Robustness of the hazard probability model

Tore Selland Kleppe[1], Hans J. Skaug[1], Hiroshi Okamura[2]

May 11, 2009

[1]Department of mathematics, University of Bergen, Norway. [2]National Research Institute of Far Seas Fisheries, 5-7-1 Orido, Shimizu, Shizuoka 424-8633, Japan.

Abstract

We compare the sensitivity of estimated effective strip half-width and detection function with respect to choice of hazard probability function (Q). The model is being fit under different erroneous assumptions about the parametric form of Q. This is done in an "infinite sample size" situation, where fitting the model by maximum likelihood amounts to minimizing the Kullback Leibler distance between the assumed and true models. The experiment is carried out in a setting which 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 around 8%, while for most situations there was almost no bias.

1 Introduction

The hazard probability models is used within the IWC to model independent observer line transect data for minke whales [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 aware of the whale. Here x and y are perpendicular and forward distance, 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 Q. For this purpose we perform a pairwise comparison of four alternative parametric forms of Q, and for each comparison treat one Q as being the truth, while the other being treated as an approximation (Q^*). Then we tune the parameters of Q^* such that the Kullback-Leibler distance between the models is minimized.

2 Hazard probability model for independent observers

Consider first a single observer with hazard probability function $Q(x, y)$. The two-dimensional detection function is given as

$$g(x, y) = \frac{\alpha}{v} Q(x, y) \exp\left(-\frac{\alpha}{v} \int_y^\infty Q(x, u)\, du\right), \tag{1}$$

while the one-dimensional detection function is given as

$$g(x) = \int_0^\infty g(x, y) dy = 1 - \exp\left(-\frac{\alpha}{v} \int_0^\infty Q(x, y)\, dy\right). \tag{2}$$

The density of the initial observations, falling within the observation strip $(0, \omega)$, is given as

$$f(x, y) = \frac{g(x, y)}{w}, \quad 0 \le x \le \omega,\ y \ge 0, \tag{3}$$

and the effective strip half width is given by

$$w = \int_0^\omega g(x) dx = \int_0^\omega \left[1 - \exp\left\{-\frac{\alpha}{v} \int_0^\infty Q(x, y)\, dy\right\}\right] dx. \tag{4}$$

Two independent observers A and B, with the same Q function, yield a combined observer $A \cup B$ with hazard probability function

$$Q_{A\cup B} = Q_A + Q_B - Q_A Q_B = 2Q - Q^2. \tag{5}$$

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

$$q(u|x, y) = \{Q_{A\cup B}(x, y)\}^{-1} \begin{cases} Q_A(x, y)\{1 - Q_B(x, y), & u = A; \\ Q_B(x, y)\{1 - Q_A(x, y)\}, & u = B; \\ Q_A(x, y) Q_B(x, y), & u = AB. \end{cases} . \tag{6}$$

The KL distance between two models for (x, y, u), induced by two different Q-choices $Q^{(T)}$ (the true model) and $Q^{(*)}$ (the approximating model), is given as

$$\begin{aligned} KL &= \sum_u \int_0^\infty \int_0^\omega \log \frac{f_T(x, y, u)}{f_*(x, y, u)} f_T(x, y, u) dx dy \\ &= \sum_u \int_0^\infty \int_0^\omega \left[\log \frac{q_T(u|x, y)}{q_*(u|x, y)} + \log \frac{f_T(x, y)}{f_*(x, y)}\right] f_T(x, y, u) dx dy \qquad (7) \\ &= \int_0^\infty \int_0^\omega KL(u|x, y) f_T(x, y) dx dy + \int_0^\infty \int_0^\omega \log \frac{f_T(x, y)}{f_*(x, y)} f_T(x, y) dx dy, \end{aligned}$$

