Asymptotic bias of the hazard probability model under model mis-specification

TORE SELLAND KLEPPE*, HANS J. SKAUG* AND HIROSHI OKAMURA+

Contact e-mail: skaug@math.uib.no

ABSTRACT

We compare the sensitivity of the estimated effective strip half-width with respect to choice of hazard probability function (Q). This is done by fitting the model under an erroneous assumption about the parametric form of Q, and comparing the estimated and true effective strip half-width. An 'infinite sample size' setting is employed, where fitting the model by maximum likelihood amounts to minimising the Kullback Leibler distance between the assumed and true models. The experiment is carried out in a situation that is relevant to minke whale sighting surveys both in the Antarctic and in the northeastern Atlantic. It is found that the hazard probability model is fairly robust with respect to the choice of parametric class for Q. The largest observed bias in the resulting effective strip half-width is less than 10%, while for most situations there is almost no bias.

KEY WORDS: ABUNDANCE ESTIMATE; g(0); SURVEY-VESSEL

INTRODUCTION

The hazard probability model has been used within the International Whaling Commission's Scientific Committee to model independent observer line transect data for minke whales, because it directly takes into account the discrete availability of the animals (Okamura *et al.*, 2003; Skaug *et al.*, 2004). The hazard probability function $Q(x, y)$ is defined as the probability of observing a cue that occurs at relative position (x, y), given that the observer is not previously aware of the whale. Here, x and y are perpendicular and forward distances (km), respectively. The purpose of the present paper is to study how sensitive quantities such as the effective strip half-width and the perpendicular distance density are to the choice of . For this purpose, we perform a pairwise comparison of four alternative parametric families for Q. For each comparison we take one Q as being the truth, with the other being treated as an approximation (Q*). We then tune the parameters of Q* such that the Kullback-Leibler (KL) distance between the models is minimised and finally we compare the corresponding effective strip half-widths, w and w*.

MATERIAL AND METHODS

Hazard probability model for independent observers

Consider first a single observer with hazard probability function $Q(x, y)$, and assume that the whales are stationary (do not move) and surface according to a Poisson process. The detection function, i.e. the probability of detecting a whale that is present at perpendicular distance x, is given as

$$g(x) = 1 - \exp\left(-\frac{\alpha}{v}\int_0^\infty Q(x, y)\, dy\right),$$

where α is the surfacing rate of the whale, and v is the speed of the observer. The probability density of the relative position of the initial observations is given as

$$f(x, y) = \frac{\alpha}{vw} Q(x, y) \exp\left(-\frac{\alpha}{v}\int_y^\infty Q(x, u)\, du\right),\ 0 \le x \le W,\ y \ge 0,$$

where w is the effective strip half-width given by

$$w = \int_0^W g(x)dx = \int_0^W \left[1 - \exp\left\{-\frac{\alpha}{v}\int_0^\infty Q(x, y)\, dy\right\}\right]dx.$$

Observations falling outside the observation strip $(0, W)$ are discarded.

There are typically two or more independent observers (or observer platforms). In the common minke whale (*Balaenoptera acutorostrata*) surveys in the northeastern Atlantic a symmetric two-platforms design is used (Skaug *et al.*, 2004), while in the surveys for Antarctic minke whales (*B. bonaerensis*), three platforms, with a partly asymmetrical configuration, have been used (Okamura *et al.*, 2003). For simplicity, we shall adhere to the setting of Skaug *et al.* (2004) and assume that there are two independent observers, which we denote by A and B, having the same Q function. The combined observer $A \cup B$, i.e. viewing A and B as being a team, has hazard probability function

$$Q_{A\cup B} = Q_A + A_B - Q_A Q_B = 2Q - Q^2.$$

To get expressions for $g(x)$, $f(x, y)$ and w for the combined observer $A \cup B$ we can directly insert in the above formulae. Further, each animal detected by sets up an experiment with trinomial outcome $u \in \{A, B, AB\}$. Conditionally on the position (x, y) the probability distribution of u is

$$q(u \mid x, y) = \{Q_{A\cup B}(x, y)\}^{-1} \begin{cases} Q_A(x, y)\{1 - Q_B(x, y)\}, & u = A; \\ Q_B(x, y)\{1 - Q_A(x, y)\}, & u = B; \\ Q_A(x, y)Q_B(x, y), & u = AB. \end{cases}$$

Via the above formulae, the true hazard probability function Q^T and its approximation Q* induce two different probability distributions for the datum (x, y, u). The KL distance between these distributions is given as

