Technical Reports of the Institute of Cetacean Research (2021) pp. 43–48

Technical Report (not peer reviewed)

An outline of the Density Surface Model for estimating abundance of baleen whales undertaken by the Institute of Cetacean Research

Takashi HAKAMADA*

Institute of Cetacean Research, 4–5 Toyomi-cho, Chuo-ku, Tokyo 104–0055, Japan

**Contact e-mail: hakamada@cetacean.jp*

ABSTRACT

The Institute of Cetacean Research (ICR) is starting the use of the Density Surface Model (DSM), a model-based approach, to estimate abundance of large baleen whales. This is an alternative approach to the design-based approach used so far for the same purpose. This paper explains the concepts of DSM, and outlines the potential utility of this approach for the work on whale abundance undertaken by the ICR.

INTRODUCTION

The Institute of Cetacean Research (ICR) conducts regular sighting surveys of large whales with the aim of estimating abundance in oceanic regions of interest. The basic sighting procedures for this purpose were described in Hakamada and Matsuoka (2017). Abundance has been estimated routinely by the 'design-based approach,' using the distance sampling method (Thomas *et al.*, 2010). Conventionally, abundance is estimated by equation (1) (Buckland *et al.*, 1993; 2001).

$$\frac{An\mathrm{E}(s)}{2wL} \quad (1)$$

where,
A is Area size of the survey area,
n is the number of detected whale schools,
E(*s*) is the expected mean school size,
w is the effective half search width (esw), and
L is the searching distance.

The validity of the abundance estimator based on equation (1) relies on the assumption of a randomized survey design (Hedley and Buckland, 2004; Miller *et al.*, 2013). In the case that this assumption is violated, an abundance estimator not requiring a randomized design is necessary.

Distribution of whales may be driven by environmental covariates such as sea surface temperature (SST), depth and salinity, each of which could vary in space. Relationships between environmental covariates and distribution of specific species is one of the interesting topics to investigate. Spatial modelling of the whale distribution can be constructed as a function of environmental covariates if dependency of the covariates on density of whales is substantial. Such constructed spatial modelling can be used to predict distribution in areas where the environmental covariates are available. Investigation of environmental covariates related to the distribution of the species can be conducted by selecting covariates using statistical criterion.

The two aims of spatial modelling are i) estimation of overall abundance in the area of interest ('model-based approach'), and ii) investigation of the relationship between density/abundance and environmental covariates (Miller *et al.*, 2013). Unlike conventional distance sampling ('design-based approach'), abundance estimates obtained by spatial modelling do not rely on the survey design. Spatial modelling can be applied to data obtained from different sources: platforms of opportunity such as ferries, fishing boats, incomplete surveys due to bad weather or accident, and surveys that were designed neither randomly nor systematically.

There are many kinds of spatial modelling. Density Surface Model (DSM) (Miller *et al.*, 2013) is one of them, and its use is starting to be utilized at the ICR. The objective of this paper is to explain the basic concepts of DSM based on Generalized Additive Model (GAM) (Wood, 2006) in the context of abundance estimates of large baleen whales. A secondary objective of this paper is to outline the future applications of the DSM in the work on whale abundance by the ICR.

GENERALIZED ADDITIVE MODEL (GAM)

DSM is expressed using the statistical model GAM. GAM allows the assumption that distribution of response vari-

ables is an 'exponential family.' A distribution belongs to the exponential family if its probability density function or probability mass function can be expressed by the formula

$$f(\theta, y) = \exp\left[\frac{\{y\theta - b(\theta)\}}{a(\phi)} + c(y, \phi)\right] \quad (2)$$

where a, b and c are arbitrary functions, ϕ is an arbitrary scale parameter and θ is known as the canonical parameter of the distribution (Wood, 2006). Properties and examples of exponential families are provided in McCullagh and Nelder (1989) and Wood (2006). In this context, GAM can model variables that follow not only normal distribution but also Poisson, binomial, and other distribution types. For example, the number of counts, probability, presence/absence can be used as a response variable in GAM. A formula of GAM can be written as:

$$\mathrm{E}(y) = g^{-1}(f_1(x_1) + \cdots + f_k(x_k)) \quad (3)$$

where,
y is the response variable,
E(y) is the expectation of y,
g^{-1} is the inverse of link function g,
$f_j(x_j)$ is a smooth function of x_j.