where

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

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

3 Experimental setup

The four parametric classes for Q that we consider are shown in Table 1. Each Q is in turn taken to be the true model ($Q^{(T)}$), while we are treating the three others as approximating models ($Q^{(*)}$). For a given $Q^{(T)}$ the parameters of $Q^{(*)}$are chosen such that the KL distance (7) is minimized. The practical interpretation of this is to use maximum likelihood estimation under an erroneous model assumption with an infinite amount of data (from the correct model). The data being fitted to consists of two parts: 1) the initial position for the combined observer $A \cup B$, i.e. the position (x,y) where the whale was first detected (regardless whether it was A, B, or both that actually made the detection), and 2) the outcome $u \in \{A, B, AB\}$ of the trinomial trial. Observations falling outside a strip of width $\omega = 1$km at each side of the transect line are discarded.

The parameter values used as the "true values" for each of the four Q functions are given in the first column of Table 2. For Q_1–Q_3 these values are based on Antarctic minke whale data [CP 3, Area 5 Okamura and Kitakado, 2009] and for Q_4 the parameter values are based on on Northeastern Atlantic minke whale data [Skaug et al., 2004].

The numerical minimization of the KL distance, with respect to the parameters of $Q^{(*)}$, is done in Matlab. All integrals occurring in (1)-(7) are being evaluated using numerical integration in Matlab (precision 10^{-6}), as well. In the integration with respect to y, the value 3km was used as the upper limit.

The main interest parameter for animal abundance estimation is w. Often, it is the single observer version of w, as opposed to $w_{\mathrm{A\cup B}}$, that is being used in the abundance calculation [e.g. Skaug et al., 2004]. So although, the parameters estimated from independent observer data, we shall measure the goodness of fit using single-observer versions of w, $g(0)$, and perpendicular distance density $f(x) = \int_0^\infty f(x,y)dy$, respectively. As a diagnostic for the fit to the trinomial trials we plot $q(AB|x,y)$, i.e. the probability that both observers detect the whale simultaneously.

4 Results and discussion

Table 2 shows parameter estimates for all pairwise comparisons of the four hazard probability functions. The corresponding comparisons of the perpendicular distance densities $f(x)$ and ratios of effective strip half-widths w_T/w_* are given in Figure 1. Figure 2 shows $q_{\mathrm{T}}(AB|x,y)/q_*(AB|x,y)$ as a function of (x,y). It is clear from both plots that Q_1 differs from the three other models, while

Q_2–Q_4 yield models with very similar properties. The reason for this is probably that the Q_1 is formulated in Cartesian coordinates (x, y), while Q_2–Q_4 are formulated in polar coordinates (r, θ). In particular, Q_2–Q_4can all be written on 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.

The most severe under-estimation of w is $w_1/w_3 = w_T/w_* = 1.07$ (Q_1 is truth, while Q_3 is the approximating model). In this case fitted perpendicular density is more spiked at $x = 0$ than the true density (Figure 1). However, it is not this phenomenon per se that is causing the misfit, as is clear from the panel $w_1/w_3 = 0.99$ in which the fitted density is also spiked. The case with the largest over-estimation is $w_1/w_4 = 0.93$. It seems that the (x, y) based Q_1 is better at approximating the three (r, θ) based models, than the other way around (compare the first column with first row in Figure 1).

In general, the smaller the KL value is, the closer w_T/w_* is to one (Table 2). This means that model selection based on a likelihood ratio test, or the AIC criterion, will perform reasonable for the purpose of picking a model that yields an unbiased estimate of w.

5 Conclusion

For the approximating effective strip half-width w the hazard probability model is fairly robust with respect to choice of Q. For all 12 pairwise comparisons considered here the fitted w is within 8% of the true value. Further, it is not the value of $f(0)$ per se that causes w_T to differ from w_* as it would be in the classical line transect literature which assumes $g(0) = 1$. All of the hazards function considered here have $g(0) < 1$, and then $f(0)$ plays a less important role.

References

H. Okamura and T. Kitakado. Abundance estimates and some diagnosis of Antarctic minke whales from the historical IDCR/SOWER survey data using the revised OK method. Paper SC/61/IA, 2009.