* *Department of Mathematics, University of Bergen, Norway.*
+ *National Research Institute of Far Seas Fisheries, 5-7-1 Orido, Shimizu, Shizuoka 424-8633, Japan.*

$$KL = \sum_u \int_0^\infty \int_0^W \log \frac{f_T(x, y, u)}{f_*(x, y, u)} f_T(x, y, u) dxdy \quad (1)$$

$$= \sum_u \int_0^\infty \int_0^W \left[\log \frac{q_T(u \mid x, y)}{q_*(u \mid x, y)} + \log \frac{f_T(x, y)}{f_*(x, y)} \right] f_T(x, y, u) dxdy$$

$$= \int_0^\infty \int_0^W KL(u \mid x, y) f_T(x, y) dxdy + \int_0^\infty \int_0^W \log \frac{f_T(x, y)}{f_*(x, y)} f_T(x, y) dxdy,$$

where

$$KL(u \mid x, y) = \sum_{u = A, B, AB} \log \frac{q_T(u \mid x, y)}{q_*(u \mid x, y)} \; q_T(u \mid x, y).$$

Here, we have exploited that $f(x, y, u) = q(u \mid x, y)\, g(x, y)$. In the expression for KL above, g denotes the density based on $Q_{A \cup B}$.

Experimental setup

The four parametric forms $Q_1 - Q_4$ considered are shown in Table 1. Each Q was in turn taken to be the true model (Q^T), while treating the three others as approximating models (Q^*). For a given Q^T the parameters of Q^* were chosen so that the KL distance (1) was minimised. The practical interpretation of this is to use maximum likelihood estimation under an erroneous model assumption, in a situation where an infinite amount of data (from the correct model) is available. The data being fit to consisted of two parts: (i) the initial position for the combined observer $A \cup B$, i.e. the position (x, y) where the whale was first detected (regardless of whether it was A, B, or both that actually made the detection); and (ii) the outcome $u \in \{A, B, AB\}$ of the trinomial trial. Observations falling outside a strip $(-W, W)$ were discarded.

The parameter values used as the 'true values' for each of the four functions are given in the first column of Table 2. For $Q_1 - Q_3$ – these values were based on Antarctic minke whale data (CP 3, Area 5, Okamura and Kitakado, 2009a) and for Q_4 the parameter values were based on northeastern Atlantic minke whale data (Skaug *et al.*, 2004). In the Antarctic setting (three upper panels of Table 2) we truncated at $W = 2$ km, while $W = 1$ km was used in the bottom panel of Table 2 due to the much shorter effective strip half-width in the Northeastern Atlantic. The vessel speed was taken to be 11.5 knots, and mean surfacing rate was 48 surfacings per hour.

The numerical minimisation of the KL distance, with respect to the parameters of Q^*, was done in Matlab. All integrals occurring above were evaluated using numerical integration in Matlab (precision 10^{-6}) as well. The integration range in the forward direction (y) was 0–6 km, except for the bottom panel of Table 2, where the range was 0–3 km.

The parameter of main interest for animal abundance estimation was, because the abundance estimate is inversely proportional to the estimate of w. Often, it is the single observer version of w, as opposed to $w_{A \cup B}$, that is being used

Table 2

Parameter estimates of approximating models (columns 2–4) that minimise the KL distance to the true model (column 1).

