.ac.uk

ABSTRACT

In this paper, two new methods are presented that enable spatial models to be fitted from line transect data. Building on preliminary work by Cumberworth *et al.* (1996) and Hedley *et al.* (1997), the first method is based on a count model and involves dividing the survey effort into small segments then modelling the number of schools in each segment. In contrast, the second method uses a model based on the intervals between detections. Its formulation is derived in detail to obtain the likelihood function for the distances between detections, conditional on an estimated detection function. Both models can be fitted using standard statistical software, although variances must be estimated using computer intensive methods. We apply the methods to data from the 1992/93 IWC/IDCR Antarctic survey of Area III, fitting generalised additive models to obtain estimates of minke whale abundance, using the parametric bootstrap to estimate variance. The results from fitting these models are compared with the results of a previous analysis by Borchers and Cameron (1995), which used conventional stratified methods.

KEYWORDS: MODELLING; DISTRIBUTION; ABUNDANCE ESTIMATE; SURVEY-VESSEL; MINKE WHALE; SOUTHERN HEMISPHERE; ANTARCTIC

INTRODUCTION

The question of how to achieve one of the long-term research objectives of the IWC, the assessment and prediction of environmental change on cetaceans, is summarised in IWC (2000) as:

> Define how spatial and temporal variability in the physical (e.g. sea surface temperature, salinity, mixed layer depth, upwelling, extent of ice cover) and biological (e.g. prey availability) environment influence cetacean species in order to determine those processes in the marine ecosystem which best predict long term changes in cetacean distribution, abundance, stock structure, extent and timing of migrations and fitness.

The conventional stratified analyses used to estimate the density of minke whales from IWC surveys (e.g. see Haw, 1993; IWC, 1994) are unsuitable for quantifying the relationships between the distribution of cetaceans, their prey and physical variables because they necessarily involve estimation at very low spatial resolution and have only a limited ability to relate density to physical variables. In contrast, spatial models are already widely applied to wildlife survey data to explain species distribution using physical variables (e.g. see Osborne and Tigar, 1992; Buckland and Elston, 1993; Augustin *et al.*, 1996), and could in principle be used to model the distribution of a predator based on the density and location of its prey. This paper describes two recent methodological developments for estimating the spatial distribution of wildlife from line transect data.

MODEL FRAMEWORK

Data collection

In common with conventional line transect surveys, a set of transects is surveyed within the region of interest. Observers record the radial distance and angle to the detected object, to enable modelling of perpendicular distances (and hence estimation of the effective strip width). Factors which can affect the sightability of objects, such as sea state, should usually also be recorded. The geographical location of the observation platform is logged at regular intervals, and is also recorded when a sighting is made. Other spatial information, for example, bottom depth and sea surface temperature, may be recorded at regular intervals; such information, if also available throughout the survey region, could be included as one or more predictor variables in a model. Unlike conventional line transect analyses, there is no requirement that transect lines should be located at random with respect to the distribution of objects, although they should provide coverage throughout the region of interest.

Using line transect count data to formulate a spatial model

In this formulation, the transects covered during a survey are divided into small sampling units or 'segments', such that the sighting conditions and geographic location do not change appreciably within a segment. The estimated number of groups (e.g. pods, schools, herds) in segment i, $\hat{N}_i$ is calculated using the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) as:

$$\hat{N}_i = \sum_{j=1}^{n_i} \frac{1}{\hat{p}_{ij}} \tag{1}$$

where n_i is the number of detected groups in the i^{th} segment, and $\hat{p}_{ij}$ is the estimated probability of detection of the j^{th} detected school in segment i.

If the 'confirmed' group size[1], s_{ij}, is available for every detected group, the estimated number of animals in segment i, $\hat{M}_i$, is:

$$\hat{M}_i = \sum_{j=1}^{n_i} \frac{s_{ij}}{\hat{p}_{ij}} \tag{2}$$

In conventional line transect analyses, the estimated detection probabilities depend only on perpendicular distance from the transect line. The Horvitz-Thompson estimator provides a framework for incorporating other explanatory variables (such as sea state or group size) which may affect detection probability. In the case of two-platform survey data, the p_{ij} may be estimated using logistic regression as in Borchers *et al.* (1998). In the case of single platform data, the covariable information may be incorporated by adopting a full likelihood approach (Ramsey

[1] A 'confirmed' group is one whose size has been determined with a high degree of confidence.

et al., 1987; Cooke and Leaper, 1998), or by using a 'covariate adjustment' method (Beavers and Ramsey, 1998), in which a log-linear regression is carried out on the observed perpendicular distances to modify the width (or in point transects, the radius) of the effective search area dependent upon the sighting conditions.

Having estimated the response variate, a spatial model is then fitted. This may take the form of either a Generalised Linear Model (GLM; McCullagh and Nelder, 1989) or a Generalised Additive Model (GAM; Hastie and Tibshirani, 1990); here we detail the latter.

Suppose the response is $\hat{N}_i$ the number of groups in the ith segment, and let z_{ik} denote the value of the kth spatial covariable in the ith segment. The expected values of the $\hat{N}_i$ are related to a predictor function of the z_{ik} via a link function. Although the response is derived from count data, in general the Poisson error distribution will not be appropriate because of over-dispersion. This is readily accounted for by specifying an error structure with variance function proportional, rather than equal to the mean. Applying a logarithmic link function, the model may be written

$$\mathrm{E}(\hat{N}_i) = \exp\left[\ln(a_i) + \theta_0 + \sum_k f_k(z_{ik})\right]$$

where the offset variable a_i is the area of the ith segment (calculated as twice the length of the segment multiplied by the perpendicular distance at which data were truncated in that segment); θ_0 is a parameter to be estimated (commonly termed the 'intercept'); and the f_k are smoothed functions of the spatial covariables. If $\hat{p}_{ij} = \hat{p}_i$ for all $j, j = 1,...,n_i$, then the counts in each segment can be modelled directly, with the probability of detection of a group being incorporated into the model via the offset, $\hat{p}_i a_i$, the effective area of the ith segment.

Using interval data to formulate a spatial model

In this formulation, interval data form the response. These may be the times or distances (along the transect line) between successive sightings, or the 'areas' between sightings, where the areas are defined as the distances along the transect line between sightings multiplied by twice the effective strip half-width at the given location. In what follows, we term these three responses 'waiting time', 'waiting distance' and 'waiting area' respectively.