H. Okamura, T. Kitakado, K. Hiramatsu, and M. Mori. Abundance estimation of diving animals by the double-platform line transect method. *Biometrics*, 59:512–520, 2003.

H. J. Skaug, N. Øien, G. Bøthun, and T. Schweder. Current abundance of minke whales (Balaenoptera acutorostrata) in the Northeast Atlantic; variability in time and space. *Canadian Journal of Fisheries and Aquatic Sciences*, 61: 870–886, 2004.

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

Table 1: Different hazard probability functions used in the study: Q_2 and Q_3 are from Okamura and Kitakado [2009] while Q_4 are 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.

True model		Approximation		Approximation		Approximation	
model 1		model 2		model 3		model 4	
σ_x	2.1815	σ_r	1.0655	σ_r	0.6166	λ_r	1.3617
σ_y	0.1661	σ_θ	2.9047	σ_θ	3.6404	λ_θ	1.1411
γ_x	1.0000	γ_r	1.1855	γ_r	1.5166	ρ_r	0.1401
γ_y	2.5100	γ_θ	0.2948	γ_θ	0.1585	ρ_θ	-33.1107
τ	1.1840	τ	-1.0861	τ	-1.3228	μ	0.4452
w	0.3407		0.3426		0.3177		0.3672
$g(0)$	0.6328		0.9183		1.0000		0.6469
KL			0.0104		0.0146		0.0340
model 2		model 1		model 3		model 4	
σ_r	1.4550	σ_x	1.2810	σ_r	1.1522	λ_r	1.4204
σ_θ	1.0811	σ_y	0.3021	σ_θ	0.9573	λ_θ	2.1382
γ_r	1.0000	γ_x	1.4409	γ_r	1.1136	ρ_r	-0.5920
γ_θ	1.5360	γ_y	2.1100	γ_θ	1.5470	ρ_θ	0.5871
τ	0.2940	τ	1.5987	τ	0.7656	μ	0.4815
w	0.3384		0.3269		0.3374		0.3369
$g(0)$	0.5720		0.4592		0.5826		0.5912
KL			0.0294		0.0001		0.0001
model 3		model 1		model 2		model 4	
σ_r	1.1096	σ_x	1.1949	σ_r	1.4725	λ_r	1.5215
σ_θ	0.9158	σ_y	0.3431	σ_θ	1.0573	λ_θ	2.4254
γ_r	1.1800	γ_x	1.6034	γ_r	1.0371	ρ_r	-0.2786
γ_θ	1.6930	γ_y	2.0790	γ_θ	1.6839	ρ_θ	0.6960
τ	0.6460	τ	1.3550	τ	0.0824	μ	0.5638
w	0.3847		0.3899		0.3860		0.3838
$g(0)$	0.6359		0.5146		0.6240		0.6499
KL			0.0313		0.0001		4.6e-5
model 4		model 1		model 2		model 3	
λ_r	5.0000	σ_x	2.1429	σ_r	2.2879	σ_r	1.7941
λ_θ	0.1000	σ_y	2.1663	σ_θ	0.0780	σ_θ	0.0051
ρ_r	0.6923	γ_x	2.0695	γ_r	1.7609	γ_r	2.0397
ρ_θ	92.7200	γ_y	1.7977	γ_θ	-0.0091	γ_θ	0.7004
μ	0.3700	τ	0.4045	τ	0.2263	τ	0.8663
w	0.3125		0.3149		0.3152		0.3137
$g(0)$	0.4519		0.4569		0.4616		0.4679
KL			0.0017		0.0017		0.0025

Table 2: Parameter estimates of approximating models (columns 2-4) that minimize the KL distance to the true model (column 1).