True model		Approximation					
Model 1		Model 2		Model 3		Model 4	
σ_x	1.1779	σ_r	0.1995	σ_r	0.1704	λ_r	0.8029
σ_y	0.0354	σ_θ	3.3828	σ_θ	2.5958	λ_θ	1.3408
γ_x	1.0000	γ_r	1.6901	γ_r	1.7262	ρ_r	0.3086
γ_y	2.5100	γ_θ	0.2452	γ_θ	0.3048	ρ_θ	–81.8074
τ	1.1840	τ	–0.9196	τ	0.0000	μ	0.5531
w	0.9960		0.9252		0.9212		1.0903
$g(0)$	0.8463		0.9953		0.9962		0.8950
KL			0.0109		0.0128		0.0379
Model 2		Model 1		Model 3		Model 4	
σ_r	0.7856	σ_x	2.5982	σ_r	0.5722	λ_r	0.8001
σ_θ	1.0811	σ_y	0.1216	σ_θ	0.9661	λ_θ	2.0854
γ_r	1.0000	γ_x	0.0687	γ_r	1.1204	ρ_r	0.0201
γ_θ	1.5360	γ_y	1.8768	γ_θ	1.5574	ρ_θ	0.5905
τ	0.2940	τ	–0.6814	τ	0.7782	μ	0.3954
w	1.0156		1.1459		1.0113		0.9972
$g(0)$	0.7940		0.9956		0.8016		0.7840
KL			0.0305		0.0001		0.0001
Model 3		Model 1		Model 2		Model 4	
σ_r	0.5362	σ_x	0.1549	σ_r	0.7632	λ_r	0.8388
σ_θ	0.9158	σ_y	0.1436	σ_θ	1.0127	λ_θ	2.3433
$_r$	1.1800	γ_x	2.5193	γ_r	1.0428	ρ_r	–0.0756
γ_θ	1.6930	γ_y	1.8313	γ_θ	1.7125	ρ_θ	0.6965
τ	0.6460	τ	1.3817	τ	0.1259	μ	0.5195
w	1.1265		1.2564		1.1291		1.1237
$g(0)$	0.8472		0.7091		0.8322		0.8493
KL			0.0227		0.0002		0.0001
Model 4		Model 1		Model 2		Model 3	
λ_r	5.0000	σ_x	2.2041	σ_r	2.2290	σ_r	1.7967
λ_θ	5.7296	σ_y	1.9247	σ_θ	0.0832	σ_θ	0.0247
ρ_r	0.6923	γ_x	1.8436	γ_r	1.7725	γ_r	2.0347
ρ_θ	1.6183	γ_y	1.9934	γ_θ	-0.0073	γ_θ	6.7873
μ	0.3700	τ	0.5042	τ	0.2769	τ	0.8671
w	0.3151		0.3226		0.3346		0.3183
$g(0)$	0.4519		0.4616		0.4551		0.4672
KL			0.0043		0.0060		0.0025

Table 1

Different hazard probability functions used in the study: Q_2 and Q_3 are from Okamura and Kitakado (2009) while Q_4 is from Skaug *et al.* (2004). Here, (r, θ) denotes polar coordinates, with $r = \sqrt{x^2 + y^2}$ is radial distance and $\theta \in [0, \pi]$ is the angle relative to the forward direction. Parameter values are given in Table 2.

	Parametric form	Parameter constraints
Model 1	$Q_1(x, y) = \left(1 + \exp(\sigma_x x^{\gamma_x} + \sigma_y y^{\gamma_y} + \tau)\right)^{-1}$	$\sigma_x, \sigma_y, \gamma_x, \gamma_y > 0$
Model 2	$Q_2(r, \theta) = \left(1 + \exp(\sigma_r r^{\gamma_r} + \sigma_\theta \theta^{\gamma_\theta} + \tau)\right)^{-1}$	$\sigma_r, \sigma_\theta, \gamma_r, \gamma_\theta > 0$
Model 3	$Q_3(r, \theta) = \exp(-\sigma_r r^{\gamma_r} - \sigma_\theta \theta^{\gamma_\theta} - \tau)$	$\sigma_r, \sigma_\theta, \gamma_r, \gamma_\theta\, \tau > 0$
Model 4	$Q_4(r, \theta) = \mu \frac{l[-\lambda_r (r - \rho_r)]\, l[-\lambda_\theta (\theta - \rho_\theta)]}{l[\lambda_r \rho_r]\, l[\lambda_\theta \rho_\theta]}$, $l[x] = \frac{\exp(x)}{1 + \exp(x)}$	$\lambda_r, \lambda_\theta > 0,\ 0 < \mu \le 1$

in the abundance calculation (e.g. Skaug *et al.*, 2004). So, although the parameters were estimated from double platform data, we measured the goodness of fit using single-observer versions of w, $g(0)$, and perpendicular distance density $f(x) = \int_0^\infty f(x, y)dy$. As a diagnostic for the fit to the trinomial trials we used, $q(AB \mid x, y)$, i.e. the probability that both observers detect the whale simultaneously.