A spline is a special function defined piecewise by polynomials, and is usually used as a smooth function in the GAM. When $f_j(x_j)$ is a linear function with respect to x_j, equation (3) becomes a Generalized Linear Model (GLM) (McCullagh and Nelder, 1989). In this sense, GAM can be regarded as an extension of GLM.

BASIC CONCEPT FOR THE CONSTRUCTION OF DENSITY SURFACE MODEL (DSM)

There are four steps in constructing a DSM from line transect survey data.

1. Fitting detection function with perpendicular distance data and other covariates that would affect detectability of whale schools.
2. Fitting count models in each grid with environmental covariates data.
3. Prediction of abundance in each grid in the study area using the models constructed above.
4. Summation of the predicted abundance in the grid over the area of interest.

The abundance estimate in the study area is obtained as follows.

Step 1

The detection function represents the relationship between the detection probability and the perpendicular distance. It is assumed that the detection probability decreases as the perpendicular distance increases. Figure 1 shows a histogram of relative frequency of detection by perpendicular distance and estimated detection function. T is the truncation distance (i.e., detections of perpendicular distance more than T are excluded from the analysis). Among the animals within T n.miles from the track line, the zone below the detection function represents animals actually detected. The red zone represents the animals missed. The detection function predicts probability for each detection. The proportion of the former is p, the detection probability and proportion of the latter is 1-p. By dividing the number of the detections by the detection probability p, the number of whales that fail to be detected can be taken into account (i.e., red zone in Figure 1).

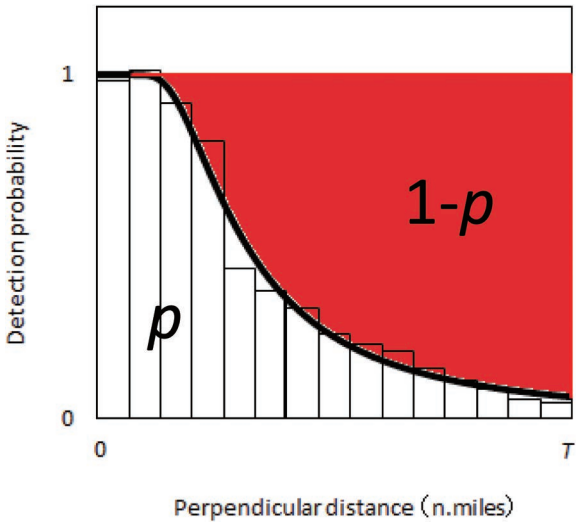

Figure 1. Histogram of relative frequency of detections by intervals of perpendicular distance and plot of detection function (bold curve). The red zone represents the proportion of animal detections missed as compared to the animals in the surveyed area within T n.miles from the track line, where T is truncation distance. Example of the histogram and the detection function models are taken from Hakamada *et al.* (2013).

The detection function derived from Multi Covariate Distance Sampling (MCDS) can be used in this method. The advantage of using MCDS is that the effect of covariates on sighting conditions, such as Beaufort scale, wind speed, and visibility, on the detection function can be assessed. As mentioned later, the detection function with assumption $g(0)<1$ can be used in this method.

Step 2

Figure 2 illustrates the division into grids to construct count models. As shown in the figure, the track line sur-

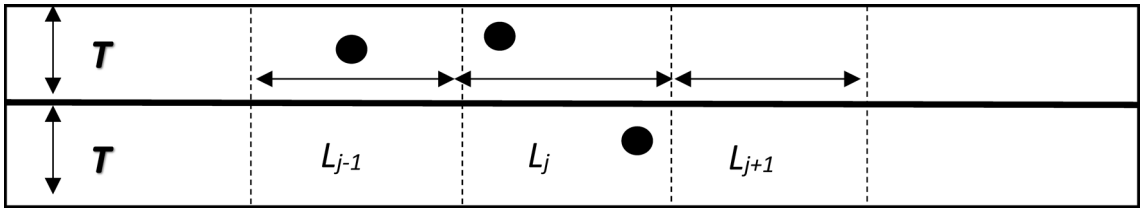

Figure 2. Illustration of the division of track line into grids to construct count models. Black circle indicates a detected whale school. Bold line indicates the track line surveyed. T is truncation distance. L_j is the length of the track line surveyed in the j th grid. Area size of the j th grid is $A_j=2L_jT$.