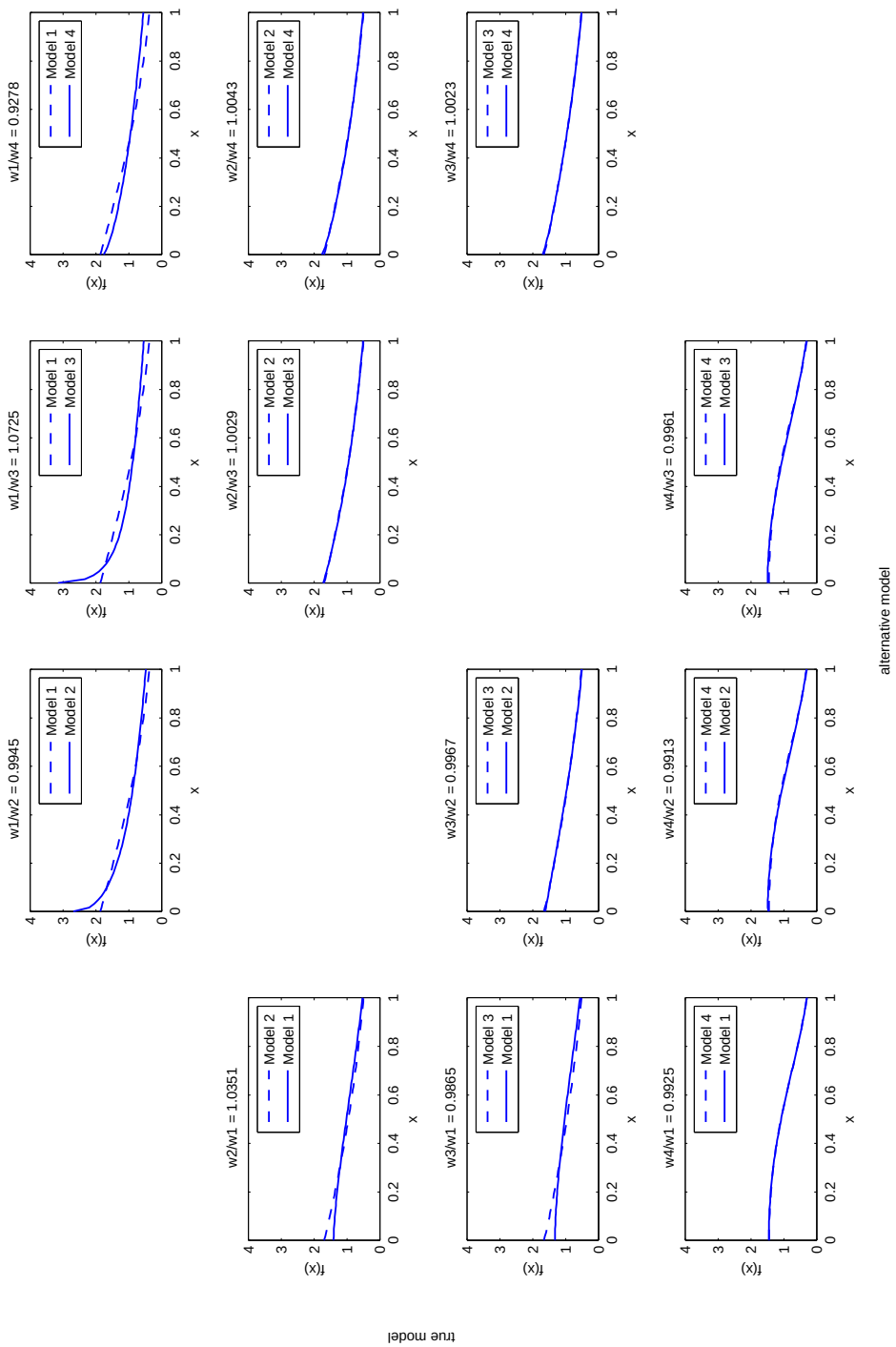

Figure 1: Comparison of perpendicular distance (km) densities for true (dashed line) and approximating density (solid line). The corresponding ratios of effective strip half widths are also given.