Animal density is not usually constant within a particular region. It typically varies according to the local environment, so it is not reasonable to assume that the locations of groups of animals are well described by a homogeneous planar Poisson process. It is more likely that the locations of groups would be better modelled by an inhomogeneous Poisson process, where the rate of the process (effectively the expected number of groups per unit area) is allowed to vary as a function of spatial location. Let this rate be denoted $D(x,y)$, the density of groups at position (x,y), where the x-axis is defined (without loss of generality) to be along the direction of the trackline and the y-axis to be orthogonal to the trackline. Consider two arbitrary points on the trackline $(x_1,0)$ and $(x_2,0)$, separated by a distance l along the line. The expected number of groups within a strip of length l and width $2w$ is:

$$\int_{-w}^{w} \int_{x_1}^{x_2} D(x,y)dxdy$$

Suppose that not all groups within the strip are detected. The expected number of detected groups within the strip is a function of how many groups are actually present in the strip and of how detectable they are. Conventional line transect estimation requires estimation of a detection function, $g(y)$; in this formulation, $g(y)$ may be thought of as a thinning function producing the observed process - the locations of detected animals. Under the additional assumption that the detection process is independent of the density of groups, the locations of detected groups within $\pm w$ of the trackline also follow an inhomogeneous Poisson process, with rate $D(x,y)g(y)$. The validity of this assumption is compromised if detection probability is greater in high density areas. This situation may occur if observers become more alert in the presence of sightings, but with appropriate survey protocol, such as frequently rotating shifts, the effect of this can be minimised. Thus the expected number of detections in the strip is given by:

$$\int_{-w}^{w} \int_{x_1}^{x_2} D(x,y)g(y)dxdy = \lambda(x_1,x_2,w) \qquad \text{(say)},$$

and:

$$\begin{aligned} &\text{P(no detections in this strip)} \\ &= \frac{[\lambda(x_1,x_2,w)]^0 \exp\{-\lambda(x_1,x_2,w)\}}{0!} \\ &= \exp\{-\lambda(x_1,x_2,w)\} \end{aligned}$$

Now let M denote the length of trackline surveyed before a detection is made, starting from $(x_1,0)$. The Cumulative Density Function (CDF) of M evaluated at $(x_1+l,0) = (x_2,0)$ is:

$$F_M(l \mid x_1) = \mathrm{P}(M \leq l \mid x_1) = 1 - \mathrm{P}(M > l \mid x_1)$$

But since $M > l$ if and only if there were no detections in the strip, then

$$\begin{aligned} F_m(l \mid x_1) &= 1 - \text{P(no detections in the strip)} \qquad (3) \\ &= 1 - \exp\{-\lambda(x_1,x_2,w)\} \end{aligned}$$

Thus, we obtain the conditional probability density function (pdf) of the waiting distance, l, from an arbitrary point $(x_1,0)$:

$$f_M(l \mid x_1) = \int_{-w}^{w} D(x_2,y)g(y)dy \; \exp\{-\lambda(x_1,x_2,w)\}$$

Suppose that there are n detections. The vessel locations at each detection are $(x_i,0) = (x_{i-1}+l_i,0)$, $i = 1,...,$n, which, given the location of the start of the surveyed trackline, are uniquely determined by the length of survey effort between each detection. Denoting the initial position by $\xi_0 = (x_0,0)$, the joint likelihood of waiting distances given ξ_0 is:

$$\begin{aligned} &f(l_{n+1}, l_n, l_{n-1}, \ldots, l_1 \mid \xi_0) = \\ &\quad \mathrm{P}(M > l_{n+1} \mid l_n, \ldots, l_1, \xi_0) f(l_n \mid l_{n-1}, \ldots, l_1, \xi_0) \ldots f(l_1 \mid \xi_0) \end{aligned}$$

where l_1 is the distance from ξ_0 to the location of the vessel when the first detection is made; l_i, $i = 2,...,n$ are the distances between the vessel locations at the (i-1)th and ith detections; and l_{n+1} is the distance surveyed on effort after the last detection. $\mathrm{P}(M > l_{n+1} \mid l_n,...,l_1,\xi_0)$ is simply $1 - F_M(l_{n+1} \mid l_n,...,l_1,\xi_0)$. Given a vector of spatial parameters θ for the density surface $D(x,y)$ and parameters β for the detection function $g(y)$, the conditional likelihood $\mathcal{L}(l|\xi_0;\theta,\beta)$ is given by:

$$\mathcal{L}(l|\xi_0;\theta,\beta) = \left[\prod_{i=1}^{n}\int_{-w}^{w} D(x_i,y)g(y)dy\right]$$
$$\exp\left[-\sum_{i=1}^{n}\lambda(x_{i-1},x_i,w)\right]\exp[-\lambda(x_n,x_n+l_{n+1},w)]$$

The conditional log-likelihood is:

$$\ln\mathcal{L}(l|\xi_0;\theta,\beta) = \sum_{i=1}^{n}\ln\left[\int_{-w}^{w} D(x_i,y)g(y)dy\right]$$
$$-\sum_{i=1}^{n+1}\lambda(x_{i-1},x_{i-1}+l_i,w)$$

which, given parametric forms for $D(x,y)$ and $g(y)$, may be maximised numerically.

In order to compare this derivation with the conventional line transect estimator (e.g. Buckland *et al.*, 1993, pp.37-39), a 'flat' density surface representing the average density across a stratum must be estimated, so $D(x,y)$ becomes simply D. Thus, the log-likelihood becomes:

$$\ln\mathcal{L}(l|\xi_0;D,\beta) = n\ln\left[D\int_{-w}^{w} g(y)dy\right] - DL\int_{-w}^{w} g(y)dy$$

where L is the total distance surveyed. Differentiating with respect to D to find the maximum likelihood estimate of density, $\hat{D}$,

$$\frac{\partial\ln\mathcal{L}}{\partial D} = \frac{n}{D} - L\int_{-w}^{w} g(y)dy$$

gives the conventional line transect estimate of group density:

$$\hat{D} = \frac{n}{L\int_{-w}^{w} g(y)dy}$$

Using standard software to fit an interval data model

In the previous section, the likelihood function of the inter-detection distances (waiting distances) was derived and it was shown that the maximum likelihood estimator of density was equivalent to the conventional line transect density estimator under the necessary constraints. We now describe how interval data can be used to fit a spatial model within a GAM framework.

We begin with a conceptual model in which the density of groups and the expected encounter rate remain constant as the observer travels from one detection to the next, but may change when a detection occurs. In reality, this is clearly an implausible model, but it serves as a starting point for a more appropriate model which must be fitted iteratively. Suppose that detections occur at $(x_{i-1},0)$ and $(x_i,0)$ where $x_i=x_{i-1}+l_i$ and assume for simplicity that the effective strip half-width, μ, is a constant. Under these conditions, the CDF given in Equation (3) becomes:

$$F_M(l_i \mid x_{i-1}) = 1-\exp\{-2\mu l_i D(x_i,0)\}$$

i.e. the distances are distributed exponentially, Exp(ϕ) say, with $\phi = 2\mu D(x_i,0)$. The mean of this distribution is $1/\phi$ and ϕ is the intensity of a homogeneous Poisson process in the area of length l_i and width 2μ between $(x_{i-1},0)$ and $(x_i,0)$. The observed density at $(x_i,0)$ is $1/2\mu l_i$. The form of the GAM is

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

where g is a monotonic differentiable function (the link function), the f_k are smoothed functions of the spatial covariables, z_{ik}, and n is the number of detections. If it is assumed that the l_i are exponentially distributed, then the gamma distribution, with dispersion parameter set to unity, is the appropriate error distribution (as this is equivalent to the exponential distribution). If there is over-dispersion, the gamma error distribution should be used and the dispersion parameter estimated. The logarithmic link ensures positive values of the mean response.