RESULTS AND DISCUSSION

Table 2 shows parameter estimates, i.e. the values that minimises the KL distance, for all pairwise comparisons of the four hazard probability functions. The corresponding comparisons of the perpendicular distance densities $f(x)$ are given in Fig. 1. This figure also gives the ratios w_T / w_*, which are the key quantity of interest in the present study. When interpreting the density plots it is useful to recall that

$$\frac{w_T}{w_*} = \frac{g_T(0)}{g_*(0)} \cdot \frac{f_*(0)}{f_T(0)}.$$

A misfit in $f(x)$ at $x = 0$ can partly be compensated for by a counteracting misfit in $g(0)$. An example of this is Truth = Q_1 and Approx. = Q_2 for which $g_1(0)/g_2(0) = 0.85$ (Table 2) and $f_2(0)/f_1(0) = 1.27$, yielding $w_1/w_2 = 1.08$. Hence, the perpendicular distance density is not fully diagnostic, and the ratio $f_*(0)/f_T(0)$ does not play the same critical role as it does when $g(0) = 1$ is assumed. Another example of this occurs when Truth = Q_4 and Approx. = Q_1, for which the two density curves are almost identical (Fig. 1; lower left corner). The proportion of ($u = AB$), on the other hand, indicate that there is a misfit (Fig. 2; lower left corner). In a 45 degree sector from the transect line the true model predicts a higher proportion of duplicates than the approximating model (light colored area in the plot), and correspondingly there are too few duplicates in the remaining 45 degree sector.

It is clear from both Figs 1 and 2 that Q_1 differs from Q_1 – Q_4, while Q_2 – Q_4 between themselves yield models with very similar properties. The reason for this is that Q_1 is formulated in Cartesian coordinates (x, y), while Q_2 – Q_4 are formulated in terms of polar coordinates (r, θ). In particular, Q_2 – Q_4 can all be written in the separable form h_1 {$h_2(r)h_3(\theta)$}, where h_1 is a decreasing function, and h_2 and h_3 are increasing functions.

Generally speaking Q_1, predicts more observations close to the vessel than do Q_2 and Q_3. This holds both when Q_1 is taken to be the truth (first row of Fig. 1) and when Q_1 is being fitted (first column of Fig. 1). Further, Q_2 and Q_1 behave very similarly in the comparison with Q_1, also when it comes to the ratios w_T/w_* (Fig. 1) and $q_T(AB \mid x, y)/q_*(AB \mid x, y)$ (Fig. 2). The picture is less clear for the comparison of Q_1 versus Q_4. It is worth noting that the effective strip half-width is over estimated, both when Q_1 is taken as the truth ($w_1/w_4 = w_T/w_* = 0.91$) and when Q_4 is taken as the truth ($w_4/w_1 = w_T/w_* = 0.98$) .

