Review of density surface modelling applied to JARPA survey data

ML BURT and CGM PAXTON
RUWPA, University of St Andrews, The Observatory, Buchanan Gardens, St Andrews, KY16 9LZ, UK

ABSTRACT

Standard line transect analyses of JARPA survey data may result in biased estimates of minke whale abundance because in high density areas more time is spent confirming and/or sampling schools and hence survey effort is reduced. Density surface models (DSM) provide an alternative method of estimating abundance without relying on the survey design *per se*. However, these models do rely on the correct model being fitted with suitable explanatory variables and good coverage of survey effort throughout the study region. Analyses are further complicated by complex irregular coastlines and predictions are sensitive to extrapolation. However, these issues are being addressed by various research groups and it is anticipated that future methodological developments in model fitting would deal better with JARPA data than the methods currently being implemented.

INTRODUCTION

The Japanese Whale Research Program under Special Permit in the Antarctic (JARPA) has been conducting line transect surveys of minke whales in Antarctic waters every austral summer since 1987/1988 – the first two years being feasibility surveys (Nishiwaki *et al*., 2005). Estimates of abundance based on these data are necessary for the estimation of biological parameters associated with population dynamics. In addition, the relatively long time series of data available may provide useful information on the status of minke whale stocks in the Antarctic, and hence abundance estimates from these data may be of interest in other contexts.

JARPA surveys were designed to ensure that the available survey area was covered spatially within an allocated time period, regardless of the conditions encountered. A pre-determined amount of effort was allocated for each day of the survey and if a vessel did not complete the allocated effort at the end of any given day, then it moved off-effort until it reached the starting point for the next day. In areas of high densities of whales, the realised survey effort was reduced as more time was spent confirming and/or sampling whale schools (IWC, 1998; Nishiwaki *et al*., 2005). Standard line transect (LT) analyses (Buckland *et al*., 2001) assume that survey effort is located independently of density, thus such analyses of JARPA data may result in biased estimates of abundance. LT estimators conventionally use design-based methods to draw inferences about the whole survey region from estimates within the searched strips.

The extent of the bias in abundance estimates due to the JARPA survey protocol depends on the degree of clustering of the whales as well as whale density. A simulation study assessed the performance of three alternative approaches to estimate abundance; the approaches considered were the 'count model' and the 'waiting-time model' developed by Hedley *et al*. (1999) and the standard LT model. This study showed that the 'count model' performed best, with no, or only small, bias evident with appropriately chosen levels of smoothing (Clarke *et al*., 2000). However, it was not clear when using these methods on real data, how best to choose the appropriate level of smoothing and using inappropriate levels of smoothing could lead to substantially biased estimates of abundance. The count method involves fitting a spatially referenced density surface to counts of schools in small areas covered by search effort. The density surface estimator is a model-based estimator and it relies on the correct model being fitted to the data. In standard LT methodology, stratum-specific density estimates are obtained whereas the density surface model (DSM) allows density to be a function of location and environmental variables.

This paper reviews the DSM approach and reflects on previous applications to the JARPA data.

SURVEY DESIGN

Brief details of the JARPA survey design are given here with full details in Nishiwaki *et al*. (2005). The main JARPA survey region encompasses the International Whaling Commission (IWC) Management Areas IV (70^{o}-130^{o}E) and V (130^{o}E-170^{o}W), south of 60^{o}S, with each Area being surveyed in alternate years. Although the whole research period ranged from the end of November to the middle of March, the majority of the research was conducted in January and February (which coincided with the IDCR/SOWER surveys).

The survey regions were divided into four strata with a north-south boundary at approximately 45 nmiles from the ice edge and an east-west boundary at 100^{o}E and 165^{o}E for Areas IV and V, respectively. Prydz Bay stratum, in Area IV, was defined as being south of 66^{o}S and the Ross Sea stratum, Area V, was defined as south of 69^{o}S. Ice conditions dictated the extent of the size and extent of the strata, particularly for Prydz Bay and the Ross Sea strata. In the northern and Ross Sea strata a zigzag trackline was used and a saw-tooth shape used in the southern strata. Two diagonally connected tracklines creating a 'z' shape were used in Prydz Bay. A pre-determined distance was allocated for each day

of the survey and if a vessel didn't complete the allocated distance at the end of any given day, the vessel moved off effort until it reached the start point for the following day. This 'skip' in the searching distance in a day may have been due to poor weather conditions and/or to sampling activity in areas of high density of whales. The skip was adopted for the first six seasons (1987/88 – 1992/93) but was abolished thereafter.