This formulation can only be fitted as a GAM because of the assumption of constant density between detections. This is equivalent to assuming a sequence of homogeneous Poisson processes, where the rate of each process may differ, but is constant between any two consecutive detections. Standard GAM software cannot be used to fit an inhomogeneous Poisson process, where the rate is a function of spatial location. This difficulty is overcome by implementing an iterative procedure which adjusts the observed waiting distances to the distances that would have occurred if the underlying Poisson process was indeed homogeneous and the density between detections was constant.

The first step of the iteration is to fit a model to the observed waiting distances as described above. The adjusted waiting distances, $\tilde{l}_i$, $i = 1,\ldots,n$ satisfy:

$$1-\exp\{-2\mu\int_{x_{i-1}}^{xi-1+l_i} D(x,0)dx\}$$
$$= 1-\exp\{-2\mu\tilde{l}_i D(x_i,0)\}$$

Taking logarithms, this simplifies to:

$$\int_{x_{i-1}}^{x_{i-1}+l_i} D(x,0)dx = \tilde{l}_i D(x_i,0)$$

Therefore, the adjusted distances are calculated as:

$$\tilde{l}_i = \frac{\int_{x_{i-1}}^{x_{i-1}+l_i} D(x,0)dx}{D(x_i,0)}$$

that is, the area under the predicted density surface between detections at $(x_{i-1},0)$ and $(x_i,0)$ is equated to the area of the rectangle of width $\tilde{l}_i$ and height $D(x_i,0)$ (Fig. 1).

Given the fitted model, the integral $\int_{x_{i-1}}^{x_{i-1}+l_i} D(x,0)dx$ may be evaluated numerically. The denominator, $D(x_i,0)$, is calculated as the reciprocal of the ith fitted value multiplied by the effective strip width, 2μ. The model is then refitted to the adjusted distances yielding an estimate of the density surface $D(x,y)$ with smaller bias. The process is repeated, each time adjusting the waiting distances, until convergence is reached.

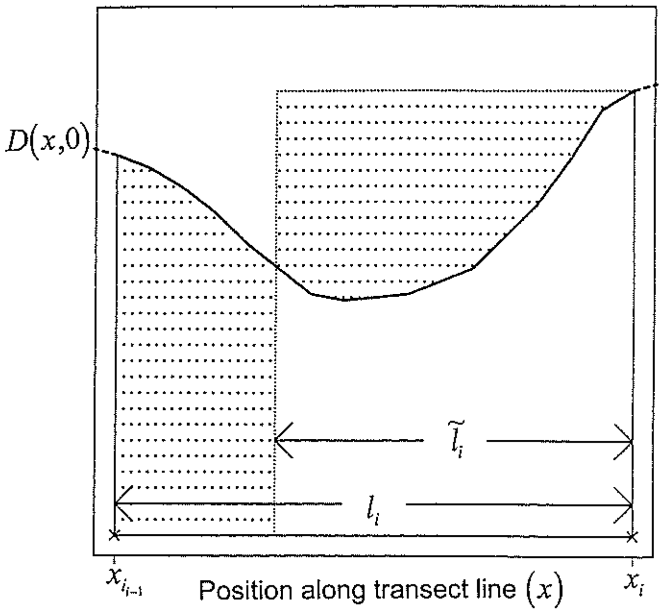

Fig. 1. Density on the trackline, $D(x,0)$, detections at x_{i-1} and x_i. The distance between detections is shown as l_i; the adjusted waiting distance is known as $\tilde{l}_i$, which is such that the area of the rectangle of width $\tilde{l}_i$ and height $D(x_i, 0)$ equals the area under the curve $D(x,0)$ (i.e. so that the dotted regions are equal in area).

VARIANCE ESTIMATION

The observations of the responses from the two models described above cannot be assumed to be independent. However, the likelihoods can still be formulated and model parameters estimated, but analytic methods will not provide valid variance estimates if the non-independence is ignored. Robust estimates of variance may be obtained using an appropriate resampling technique. An obvious choice is the non-parametric bootstrap, with transect lines, or perhaps days, as the sampling units (which are assumed to be independent). With this technique however, samples are drawn with replacement, hence each bootstrap resample is likely to produce reduced spatial coverage compared to the original survey effort. Two alternative estimators were considered: the jackknife and the parametric bootstrap. The jackknife estimator performed favourably compared to the variance estimate from the conventional line transect analysis for the count data model, but poorly for the interval data model. This result was not entirely unexpected, particularly given the low number of sampling units (transects) on which the jackknife was based.

The jackknife estimator operates correctly only if the estimated statistic (say $\hat{\theta}$) has a locally linear quality (Miller, 1974). Intuitively this means that $\hat{\theta}$ should exhibit 'smoothness', whereby small changes in the data cause only small changes in the value of the estimated statistic (Efron and Tibshirani, 1993). The jackknife estimator of abundance satisfies the smoothness criterion, perhaps explaining why it performs well in the case of the count data model, despite the low number of sampling units. Its poor performance in the case of the interval data model might therefore be better explained by looking at the structure of the model formulation itself. In this regard, the interval data may be thought of as a series of observations indicating some overall trend in density across a region. In practice, these observations are likely to be overdispersed, represented by some clustered observations which are serially correlated. Miller (1974) notes that the jackknife is rarely successful in such situations. Therefore in this paper, we discuss the merits of implementing the parametric bootstrap as a means of estimating variance from the spatial models.

The parametric bootstrap

Instead of resampling directly from the observations as in the non-parametric bootstrap, for the parametric bootstrap a model is fitted to the data from which new data values are then generated. Unlike the non-parametric bootstrap (and most analytical methods), the parametric bootstrap does not require the assumption of independently and identically distributed (IID) observations. As noted above, with both the count data model and the interval data model the observations are not independent, and in general will not be identically distributed. We describe an algorithm to generate parametric bootstrap resamples, noting that in practice, suitable models for estimating the pdf for detections are unlikely to be sufficiently flexible to fully incorporate the serial correlation between successive observations.

The first step of the parametric bootstrap is to fit the spatial model to the original data. Density at every point along the trackline may then be estimated from the fitted model. The pdf for detections along the line is obtained by dividing the estimated densities by their total integral along the line (which is calculated numerically). For each bootstrap pseudosample, the number of values to be generated from this pdf is a deviate from the Poisson distribution with rate E(n), approximated by n, the total number of detections. A rejection sampling method could be used to generate the values in the resamples. Thus, two variates from uniform distributions are generated - one (representing the surveyed effort) from a uniform distribution on (0, *total transect length*) and another (representing the variation in density along the transects) from a uniform distribution on (0, *maximum pdf value*). The two uniformly-distributed variates yield a point in two dimensions which, if it is located beneath the curve given by the pdf, is accepted in the pseudosample; otherwise it is rejected. Accepted values are then translated onto the transect line and their positions are calculated, enabling the number of schools in each segment (with segment boundaries remaining the same as for the original data) to be calculated for the count data model, or waiting distances to be calculated for the interval data model. This simple implementation has the advantage that observations do not need to be generated sequentially along the transect line. If the fitted model is sufficiently flexible to model any clustering in the data, so that, for example, high pdf values adequately represent locally high density clusters, then the implementation, and any consequential inferences, are entirely valid. If not, then inferences must take account of any unmodelled serial correlation neglected in the construction of the pseudosamples.