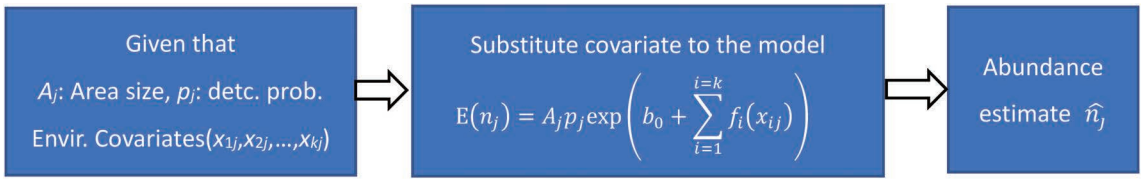

Figure 3. Scheme to predict abundance estimate in each grid.

veyed were divided into several sections.

The j th grid has a width of $2T$ and length of L_j. L_j is not necessarily constant. The area size of the j th grid is $A_j=2L_jT$. The area size should be chosen, such that it is small enough to ensure that density in the grid can be regarded as almost uniform. The number of the detected school in the j th grid is n_j. The count model has covariates that could change in response to space and other factors (e.g., longitude, latitude, SST, depth, salinity) as explanatory variables and the number of counts per unit area size as a response variable.

For simplicity, it is assumed that the values of the explanatory variables x_js are available for any position in the area of interest. When fitting the model, explanatory variables in the grid are usually obtained by averaging all observed values in the grid. One example of the count model can be expressed by:

$$\mathrm{E}(n_j) = A_j p_j \exp\left(b_0 + \sum_{i=1}^{i=k} f_i(x_{ij})\right) \quad (4)$$

where,

$\mathrm{E}(n_j)$ is the expected the number of the detection in the grid j,

A_j is the area size of the grid j as an offset,

p_j is the averaged detection probability for detected schools in the j th grid derived from the detection function constructed in step 1,

b_0 is intercept,

k is the number of the explanatory variables, and

$f_i(x_{ij})$ is the i th smooth function of the i th covariate x_{ij} at the j th grid.

When the detection function has covariates on sighting condition derived by MCDS, it can be considered that p_j may be different among the grids.

Step 3

The grids for prediction of abundance are defined in order to cover the study area of interest. The shape of the grid is usually almost a square. The size of the grid should be small enough so that density can be regarded as almost uniform. These grids are different from those defined in Step 2.

The expected numbers of detection (abundance estimate) in the grid can be predicted by using equation (4), if A_j, p_j and x_{ij} for all i are available (Figure 3). A_j is the area size of the grid j and is clearly available. p_j can be predicted in all the grids if one of the two following conditions are satisfied. The first is that the covariate of the detection function derived in Step 1 is only the perpendicular distance (in this case, p_j is constant for all grids). The second is that the detection function has covariates on sighting conditions and the value of the covariates are available for all the grids. The explanatory variables in the grid are usually obtained by averaging all observed values in the grid. Thus, abundance estimate in all the grids can be predicted.

Step 4

The predicted abundance estimate in each grid are totaled to obtain the abundance estimate for the whole area of interest. When the shape of the area of interest is complex, only a part of the grid belongs to the area for some of the grids at the edge of the area. In such case, equation 4 can be applied to the intersect of the grid and the area.