During the surveys, two or three sighting and sampling vessels (SSVs) followed parallel transect lines, at fixed distances from each other. All sightings made while the vessels were on-effort (primary sightings) were recorded. Sightings made within 3nm of the transect line were approached by the vessel to confirm group size and species and to sample individuals from the group. Sightings made while the vessel was confirming and/or sampling were considered secondary sightings, and were not included in the analyses. After confirmation or sampling was completed, the vessel returned on-effort to the transect line at a 45º angle relative to the transect line. If during confirmation or sampling the vessel moved beyond 3nm from the line, it returned from the most advanced point reached during confirmation or sampling to the transect line at a 90º angle, and no search effort took place until it reached the transect line.

A dedicated sighting vessel (SV) was introduced in 1991/1992. The SV travelled at least 12 nm ahead of the SSVs so that it was unaffected by sampling activities by the SSVs (Nishiwaki *et al*., 2005). The survey procedure by the SV consisted of closing mode (the vessel approached schools to confirm species and group size) and also passing mode (continuous search effort, with schools not being approached). In later years, the SSVs also operated in passing mode. Thus, vessels were either classed as SSV or SV and could operate in both closing and passing modes.

ANALYSIS METHOD

There are essentially four components to the DSM methodology; 1) fitting a detection function, 2) modelling school density, 3) estimating school size and 4) estimating variance. Each component is described below, along with further modifications of this basic DSM approach.

Detection function estimation using MCDS

Explanatory variables which influence detectability, in addition to perpendicular distance, are included in the detection function using the multiple covariate distance sampling (MCDS) approach developed by Marques (2001). This was achieved by setting the scale parameter in the detection function model to be an exponential function of the covariates (Marques, 2001; Marques and Buckland, 2003). In this way, a single model can be fitted with strata and vessel survey mode as the explanatory variables for example, rather than fitting separate detection functions to each strata/survey mode combination as in a conventional stratified LT analysis. The MCDS methods parameterise only the detection function scale parameter as a function of explanatory variables, the explanatory variables do not affect the shape of the detection function. Using the fully stratified LT method, both the scale and shape parameters can change because separate models can be fitted to each stratum/mode combination. Two forms of detection function were considered; the hazard-rate and the half-normal models and AIC was used to choose between different models.

Spatial modelling of school density

The 'count model' of Hedley *et al*. (1999) was used to model the trend in spatial distribution of minke whale schools. The response variable was the number of minke whale schools in segment *i*, $\hat{N}_i$, estimated using the Horvitz-Thompson estimator (Horvitz and Thompson, 1952):

$$\hat{N}_i = \sum_{j=1}^{n_i} \frac{1}{\hat{p}_{ij}}, \qquad i = 1, \ldots, \nu, \tag{1}$$

where n_i is the number of schools detected in segment *i*, $\hat{p}_{ij}$ is the estimated probability of detection of school *j* in segment *i*, obtained from the fitted model for the detection function described above, and ν is the total number of segments (i = 1, …, ν). To model the response as a function of spatial covariates, we used a GAM with spatially referenced covariates, with the following general formulation.

$$E[\hat{N}_i] = \exp\left[\ln(a_i) + \beta_0 + M_i + \sum_{k=1}^{q} f_k(z_{ik}) \right] \tag{2}$$

Here a_i is an offset (parameter with known regression coefficient) that corresponds to the area of the *i*th segment, β_0 denotes the intercept and the f_k are one-dimensional smooth functions (cubic smoothing splines) of the q spatial covariates **z**. Other variables are easily incorporated into this framework; for example, vessel survey mode was incorporated as a factor variable and denoted by M_i. Inclusion of a factor variable in this way created parallel predicted density surfaces. Two-way interactions between the spatially referenced covariates were also considered for inclusion in the model via two-dimensional smooths (Wood, 2003). In particular, a two-dimensional smooth of latitude and longitude intuitively has more appeal than the sum of the one-dimensional effects.