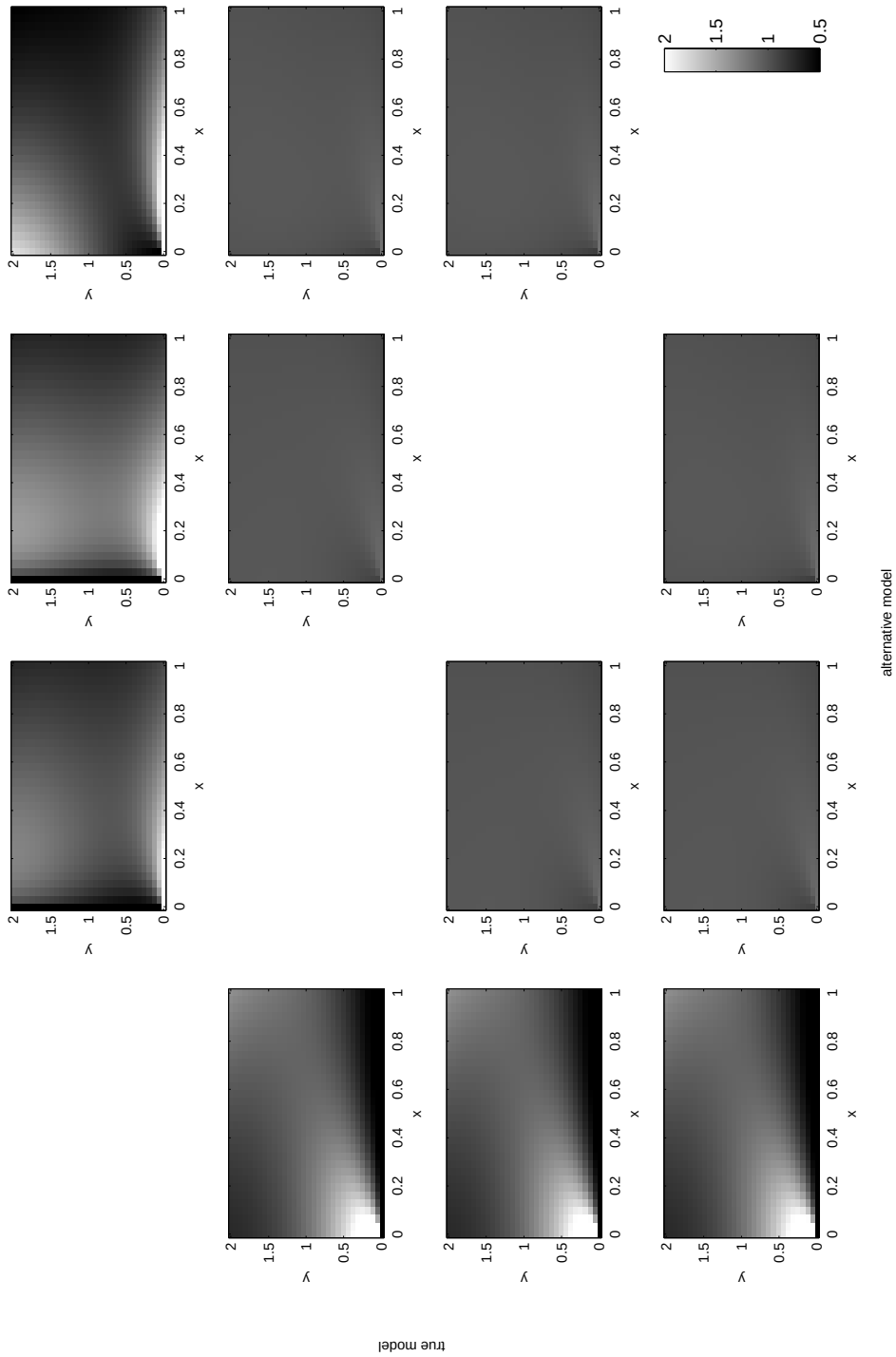

Figure 2: Ratio $q^{(T)}(AB|x,y)/q^{(*)}(AB|x,y)$ between true and approximating probability of the initial observation being made by both platforms. The layout of the plot corresponds to that in Figure 1.