From a conservation perspective a negative bias in is more

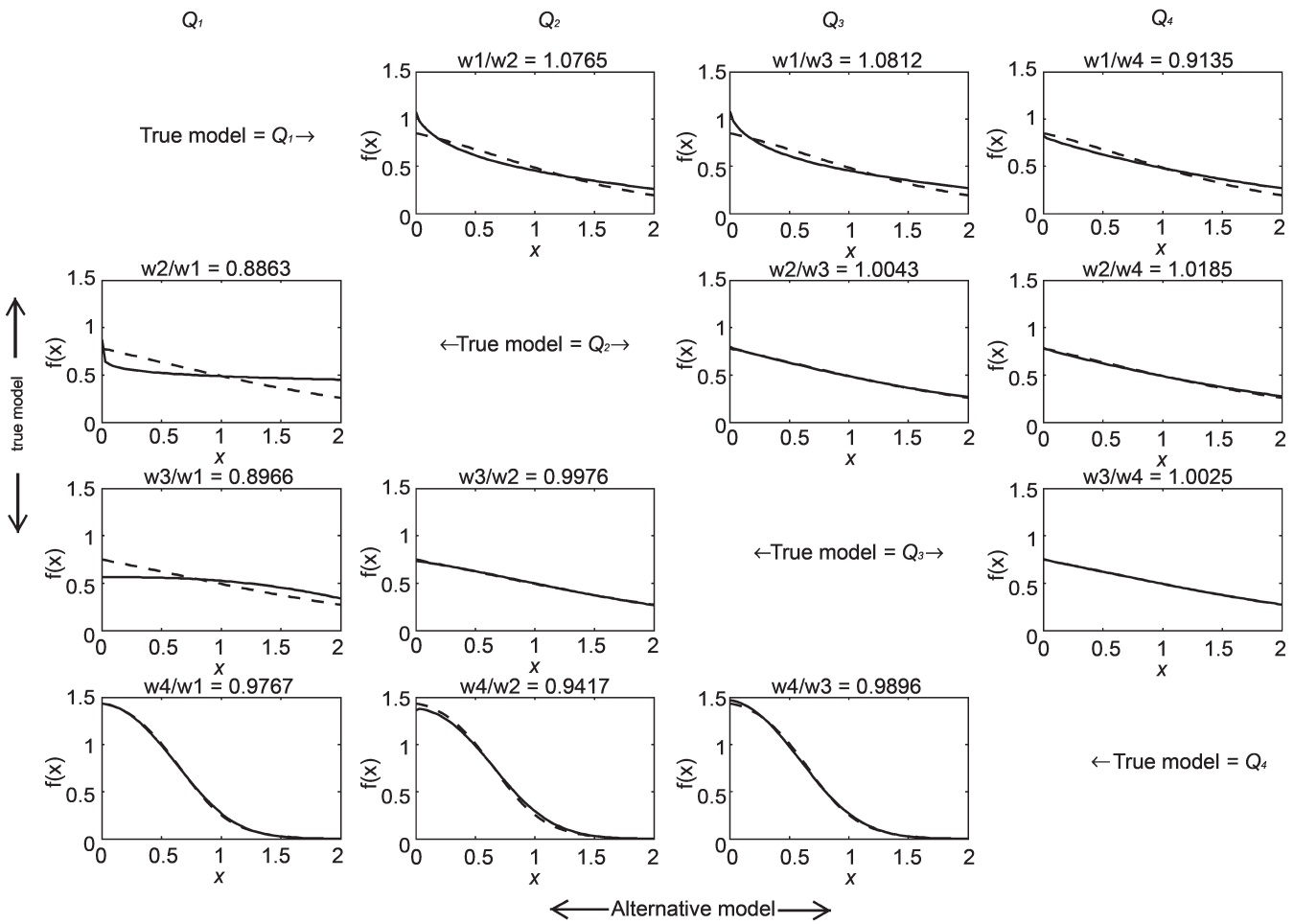

Fig. 1. Comparison of perpendicular distance (km) densities for true (dashed line) and approximating density (solid line), where x is the perpendicular distance. The corresponding ratios of effective strip half widths (w) are also given.