The formulation shown in equation (2) assumed a logarithmic link function for the GAM. An appropriate form for the variance-mean relationship must also be selected and a quasi-likelihood formulation equivalent to an overdispersed

Poisson distribution was used so that the scale parameter was estimated as the constant of proportionality between the variance and the mean of the observations.

Estimation of smoothing parameters in GAMs, as implemented in the software R (Ihaku and Gentleman, 1996) through the mgcv package (Wood, 2001), was done using Generalised Cross Validation (GCV). Model selection was made on the basis of the lowest GCV score and diagnostic plots.

School size estimation

The expected school size was estimated by stratum. It was calculated as the ratio of the Horvitz-Thompson estimate (Horvitz and Thompson, 1952) of the abundance of individuals to the Horvitz-Thompson estimate of the abundance of schools:

$$E(s_t) = \frac{\sum_{j=1}^{n_t} \frac{s_{tj}}{\hat{p}_{tj}}}{\sum_{j=1}^{n_t} \frac{1}{\hat{p}_{tj}}}. \quad (3)$$

Here $E(s_t)$ indicates the expected school size, with the subscript t identifying the stratum. The parameter s_{tj} is the observed size of the jth detected school in stratum t, with n_t corresponding to the total number of detected schools in that stratum. The term $\hat{p}_{tj}$ is the estimated detection probability of the jth detected school in stratum t, obtained from the fitted model for the detection function. As in the stratified LT method, only schools with confirmed school sizes in closing mode were included in the school size estimation. Estimates of expected school sizes were obtained separately for SSV and SV mode sightings and also for all closing mode sightings (i.e. SSV+SV sightings). These school sizes were applied to passing mode data to obtain individual abundance from passing mode.

Variance estimation

Variances were estimated using a non-parametric bootstrap (Hedley and Buckland, 2004) which combined the three elements of the modelling; detection function estimation, density surface fitting and school size estimation. Having selected the terms of each element of the modelling using the original data, the whole modelling procedure was repeated on the bootstrap resample conditional upon these terms being in the model. The bootstrap resamples were generated using 'day' as the sampling unit, resampling with replacement.

Alternative DSM formulations

This general framework was readily adapted to different formulations of the model, thus providing additional flexibility should the nature of the problem and data dictate. Two alternative formulations of the basic methodology are described below.

Multi-stage modelling of school density

Instead of modelling density as shown in equation (2), the predicted densities of schools can also be obtained in three stages; first, fitting a logistic regression model to the presence/absence of schools; secondly, using a GAM to model the non-zero school densities; and then the product of the resultant surfaces from these previous steps gives a predictive map of school densities. For details see Paxton *et al*. (2006)

Modelling density directly

In equation (2), the response variable is $\hat{N}$ with the area of the segment being taken into account in the model as an offset term. Density can be modelled directly by rearranging equation 2 as follows (again assuming a log-link formulation).

$$E\left[\frac{\hat{N}_i}{a_i}\right] = E[\hat{D}] = \exp\left[\beta_0 + M_i + \sum_{k=1}^{q} f_k(z_{ik})\right]$$

APPLICATION TO JARPA DATA

Detection function estimation

The probability of detection was likely to be dependent on many things and in their analyses of survey data from Area IV, Marques *et al*. (2003) considered school size, sighting cue, Beaufort sea state, survey mode, stratum and vessel in the detection function model. Of these they found school size to be the most important. Beaufort sea state, sighting cue and, to a lesser extent, vessel mode were also included in the models. In contrast, Burt *et al.* (2005) restricted the potential covariates used in the detection function model for Area V data to strata and vessel survey mode in order to compare results more easily with estimates obtained from the standard LT estimates of Hakamada *et al.* (2005). The

effective strip half-widths (esw) from the two methods (LT and MCDS) were generally similar although some differences did occur when the number of sightings in a stratum/survey mode combination was small.

Density surface modelling

The geographic coordinates, latitude and longitude, and distance from ice edge were used as explanatory variables in the DSM. The location of the ice edge was based on the position of the ice edge that was carefully charted by the ships' officers while the vessels were in the southern strata. The fitted spatial models for Area IV comprised of smooth functions of the geographic coordinates and described a general decreasing trend in density with distance from the ice edge.