Given the bootstrap samples, the model selected for the original data is then refitted to obtain density and abundance estimates from each pseudosample. The sample variances of these estimates provide bootstrap estimates of the components of variance of $\hat{D}$ and $\hat{N}$ from the spatial modelling. The component of variance due to the estimation of $\hat{p}_{ij}$ in the count data model or effective strip width for the interval data model must also be incorporated to obtain the

Table 1

Stratum estimates of detection probability, p, of pods within the strip of half-width 1.5 n.miles, and effective strip half-width, $\hat{ESW}$, with coefficients of variation (calculated using DISTANCE software).

Pooled strata	p	%CV	$\hat{ESW}$(nm)	%CV
WN and EN	0.742	7.61	1.112	7.61
WS and ES	0.360	15.42	0.540	15.42

overall variance estimates of $\hat{D}$ and $\hat{N}$. We use the delta method (Seber, 1982, pp.7-9) to combine the components of variance in the estimation.

APPLICATION OF THE METHODS TO IWC/IDCR ANTARCTIC MINKE WHALE DATA

In this example, we apply the spatial modelling methods described above to Independent Observer (IO) mode minke whale data from the 1992-3 IWC/IDCR Antarctic Survey in Area III. Transects covered in this region are shown in Fig. 2, together with locations of sighted pods.

Count data model

Following Borchers and Cameron (1995), a conventional stratified analysis was conducted in DISTANCE (Laake *et al.*, 1993) to estimate the probability of detecting a minke whale pod within a truncation width w of the trackline. As in Borchers and Cameron (1995), data were truncated at 1.5 n.miles, and effective strip widths were calculated separately for the Northern strata (WN and EN) and the Southern strata (WS and ES). The results are shown in Table 1.

Effort legs were divided into segments of 16 minutes, or approximately 3 n.miles assuming a vessel speed of 11.5 knots. Therefore, with a truncation distance of 1.5 n.miles, the sampling units were approximately squares of side 3 n.miles, bisected by the trackline. The estimated number of minke whale pods in each segment, $\hat{N}_i$ was calculated according to Equation (1), with $\hat{p}_{ij}$ equal to 0.742 if the segment was located in one of the Northern strata, and 0.360 if it was in a Southern stratum.

A GAM with a logarithmic link function and overdispersed Poisson error distribution was fitted to the $\hat{N}_i$ to obtain a smooth density surface of minke whale pods throughout Area III (Fig. 4). Possible covariables were distance from the ice edge (*ice*), latitude (*lat*) and longitude (*lon*). Each of these was considered for inclusion in the model as a cubic smoothing spline with either 8, 4 or 2 degrees of freedom, or as a linear term. An automated stepwise procedure using a version of Akaike's Information Criterion (AIC) that adjusts for overdispersion for model selection was adopted (Chambers and Hastie, 1993, p.282). The final model was highly flexible, with all three covariables selected as smoothed terms with eight degrees of freedom, and is shown below:

$$E(\hat{N}_i) = \exp[\log(a_i) + \theta_0 + s(ice_i, 8) + s(lat_i, 8) + s(lon_i, 8)],$$

where as noted previously, the offset variable a_i is the area of the ith segment. The nonlinear form of the dependence of pod density on the covariables is shown in Fig. 3. In interpreting the plots in this figure, it is important to recognise that they show the additional effect of the covariable being plotted, *given* that the other (smoothed) covariables are included in the model. For example, the second peak in density in the smoothing spline of latitude at around 62°S does not correspond to the region of highest density in Fig. 4 because, as can be seen from the smoothing spline of distance from the ice edge, densities generally

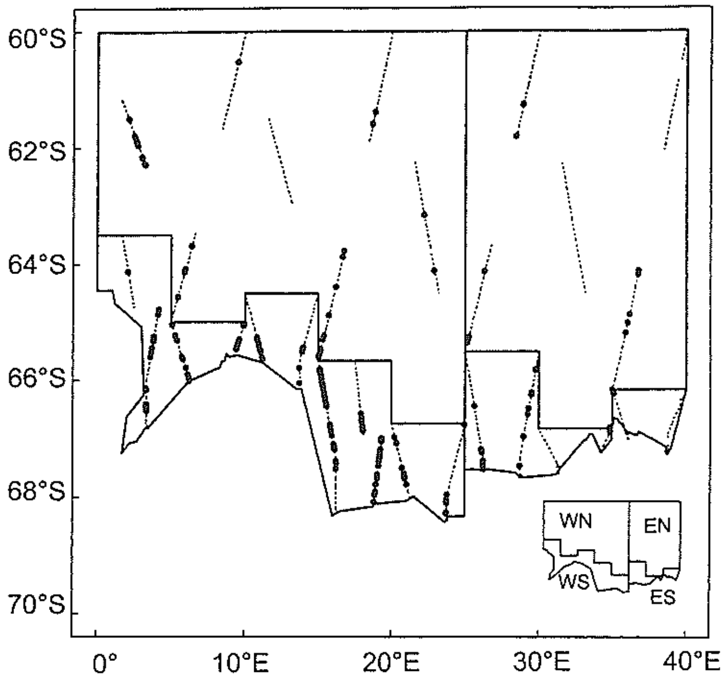

Fig. 2. Realised survey effort in IO mode and minke whale school sightings in Area III during the 1992-3 IWC/IDCR Antarctic Survey. Subplot shows the division of the region into four strata: WN, WS, EN and ES.