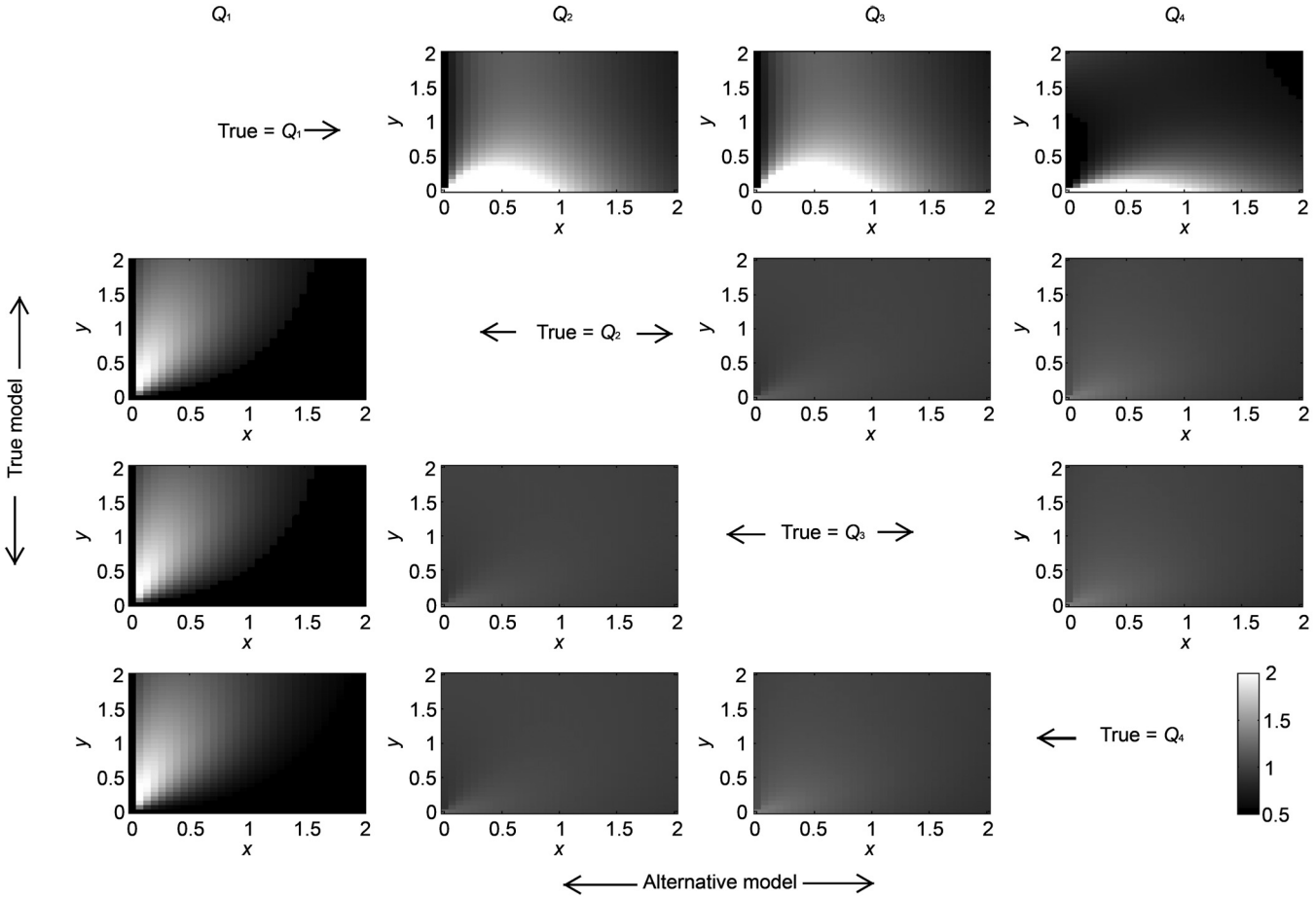

Fig. 2. Ratio $q^{(T)}(AB \mid x, y)/q^{(*)}(AB \mid x, y)$, where $q^{(T)}(AB \mid x, y)$ is the probability (under the true model) that a whale which is initially is observed at location x, y is detected by both A and B (duplicate sighting). Similarly, $q^{(*)}(AB \mid x, y)$ is the probability under the approximating model. The layout of the plot corresponds to that in Fig. 1. The darker the cell, the smaller the ratio.

critical than a positive bias, because the former will lead to a positive bias in the abundance estimate. The most severe underestimation of w found in the present experiment is $w_1/w_3 = w_T/w_* = 1.08$ (Fig. 1; upper row, second column from right). This occurs when Q_1 is the truth and Q_3 is the approximation. For this case Q_3 predicts a too large proportion of duplicate sightings (seen by both A and B) at short radial distance (Fig. 2 upper row, second column from right). The case with the largest over-estimation of w is $w_1/w_4 = 0.91$.

Within each row in Table 2 the smaller the KL distance is, the closer w_T/w_* is to unity. This means that model selection based on a likelihood ratio test, or the AIC criterion, will perform reasonable well for the purpose of picking a model that yields an unbiased estimate of w. The value of the KL statistic does not say anything about the direction of the bias of the estimated w, however.

CONCLUSION

For the purpose of estimating the effective strip half-width (w) the hazard probability model is fairly robust with respect to choice of Q. For all 12 pairwise comparisons considered in this study the fitted w falls within 10% of the true value in all cases. Strictly speaking these conclusions apply only to the version of the hazard probability model used in Skaug *et al.* (2004), and it has not been investigated they hold in the setting of Okamura *et al.* (2003).

We have chosen to use an infinite sample-size setting, which allowed bias arising from mis-specification of Q to be separated from the finite-sample properties of the maximum likelihood estimator. The latter can be studied by simulating data from the hazard probability model, and then applying the estimator on each simulated dataset. This has recently been done for the method of Okamura *et al.* (2003) which did not show any severe biases as a result of finite sample size alone (Okamura and Kitakado, 2009b).

REFERENCES

Okamura, H. and Kitakado, T. 2009a. Abundance estimates and diagnostics of Antarctic minke whales from the historical IDCR/SOWER survey data using the OK method. Paper SC/61/IA6 presented to the IWC Scientific Committee, June 2009, Madeira, Portugal (unpublished). 58pp. [Paper available from the Office of this Journal].

Okamura, H. and Kitakado, T. 2009b. Summary of simulation trials for Antarctic minke whale abundance surveys using the revised OK method. Paper SC/61/IA7 presented to the IWC Scientific Committee, June 2009, Madeira, Portugal (unpublished). 8pp. [Paper available from the Office of this Journal].

Okamura, H., Kitakado, T., Hiramatsu, K. and Mori, M. 2003. Abundance estimation of diving animals by the double-platform line transect method. *Biometrics* 59: 512–20.

Skaug, H.J., Øien, N., Schweder, T. and Bothun, G. 2004. Abundance of minke whales (*Balaenoptera acutorostrata*) in the northeastern Atlantic; variability in time and space. *Can. J. Fish. Aquat. Sci.* 61(6): 870–86.

Date received: September 2009
Date accepted: May 2010