Analysis of the Area V data proved less straightforward which was due in part to the more complex nature of the ice edge. Given the similarities in the survey region and methodology between years it was suggested that a more robust estimate may be achieved by combining all surveys together and including year as an additional explanatory variable (IWC, 2005). In Paxton *et al*. (2006) survey data in Area V was combined and a single model was fitted with survey year being incorporated into the model as an additional explanatory variable. However, the DSM could not simply be extended to the combined dataset because of computational and model fitting problems due to its large size and distributional properties (i.e. a large number of zeros). Therefore, the multi-stage modelling approach was used. However, this combined model was not entirely satisfactory. Unsurprisingly, the models fitted to individual survey years were generally simpler and also explained more of the variation in the data. More surprising was the failure of the combined model to model whale presence in narrow regions of open water and this may have been because the limited set of explanatory variables was inadequate to model substantial changes which occurred in the study region from year to year. Large errors associated with the multi-stage models were a result of combining two surfaces which each contained large estimated values; both surfaces were sensitive to the omission of data points which occurred in the bootstrap samples.

School size estimates

In Burt *et al*. (2005) school sizes were obtained separately for each stratum and vessel mode. Bearing in mind that school size estimates in SV closing mode were sometimes based on small number of sightings, the estimates tended to be higher for SV sightings than for SSV sightings although it is not clear that this tendency is significant or not. This was thought to occur because larger schools were missed during the sampling activities of the SSV rather than the SV missing more smaller schools than the SSVs. School size is an important influence on the estimates of whale abundance and a substantial difference between the abundance for the different survey modes was due to the differences in school size.

Variance estimation

The DSM has the potential to produce smaller estimates of variance than the standard LT approach because more of the variation in encounter rate is being explained by the model. In practice, the variance estimation procedure can give extreme bootstrap samples resulting in unrealistic bootstrap abundance estimates. These extreme values can then have a large influence on the coefficient of variation although the 95% 'percentile confidence limits (using the lower and upper quartiles of the distribution) are less affected by extreme values.

DISCUSSION

There is no doubt that the DSM approach provides a flexible structure with which to estimate abundance when design-based estimators may result in biased estimates. It also provides a framework to explore the underlying environmental factors that drive animal density and distribution. Their utility is compromised if inadequate explanatory variables are used and this application was limited by the lack of suitable explanatory variables. In these applications to JARPA data, latitude and longitude served as proxies for environmental and biological variables that were not available and so cannot be expected to fully explain the complex variability in minke whale distribution. Also, the ice edge could change substantially throughout the time period of the survey and this would affect not only the effectiveness of the distance from ice edge variable, but also the size of the study region. This could have consequences for prediction especially if densities varied as a consequence of a currently unmeasurable interaction of ice edge position with other variables such as the shelf edge. Data on the latter may be available but data on more transient variables (e.g. presence of krill) may not be available.

Some of the computational and modelling difficulties associated with combining all surveys together and fitting a multi-stage model could be alleviated. Using longer segment lengths, possibly varying the segment lengths by strata and appropriately weighting the observations, would reduce the size of the dataset and thus a single-stage model could be used. Other types of smoothers designed specifically for complex topographical regions may improve model fits, particularly for the complex survey boundaries of Area V. This is an area of ongoing research by various research groups throughout the world.

Although the models do not rely on the design of the survey *per se*, they do require good coverage throughout the study region. The extension of the prediction grid to 60^{o}S where, in most years, there was little search effort was a source of

variation in the bootstrap sample. GAMs are susceptible to edge effects and extrapolation (especially on a logarithmic scale) into a region where there is little, or no, data can generate unrealistic predicted values, both high and low.

Despite the limitations the models did appear to have captured some of the spatial variation but it may be worth exploring the data set within a GAM framework that deals explicitly with spatial correlation in the error structure of the model.

By applying the methods to JARPA data, problems in the spatial models were identified and have been improved. In this sense, JARPA contributed to the development of the spatial modelling.