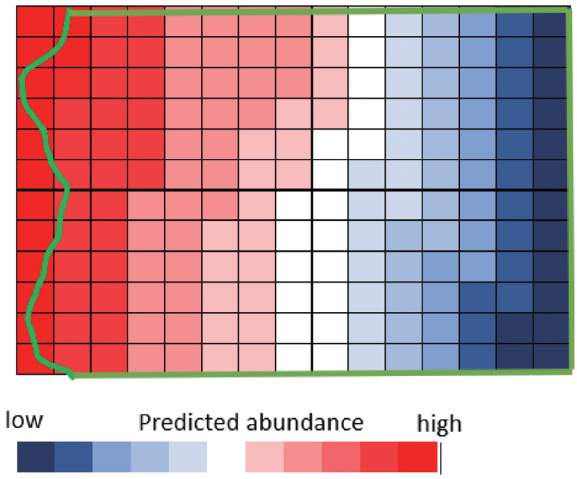

Figure 4. Illustrative example of a map for the predicted abundance in each grid. Strong red indicates high density and strong blue indicates low density. Green line indicates the boundary of the area of interest.

Figure 4 is an illustrative map for the predicted abundance in each grid. In this case, the model shows a tendency whereby the predicted abundance is getting higher from the right to the left. For some of the grids, the whole of the grid is not included in the area of interest. Prediction can be conducted using the data in the intersect of the grid and the area. One of the advantages of predicting abundance by the grids is that it is easy to calculate the abundance estimate for any subset of the surveyed area by summing the predicted abundance over the grids in the subset.

Further aspects on DSM

One and two stage approaches

Explanation in the previous section is based on the two-stage approach, which means fitting DSM and detection model at each stage. A one-stage approach means fitting both models simultaneously (Royle *et al.*, 2004). The disadvantage of the one-stage approach is that the computation to estimate and check is more difficult than in the two-stage approach, because both steps must be conducted at once. The disadvantage of a two-stage approach is that to estimate uncertainty for the abundance estimate, the uncertainties in the detection function and the spatial model must be combined appropriately (Miller *et al.*, 2013).

Distribution of response variable

In the case of count models, Poisson distribution is usually assumed as the distribution of the count. In many cases, there is a high proportion of the grids with zero count (i.e., there is no sightings in the grid), because the numbers of grids constructed in Step 2 is larger than the numbers of animals detected. Alternative distributions of count are, for example, quasi-Poisson, negative binomial and Tweedie distribution (Tweedie, 1984), which deal with the higher proportion of zero data. The latest version of R library dsm, which is software that can conduct DSM, can deal with quasi-Poisson, negative binomial and Tweedie distribution as the distribution of a response variable (Miller *et al.*, 2021).

Variance estimation

In previous studies on abundance estimates derived from DSM, the variance was estimated using parametric bootstrap (Hedley et al., 1999; Hedley and Buckland 2004). When the detection function model and the count model are independent, the variance of the abundance can be approximated as the sum of variance due to detection function and variance due to count model using delta method (Seber, 1982). When the detection function and the count model are not independent, the approximation by the delta method cannot be applied to estimate the variance of abundance. In this case, variance propagation can be applied to estimate the variance of the abundance. Details of the variance propagation are provided in Bravington *et al.* (2021).

Evaluation of extrapolation using DSM

When applying DSM to areas outside of the reference data (i.e., data used to construct the model), attention should be paid to the extrapolation of DSM (Miller *et al.*, 2013). For example, when the DSM that has SST as a covariate is fitted using data from warm waters in a temperate zone, predictions for cold waters in the sub-arctic by the model are unlikely to be reliable due to difference in the range of SST. Bouchet *et al* (2020a) reviewed several quantitative extrapolation diagnostics and recommended to use two of them, principally, the percentage of data nearby, %N (King and Zeng, 2007) and the Extrapolation Detection, ExDet (Mesgaran *et al.*, 2014), as standard tools for assessing extrapolation in abundance models. %N is a metric that represents how close the data for extrapolation and the reference data are. The higher the %N is, the more reliable the extrapolation is. ExDet is a metric that characterizes both univariate and combinatorial extrapolation and provides geographical distribution of the most influential covariate among the covariates (Bouchet *et al.*, 2020a). Assessment of extrapolation using the two metrics can be implemented by dsmextra library of R (Bouchet *et al.*, 2020b).