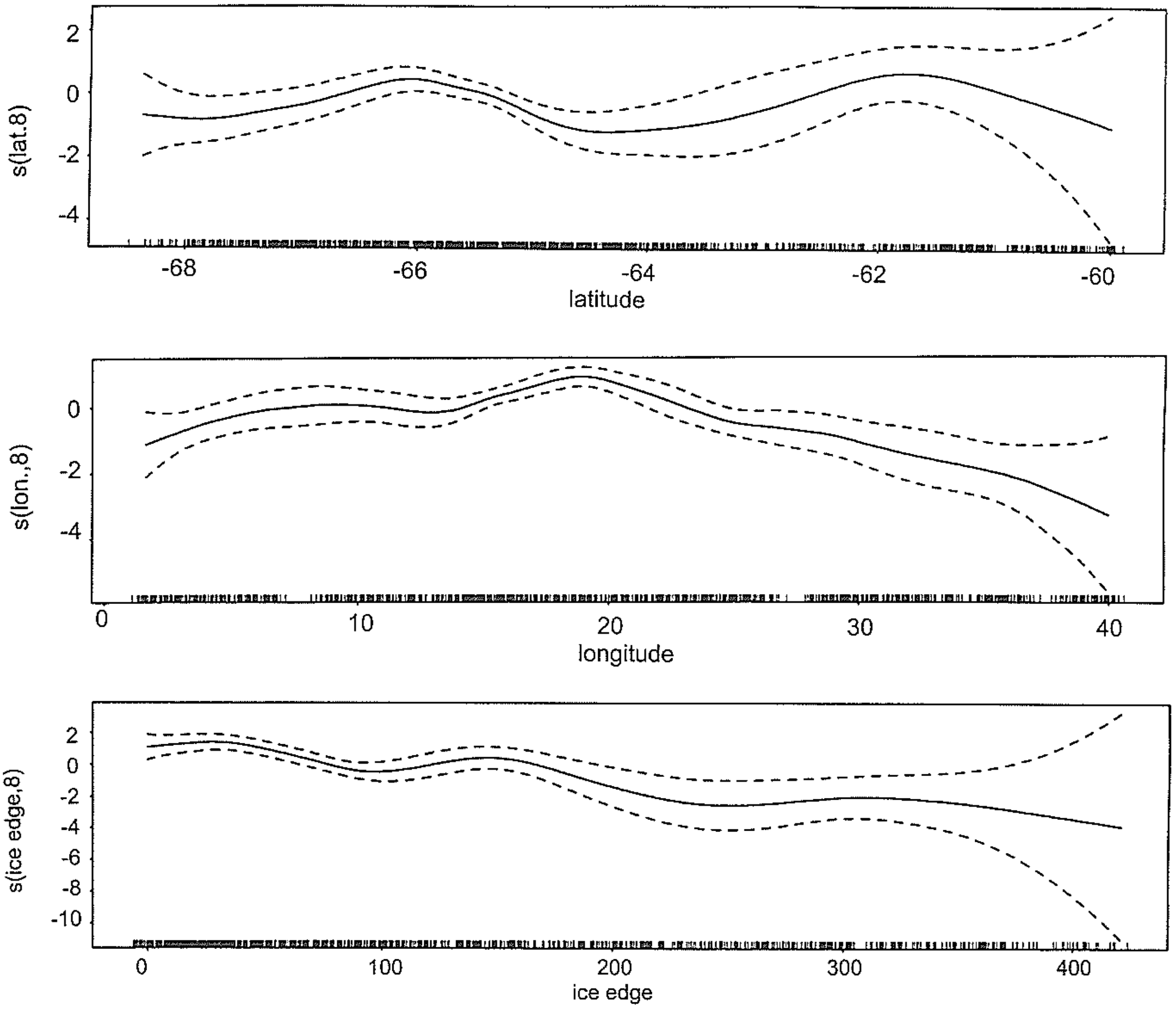

Fig. 3. Shapes of the functional forms for the smoothed covariables used in the count data model example. Zero on the vertical axes corresponds to no effect of the covariable on the estimated response (here, pod density). The locations of the observations are plotted as small ticks along the horizontal axes.

decrease with increasing distance from the ice edge, which was located considerably further south than (and hence at large distances from) 62°S during this survey.

Estimates of abundance by stratum are given in Table 2. Coefficients of variation were estimated using the parametric bootstrap. The predicted density surface of minke whale pods, as shown in Fig. 4, is quite well supported by the observed data, indicating that the model has provided a good description of the spatial variation in pod density. The highest density area is located around 66°S, 16°E. However, a moderately high density patch is also predicted to occur at around 20°E, just north of 66°S, on the interstratum boundary between the WN and WS strata which, since there were no transects in this region, is not apparent from the sightings data. It is easy to see from Fig. 3 why high densities are being predicted here. The middle plot, showing how density varies with longitude, displays a global peak at about 20°E, whilst the top plot, indicates a local peak at around 66°S. These two effects together with relatively high densities predicted from the distance from ice edge smoothed combine to produce this patch. (The centre of the patch is approximately 125 n.miles from the ice edge boundary). The 'truth' is of course unknown in this example, but the scenario serves to remind us of a possible pitfall when combining one-dimensional smoothers to produce a two-dimensional surface. An alternative is to model the surface directly using a bivariate smoothing function, although the increase in complexity can lead to difficulties in interpretation and computation (Hastie and Tibshirani, 1990).

Interval data model

Using the stratum estimates of effective strip half-width shown in Table 1, waiting areas were calculated as twice the estimated effective strip half-width multiplied by the distance along the trackline between consecutive sightings. A generalised additive model with a logarithmic link function was fitted to the estimated waiting areas, assuming a Gamma error distribution. As for the count data model, possible covariables were distance from the ice edge, latitude and longitude, and again these were considered for inclusion in the model as cubic smoothing splines with either 8, 4 or 2 degrees of freedom, or as linear terms. Stepwise automated model selection, based on the AIC (adjusted for overdispersion), led to the following final model:

$$E(\hat{W}_i) = \exp[\theta_0 + s(lat_i,8) + s(lon_i,2)] \qquad i = 1,\ldots,n$$

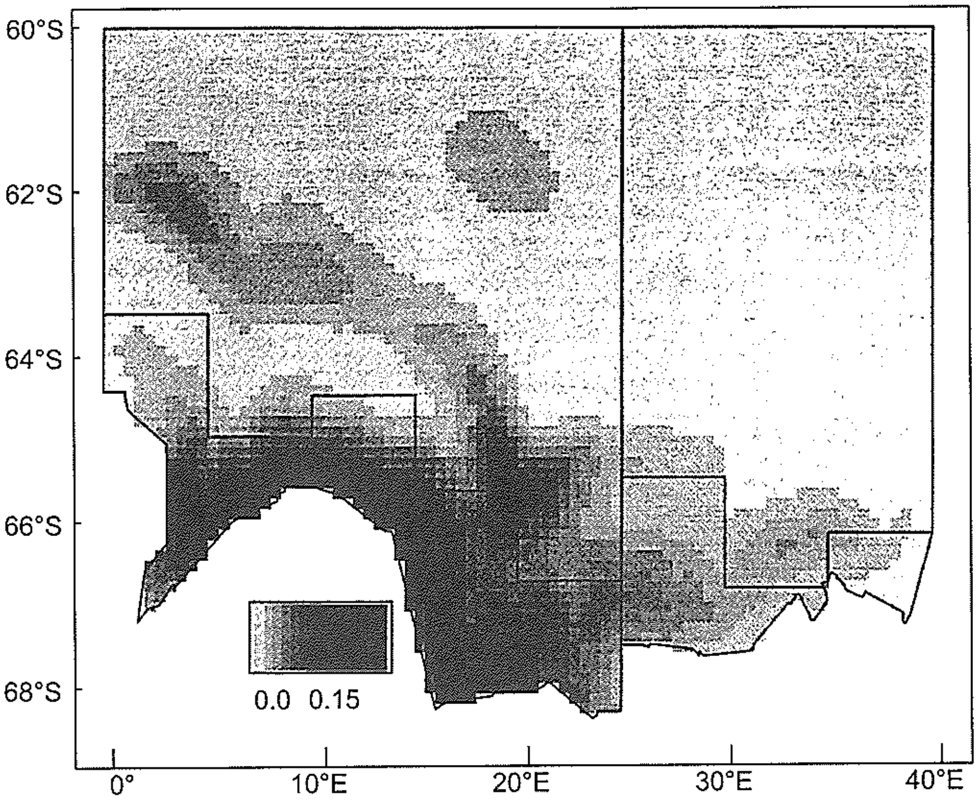

Fig. 4. Density of minke whale pods in Area III, estimated using the count data model.

where $\hat{W}_i$ is the estimated waiting area between sighting (i-1) and sighting i, and n is the total number of sightings. The failure to select the term representing the distance from the ice edge may at first seem counter to *a priori* expectations, but in this case the high flexibility of the smoothed term in latitude, coupled with the strongly smoothed longitudinal term, is sufficient to model the spatial variation in density. The non-linear form of the dependence of the estimated waiting areas on these two covariables is shown in Fig. 5, so for example, the increasing trend with decreasing latitude seen in the latitude smooth represents a decreasing trend in density (because expected densities are given by the reciprocal of the expected waiting areas).