CONCLUSIONS AND RECOMMENDATIONS

1. DSM methods provide the means to correct for the non-random sampling design.
2. Obtaining approximately unbiased estimates requires that:
   a) the degree of smoothing is estimated correctly;
   b) the methods can deal with irregularly shaped coastlines and
   c) the methods prevent unrealistic behaviour at the edges of the sampled region.
3. However,
   a) with non-independent sampling units (segment of effort), model selection tools which assume independence cannot be replied upon and so the degree of smoothing may be estimated incorrectly.
   b) the present methods don't deal well with complex irregular coastlines and
   c) can go wrong at the edges of the study region (particularly when bootstrapping).
4. Wood *et al.* (in prep) are developing methods to deal with the problems outlined in point 3. Given this we feel it wise to delay further work on DSM methods to deal with sampling from JARPA until their methods are developed sufficiently to be applied to these data. The prospects for robust DSM-based estimates from these data using the new methods are better than with the currently available methods.

REFERENCES

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L. (2001) *Introduction to distance sampling: estimating abundance of biological populations.* Oxford University Press, London. 432pp.

Burt, M.L., Hedley, S.L. Hakamada, T. and Matsuoka, K (2005) Comparison of abundance estimates of JARPA survey data in Area V from standard line transect analysis and density surface fitting *Paper SC/57/IA18 presented to the Scientific Committee of the International Whaling Commission, June 2005, Ulsan (unpublished).* 13 pp.

Clarke, E.D., Burt, M.L. and Borchers, D.L. (2000) Investigation of bias in GAM-based abundance estimation methods and their suitability for JARPA survey data. *Paper SC/52/IA19 presented to the Scientific Committee of the International Whaling Commission, June 2000, Adelaide.* 15pp. (unpublished)

Hakamada, T., Matsuoka, K., Nishiwaki, S. (2005) *An update of Antarctic minke whales abundance estimate based on JARPA data including comparison to IDCR/SOWER estimates*. Paper JA/J05/PJR4 presented to JARPA review meeting January 2005, Tokyo. (unpublished)

Hedley, S.L. and Buckland, S.T. (2004). Spatial models for line transect sampling. *Journal of Agricultural, Biological, and Environmental Statistics.*9:181-199

Hedley, S.L., Buckland, S.T. and Borchers, D.L. (1999) Spatial modelling from line transect data. *Journal of Cetacean Research and Management* 1(3): 255-264.

Horvitz, D.G. and Thompson D.S. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47: 663-685

Ihaku, R., and Gentleman R. (1996) R: A Language for data analysis and graphics. *Journal of Computational and Graphical Structures* 5,3 299-314

IWC (2005) Remarks in Annex G of the Meeting of the IWC Scientific Committee 2005.

Marques, F.F.C. (2001) *Estimating wildlife distribution and abundance from line transect surveys conducted from platforms of opportunity*. PhD Dissertation, University of St Andrews, St Andrews, Scotland. 157pp.

Marques, F.F.C. and Buckland, S.T. (2003) Incorporating covariates into standard line transect analyses. *Biometrics* 59:924-935

Marques, F.F.C. Hedley, S.L., Hakamada, T. and Matsuoka, K. (2003) Spatial modeling of JARPA survey data in Area IV. *Paper SC/55/IA3 presented to the Scientific Committee of the International Whaling Commission, Berlin 2003 (unpublished).* 25 pp.

Nishiwaki, S., Ishikawa, H. and Fujise, Y. (2005) Review of general methodology and survey procedure under the JARPA. *Paper JA/J05/JR2 presented to the JARPA Review meeting, Tokyo, 2005*. Unpublished.

Paxton, C.G.M., Burt, M.L., Hakamada, T. and Matsuoka, K (2006) Spatial modelling of JARPA survey data in Area V: fitting all years in a single model. *Paper SC/58/IA20 presented to the Scientific Committee of the International Whaling Commission, May 2006, St Kitts and Nevis (unpublished).* 17 pp.

Wood, S.N. (2001) mgcv: GAMs and generalized ridge regression for R. *R News* 1: 20-25.

Wood, S.N. (2003) Thin-plate regression splines. *Journal of the Royal Statistical Society (B)* 65: 95-114.

Wood, S.N., Bravington, M.V. and Hedley, S.L. (in prep) Soap film smoothing. *Journal of the Royal Statistical Society*