POTENTIAL UTILITY OF THE DSM FOR THE WORK ON WHALE ABUNDANCE BY THE INSTITUTE OF CETACEAN RESEARCH

The model-based approach can be applied to the case of Antarctic minke whales and common minke whales in the western North Pacific. These are two of the target species for abundance research by the ICR.

In the case of Antarctic minke whales, the abundance estimation in polynyas is problematic. The pack ice prevents the research vessels entering the polynyas. The shape of the polynyas is often complex and narrow, therefore a random and/or systematic survey design cannot be realized even if platforms other than the vessel (e.g., drone) are used for the survey. In such cases, the data obtained can be used to estimate abundance based on DSM.

The common minke whales in the western North Pacific are distributed in the Okhotsk Sea, the western North Pacific Ocean and the Sea of Japan. Some areas of the distribution of this species cannot be surveyed, because they are in the Exclusive Economic Zone (EEZ) of third countries, from which permits are difficult to obtain. This situation results in underestimation of the abundance because not all areas of distribution of the species or stock are covered. If some environmental data in the unsurveyed areas are available, extrapolation by model-based approach can be applied. This study would be ambitious and challenging, because it is difficult to obtain reliable DSM given that the ranges of environmental covariates in areas where the DSM was extrapolated would be expected to be different from the reference data (i.e., the data used to construct the model). However, the reliability of the extrapolation can be evaluated using the indices mentioned above. By assessing conditions under which models are likely to fail or succeed in extrapolation through this exercise, a better understanding of distribution patterns and their underlying drivers for this species may be obtained (Yates *et al.*, 2018).

The latest version of R library dsm can deal with detection function derived from mark-recapture distance sampling (MRDS) that allows $g(0)<1$ (Miller *et al.*, 2021). Concept and example of such detection function derived from the MRDS are provided in Takahashi (2019). From this update, DSM derives abundance estimates that consider $g(0)$ estimate. This is particularly relevant for species that cannot be detected due to their behavior and whose distribution is driven by environmental covariates varying in space. DSM that allows $g(0)<1$ can be applied to whales, such as common minke whales and Antarctic minke whales, whose $g(0)$ estimates have been shown to be less than 1 in previous studies (Okamura *et al.*, 2010; Okamura and Kitakado, 2012).

ACKNOWLEDGEMENTS

I thank Luis A. Pastene (ICR) for his assistance in preparing this manuscript.

REFERENCES

Bouchet, P.J., Miller, D.L., Roberts, J.J., Mannocci, L., Harris, C.M. and Thomas, L. 2020a. From here and now to there and then: Practical recommendations for extrapolating cetacean density surface models to novel conditions. Technical report 2019-01 v2.0, Centre for Research into Ecological & Environmental Modelling (CREEM). University of St Andrews, St Andrews. 59 pp.

Bouchet, P.J., Miller, D.L., Roberts, J.J, Mannocci, L., Harris, C.L. and Thomas, L. 2020b. dsmextra: Extrapolation assessment tools for density surface models. *Methods in Ecology and Evolution* 11: 1464–1469.

Bravington, M.V., Miller, D.L. and Hedley, S.L. 2021. Variance propagation for density surface models. *JABES* 26: 306–323.

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

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L. 2001. *Introduction to Distance Sampling*. Oxford University Press, Oxford. 595 pp.

Hakamada, T. and Matsuoka, K. 2017. Sighting survey procedures for abundance estimates of large whales in JARPA and JARPAII, and results for Antarctic minke whales. *Technical Reports of the Institute of Cetacean Research (TEREP-ICR)* No. 1: 28–36.

Hakamada, T., Matsuoka, K., Nishiwaki, S. and Kitakado, T. 2013. Abundance estimates and trends for Antarctic minke whales (*Balaenoptera bonaerensis*) in Antarctic Areas IV and V based on JARPA sighting data. *J. Cetacean Res. Manage.* 13(2): 123–151.