Estimates of abundance by stratum with corresponding coefficients of variation (estimated using the parametric bootstrap) are given in Table 2. The predicted density surface of minke whale pods is shown in Fig. 6.

Discussion

Given the differences between the stratified analysis and the form of the two spatial methods presented in this paper, the point estimates of total Area III abundance are remarkably similar. Whilst the estimate of abundance in the WN stratum from the interval data model is higher than the corresponding estimates from the other two methods, the differences are small relative to the precision of the estimates. However it may be that in this case the model is a poor fit to the data, and influenced by the eleven sightings clustered on the westernmost transect of the WN stratum, fails to capture the more expected scenario that density decreases with distance from the ice edge. Perhaps this highlights the possible shortcomings associated with reliance on an automated model selection procedure – an alternative model might well have been selected had the choice been augmented with other available tools, such as graphical methods. However,

Table 2

Estimates of abundance of minke whale pods ($\hat{N}_s$) from a conventional stratified analysis (Borchers and Cameron, 1995) and the two spatial modelling approaches described in this paper. For the spatial models, the CVs were estimated using the parametric bootstrap (%CV$_{PB}$).

	Stratified analysis		Count data model		Interval data model	
Stratum	$\hat{N}_s$	%CV	$\hat{N}_s$	%CV$_{PB}$	$\hat{N}_s$	%CV$_{PB}$
WN	4,810	40.13	4,621	19.90	6,386	30.52
EN	1,460	49.53	819	28.94	1,058	31.24
WS	7,412	25.14	8,415	17.62	6,895	23.04
ES	636	44.25	885	25.06	686	29.10
Total	14,318	22.99	14,740	16.03	15,025	21.23

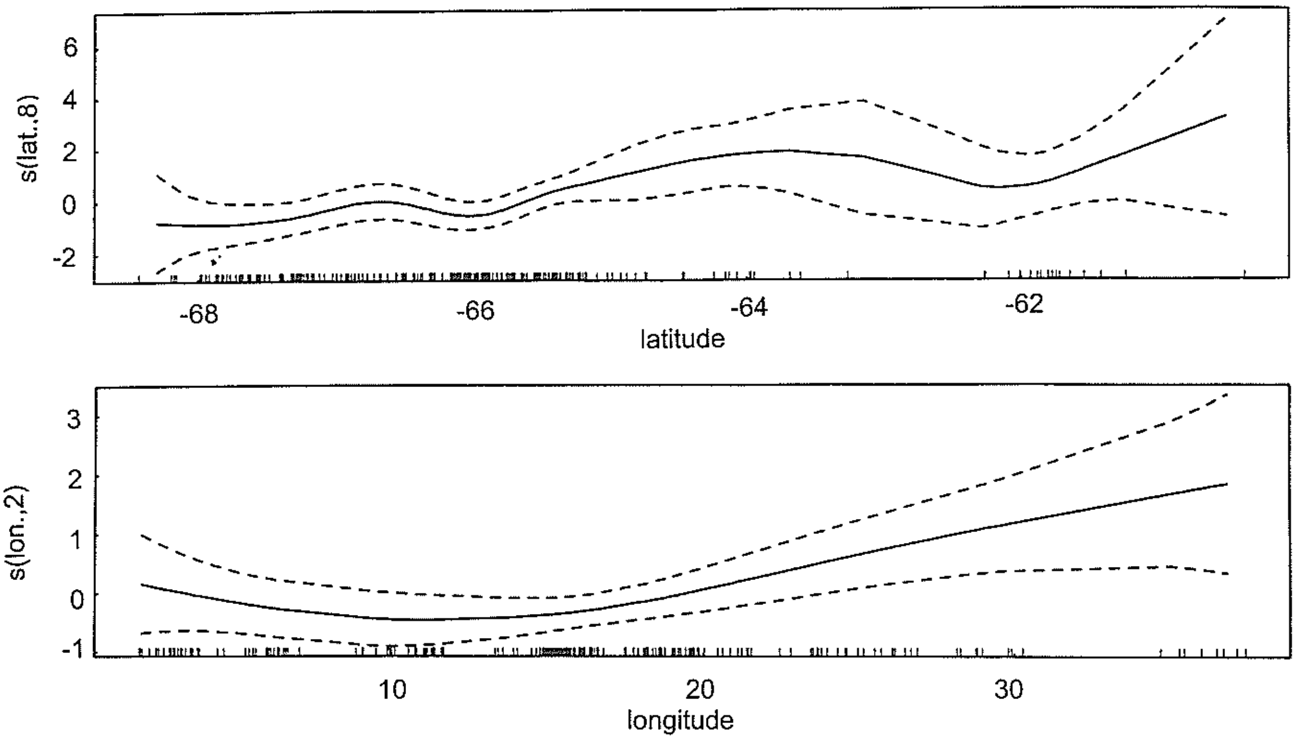

Fig. 5. Shapes of the functional forms for the smoothed covariables used in the interval data model example. Zero on the vertical axes corresponds to no effect of the covariable on the estimated response (which for this model is waiting area = [pod density] $^{-1}$). The locations of the observations are plotted as small ticks along the horizontal axes.

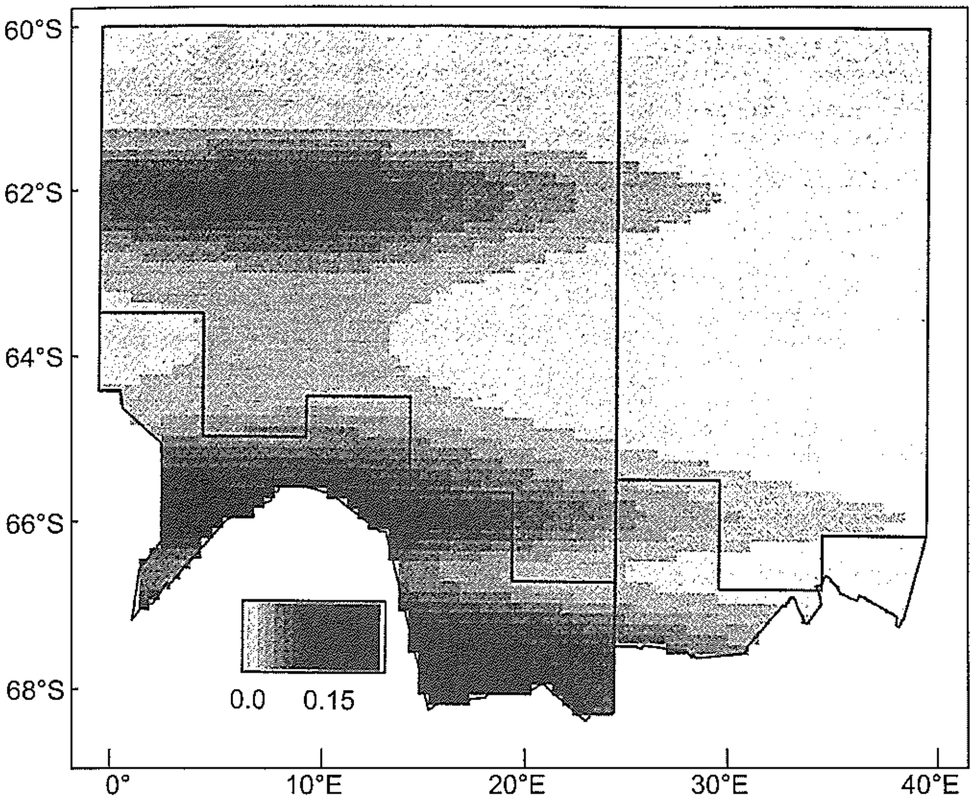

Fig. 6. Density of minke whale pods in Area III, estimated using the interval data model.

in this paper, we are seeking to demonstrate the methodology rather than select the 'best' model, so the automated procedure is adequate for these purposes. A better comparison of the methodologies would be gained from using simulated data to examine possible biases and further evaluate conditions under which the spatial models might be expected to perform favourably in comparison to a stratified analysis. Substantial progress in this regard has been made by Clarke *et al.* (1999), who in their simulation of survey data from the Japanese Whale Research Programme under Special Permit in the Antarctic (JARPA), compare conventional line transect estimates, corrected estimates that attempt to account for a bias in the JARPA survey design (Burt and Borchers, 1997), and estimates under the interval data model.

One of the potential advantages gained by using a spatial model to estimate density is an improvement in the estimated precision, since variation in density can be explained by relatively few spatial covariates. In this example, we applied a resampling technique, the parametric bootstrap, to estimate the precision of the estimates of abundance by stratum and of the total abundance in the area surveyed. The results from using this technique look extremely promising. For the count data model, appreciable improvements in estimated precision were obtained in both the stratum estimates and the overall abundance estimate compared to the analytical estimates from the stratified analysis. Some questions remain about the possible inducement of negative bias in these estimates, for example, by neglecting unmodelled correlation in constructing the bootstrap resamples, as demonstrated by the relatively low mean estimated dispersion parameter seen in the pseudosamples compared to the observed data (Table 3). A Monte Carlo test performed on the estimates in Table 3 indicated significant differences (at significance levels 0.02 and 0.04 for the count data and interval data models respectively) between the dispersion parameter estimates from the observed data and the means of the estimates from the two sets of pseudosamples. For the interval data model, there was no improvement in precision for the overall abundance estimate, but the precision of the stratum estimates was increased, most notably in the two strata with fewest sightings: the EN and ES strata. Model-based estimates can be expected to provide increased precision over stratified estimates in such circumstances, since conditional on an appropriate model, data from outside a stratum can still provide information relevant to the estimation of density within that stratum.

Table 3

Dispersion parameter estimates from the observed data and from the parametric bootstrap resamples.

	Dispersion parameter estimated from the observed data	Mean of the dispersion parameter estimates from 100 pseudosamples
Count data model	2.767	1.683
Interval data model	3.304	1.922

Model selection uncertainty can be readily incorporated into the variance estimator. Rather than conditioning on the original spatial model, the explanatory variables and their degree of smoothness would be independently selected for each bootstrap resample (e.g. see Buckland *et al.*, 1997). This will tend to inflate the variance of the estimators, and arguably better reflects their true variance, but for purposes of comparison with conventional line transect estimates of D and N, we do not incorporate such uncertainty here.

CONCLUSIONS

Although the unbiasedness of the parametric bootstrap method needs to be established before the associated variance estimator can be used with confidence, the spatial modelling methods presented here represent a very promising improvement over traditional line transect estimation methods in several respects:

(1) they provide a statistically sound means for estimating abundance at any spatial resolution (e.g. by *Small Management Area*, see IWC, 1999) with relatively high precision (because they exploit data from outside the small area);

(2) by modelling the spatial variation in density, they may provide higher precision for abundance estimation in the whole survey area than stratified estimation methods;

(3) they provide a powerful tool for relating cetacean distribution and abundance to spatial and other explanatory variables.

In addition, unlike conventional line transect analyses which rely on statistical sampling theory for the estimation, the spatial methods adopt a model-based framework. This is potentially advantageous because there is no requirement for random placement of transect lines, although it remains inadvisable to extrapolate the predicted density surface beyond the range of the region where the data are collected. In particular, provided the spatial coverage is reasonably representative within the region of interest, spatial models have potential value in modelling data from Platforms of Opportunity (see Bravington, 1999). The only substantial disadvantage of the methods relative to traditional stratified methods for estimating abundance is their complexity; in particular, model selection issues remain largely unresolved. However, the spatial modelling methods have useful applications beyond simply estimating abundance.

One of the long-term research objectives of the IWC-SOWER 2000 research programme is to relate spatial and temporal variability in oceanographic variables and prey distribution to cetacean distribution and abundance. As noted above, the methods presented in this paper provide a useful tool for addressing this objective. With sufficient coverage over a time period, it would be quite straightforward to incorporate a temporal component to assess changes in spatial distribution with time. Whilst this may result in useful inferences over time periods of less than a year, perhaps more useful inferences on inter-annual changes in distribution and abundance are likely to result from methods which integrate process models and survey data (e.g. Sullivan, 1992; Fewster, 1999). Spatial models like those presented here are likely to be an important component of such integrated models.

The spatial methods described have already been applied to several quite different datasets, yielding informative results. The count data methodology has been used to describe the spatial distribution of minke whales from an aerial survey off the coast of West Greenland (Hedley *et al.*, 1997), whilst applications of the interval data methodology include modelling Antarctic minke whale distribution from data collected on JARPA surveys (Clarke *et al.*, 1998), and modelling the distributions of harbour porpoise and minke whales in the North Sea and surrounding waters using data from the 1994 SCANS survey (Burt *et al.*, 1999). However, a fuller understanding of the performance of the methods would be achieved by using simulated data. Some questions that such research would attempt to address include testing the sensitivity of the count data model to segment length selection and testing the robustness of the models to the selection of smoothing parameters, together with further investigation of variance estimators.

Finally we note some problems in estimating individual, rather than group, abundance. Equation (2) gives a response variable which could be used in estimating individual abundance from the count data model, but the use of such a response in a spatial model will generally induce even more overdispersion than the Horvitz-Thompson estimator of Equation (1). Further, variance estimation becomes more difficult since any resampling algorithm should take into account spatial variation in school size. The estimation of individual abundance from the interval data model has not been addressed in this paper. One possible solution is to fit a spatial model to the observed group sizes and multiply the resulting surface by the group density surface given by the spatial model. However, if size bias is present, estimation of a group size surface in order to estimate absolute abundance is not straightforward. Bravington (1999) describes difficulties in modelling group size even when the objective is to obtain relative abundance estimates. One solution is to revert to estimating mean group size as in conventional line transect estimation, and simply multiply the group abundance estimates by the estimated mean group size to obtain estimates of the number of individuals. This has the disadvantage of not modelling spatial variation in group size (except, where sample sizes permit, by stratum), but it has the advantage of being able to utilise the methods that have been developed for accounting for size-bias in conventional line transect estimation, such as those described in Buckland *et al.* (1993, pp.125-135). Research is ongoing to examine whether group size could be incorporated directly into the interval data model framework, via marked point process modelling, but it appears that this is only a solution when certain somewhat unrealistic assumptions can be met. The models would benefit from further research aimed at the development of methods which are capable of modelling the group size surface separately from the group density surface, whilst taking account of any size bias.