Hedley, S.L., Buckland, S.T. and Borchers, D.L. 1999. Spatial modelling from line transect data. *J. Cetacean Res. Manage.* 1: 255–264.

Hedley, S.L. and Buckland, S.T. 2004. Spatial models for line transect sampling. *JABES* 9: 181–199.

King, G. and Zeng, L. 2007. When can history be our guide? The pitfalls of counterfactual inference. *Int. Stud. Q.* 51: 183–210.

McCullagh, P. and Nelder, J.A. 1989. *Monographs on Statistics and Applied Probability. 37. Generalized linear model, 2nd edition.* Chapman and Hall, London. 511 pp.

Mesgaran, M.B., Cousens, R.D. and Webber, B.L. 2014. Here be dragons: A tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Divers. Distrib*. 20: 1147–1159.

Miller, D.L., Burt, M.L., Lexsdat, E.A. and Thomas, L. 2013.

Spatial models for distance sampling data: recent developments and future directions. *Methods in Ecology and Evolution* 4: 1001–1010.

Miller, D.L., Rexstad, E.A., Burt, M.L., Bravington, M.V. and Hedley, S.L. 2021. dsm: Density surface modelling of distance sampling data. Ver. 2.3.1. URL http://github.com/dill/dsm.

Okamura, H., Miyashita, T. and Kitakado, T. 2010. *g*(0) estimates for western North Pacific common minke whales. Paper SC/62/NPM9 presented to the IWC Scientific Committee, June 2010 (unpublished). 7 pp. [Available from the IWC Secretariat].

Okamura, H. and Kitakado, T. 2012. Abundance estimates of Antarctic minke whales using the OK method. Paper SC/64/IA2 presented to the IWC Scientific Committee, June 2012 (unpublished). 24 pp. [Available from the IWC Secretariat].

Royle, J., Dawson, D. and Bates, S. 2004. Modeling abundance effects in distance sampling. *Ecology* 85: 1591–1597.

Seber, G.A.F. 1982. *The estimation of animal abundance and related parameters*. Macmillian, New York. 654 pp.

Takahashi, M. 2019. A note on *g*(0) estimates derived from vessel-based sighting surveys. *Technical Reports of the Institute of Cetacean Research (TEREP-ICR)* No. 3: 14–20.

Thomas, L., Buckland, S.T., Rexstad, E.A., Laake, J.L., Strindberg, S., Hedley, S.L., Bishop, J.R., Marques, T.A. and Burnham, K.P. 2010. Distance software: design and analysis of distance sampling surveys for estimating population size. *J. Appl. Ecol.* 47: 5–14.

Tweedie, M.C.K. 1984. An index which distinguishes between some important exponential families. pp. 579–604. *In*: J.K. Ghosh and J. Roy (eds.). *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*. Indian Statistical Institute, Calcutta. 609 pp.

Wood, S.N. 2006. *Generalized Additive Models: An introduction with R*. Chapman and Hall/CRC, Boca Raton, FL, USA. 391 pp.

Yates, K.L., Bouchet, P.J., Caley, M.J., Mengersen, K., Randin, C.F., Parnell, S., Fielding, A.H., Bamford, A.J., Ban, S., Barbosa, A.M., Dormann, C.F., Elith, J., Embling, C.B., Ervin, G.N., Fisher, R., Gould, S., Graf, R.F., Gregr, E.J., Halpin, P.N., Heikkinen, R.K., Heinänen, S., Jones, A.R., Krishnakumar, P.K., Lauria, V., Lozano-Montes, H., Mannocci, L., Mellin, C., Mesgaran, M.B., Moreno-Amat, E., Mormede, S., Novaczek, E., Oppel, S., Ortuño Crespo, G., Peterson, A.T., Rapacciuolo, G., Roberts, J.J., Ross, R.E., Scales, K.L., Schoeman, D., Snelgrove, P., Sundblad, G., Thuiller, W., Torres, L.G., Verbruggen, H., Wang, L., Wenger, S., Whittingham, M.J., Zharikov, Y., Zurell, D. and Sequeira, A.M.M. 2018. Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol.* 33: 790–802.