ACKNOWLEDGEMENTS

We thank Fernanda Marques and Samantha Strindberg, both from the Research Unit for Wildlife Population Assessment, University of St Andrews, and reviewers Justin Cooke and Mark Bravington, for their insightful comments on earlier versions of this paper. We also would like to express thanks to the IWC for funding Sharon Hedley's contribution to this work.

REFERENCES

Augustin, N.H., Mugglestone, M.A. and Buckland, S.T. 1996. An autologistic model for the spatial distribution of wildlife. *J. Appl. Ecol.* 33:339-47.

Beavers, C.S. and Ramsey, F.L. 1998. Detectability analysis in transect surveys. *J. Wildl. Manage.* 62(3):948-57.

Borchers, D.L. and Cameron, C. 1995. Analysis of the 1992/93 IWC Minke Whale Sightings Survey in Area III. Paper SC/47/SH16 presented to the IWC Scientific Committee, May 1995 (unpublished). 16pp. [Paper available from the Office of this Journal].

Borchers, D.L., Buckland, S.T., Goedhart, P.W., Clarke, E.D. and Hedley, S.L. 1998. Horvitz-Thompson estimators for double-platform line transect surveys. *Biometrics* 54:1221-37.

Bravington, M. 1999. Modelling relative abundance and covariate effects from sightings data. Paper SC/M99/SOWER12 presented to the IWC Scientific Committee SOWER 2000 Workshop, March 1999, Edinburgh (unpublished). [Paper available from the Office of this Journal].

Buckland, S.T. and Elston, D.A. 1993. Empirical models for the spatial distribution of wildlife. *J. Appl. Ecol.* 30:478-95.

Buckland, S.T., Anderson, D.R., Burnham, K.P. and Laake, J.L. 1993. *Distance Sampling: Estimating Abundance of Biological Populations.* Chapman and Hall, New York and London. xii+446pp.

Buckland, S.T., Burnham, K.P. and Augustin, N.H. 1997. Model selection: an integral part of inference. *Biometrics* 53:603-18.

Burt, M.L. and Borchers, D.L. 1997. Minke whale abundance estimated from the 1991/92 and 1992/93 JARPA sighting surveys. Paper SC/M97/23 presented to the IWC Intersessional Working Group to Review Data and Results from Special Permit Research on Minke Whales in the Antarctic, May 1997 (unpublished). 16pp. [Paper available from the Office of this Journal].

Burt, M.L., Hedley, S.L., Borchers, D.L. and Buckland, S.T. 1999. Spatial modelling of data from Project SCANS. Internal RUWPA report (unpublished). [Available from the author].

Chambers, J.M. and Hastie, T.J. 1993. *Statistical models in S.* Chapman and Hall, New York. 608pp.

Clarke, E.D., Ashbridge, J., Burt, M.L., Hedley, S.L. and Borchers, D.L. 1998. GAM-based abundance estimation from JARPA survey data: progress and simulation model design. Paper SC/50/CAWS33 presented to the IWC Scientific Committee, April 1998 (unpublished). 51pp. [Paper available from the Office of this Journal].

Clarke, E.D., Burt, M.L. and Borchers, D.L. 1999. Simulation of JARPA surveys to test abundance estimation methods. Paper SC/51/RMP16 presented to the IWC Scientific Committee, May 1999, Grenada, WI (unpublished). 17pp. [Paper available from the Office of this Journal].

Cooke, J.G. and Leaper, R. 1998. A general modelling framework for the estimation of whale abundance from line transect surveys. Paper SC/50/RMP21 presented to the IWC Scientific Committee, June 1998 (unpublished). 16pp [Paper available from the Office of this Journal].

Cumberworth, S.L., Buckland, S.T. and Borchers, D.L. 1996. A spatial modelling approach for the analysis of line transect data. Paper SC/48/O 12 presented to IWC Scientific Committee, June 1996, Aberdeen (unpublished). 11pp. [Paper available from the Office of this Journal].

Efron, B. and Tibshirani, R.J. 1993. *An Introduction to the Bootstrap.* Chapman & Hall, New York. 436pp.

Fewster, R.M. 1999. Population dynamics models in a spatial oceanographic setting. Paper SC/M99/SOWER2 presented to the IWC Scientific Committee SOWER 2000 Workshop, March 1999, Edinburgh (unpublished). [Paper available from the Office of this Journal].

Hastie, T.J. and Tibshirani, R.J. 1990. *Monographs on Statistics and Applied Probability.* 43. *Generalized Additive Models.* Chapman & Hall, London. 335pp.

Haw, M.D. 1993. Estimation of minke whale abundance from the 1990/91 IWC/IDCR Antarctic assessment cruise in Area VI. Paper SC/45/SHBa1 presented to the Scientific Committee, April 1993 (unpublished). 33pp. [Paper available from the Office of this Journal].

Hedley, S., Barner Neve, P. and Borchers, D.L. 1997. Abundance of minke whales off West Greenland, 1993. Paper SC/49/NA7 presented to the IWC Scientific Committee, September 1997, Bournemouth (unpublished). 11pp. [Available from the Office of this Journal].

Horvitz, D.G. and Thompson, D.J. 1952. A generalisation of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47:663-85.

International Whaling Commission. 1994. Report of the Scientific Committee, Annex E. Report of the sub-committee on Southern Hemisphere baleen whales. *Rep. int. Whal. Commn* 44:93-107.

International Whaling Commission. 1999. Report of the Scientific Committee. Annex N. The Revised Management Procedure (RMP) for Baleen Whales. *J. Cetacean Res. Manage. (Suppl.)* 1:251-8.

International Whaling Commission. 2000. Report of the Intersessional Planning Meeting to Discuss Research on Climate Change and Cetaceans. *J. Cetacean Res. Manage. (Suppl.)* 2:In Press.

Laake, J.L., Buckland, S.T., Anderson, D.R. and Burnham, K.P. 1993. *Distance User's Guide, Version 2.0.* Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, CO. 72pp.

McCullagh, P. and Nelder, J.A. 1989. *Monographs on Statistics and Applied Probability.* 37. *Generalized Linear Models.* 2nd Edn. Chapman & Hall, London. 511pp.

Miller, R.G. 1974. The jackknife - a review. *Biometrika* 61:1-17.

Osborne, P.E. and Tigar, B.J. 1992. Interpreting bird atlas data using logistic models: an example from Lesotho, Southern Africa. *J. Appl. Ecol.* 29:55-62.

Ramsey, F.L., Wildman, V. and Engbring, J. 1987. Covariate adjustments to effective area in variable-area wildlife surveys. *Biometrics* 43:1-11.

Seber, G.A.F. 1982. *The Estimation of Animal Abundance and Related Parameters.* 2nd Edn. Edward Arnold, London. xvii+654pp.

Sullivan, P.J. 1992. A Kalman Filter approach to catch-at-length analysis. *Biometrics* 48:23757.