SC/61/RMP3

Description and performance of an automatic duplicate identification routine

GJERMUND BØTHUN, HANS J. SKAUG

Institute of Marine Research, Box 1870 Nordnes, N-5817 Bergen, Norway
Email: gjermund@imr.no

ABSTRACT

The routine described in this document tries to identify duplicate observations of surfacing animals based on similarities in spatial and temporal appearance. The routine is tested on simulated data with properties inherited from survey data. The parameters in the routine is set by tuning. Given the suggested parameters setting the routine is able to correctly identify around 90% of true duplicate and still keep the number of falsely identified duplicate as low as possible. This is an improvement over earlier version of the routine.

INTRODUCTION

An important part of analyzing data from distance sampling survey with symmetrical double platform is identification of duplicate observations. In cue based methods, identification of an individual is sufficient, but in surfacing based methods, identification of observed duplicate surfacing is also vital. This can be done manually either in real-time during the survey, after the survey is finished or automatically by computer program. Schweder et al., (1997) developed and used an automatic routine called “duplicate identification routine, version 1 & 2”. (DIR 1 & 2). Skaug et al., (2002) developed this routine further. Both routines was mainly based on differences in time between surfacing observed by the two platforms. In this paper we develop this routine further by taking into account information in radial distances and bearing.

MATERIALS AND METHODS

Even if the main use of this routine is on data collected from symmetrical double platform, the routine doesn’t assume anything about the configuration of the platforms. We will refer to platforms as A and B, without any ordinal meaning. The routine described in this paper is implemented in the computer language C, and is a part of the software package developed by Norwegian Computer Centre in 1996 called “NAILS”. The routine is tested on simulated data under different condition and also manually jugged against real data from the Norwegian sighting survey 2002 – 2007.

Description of the routine

The routine defines a (whale) track as a series of observation of a whale surfacing within sight, done by one observer. An observer may miss an observation of surfacing. Further observation of surfacing may be incomplete and with error. A duplicate surfacing is defined as one surfacing seen from both platforms. A duplicate track consist of at least one duplicate surfacing.

The objective of the routine is to detect duplicate surfacing within tracks, so the experiments in the Bernoulli trails can be correctly defined.

This duplicate routine search for possible observed duplicate surfacing within tracks considered as duplicate candidate. To be considered as candidates at least one pair of observed surfacing within the two tracks must fulfil the legality criteria defined further below.

For each candidate of duplicate track meeting the legality criteria, two alternatives are considered and a score value is calculated for each alternative. The alternative with the highest score is chosen as the most likely observed duplicate surfacing, and the score is assigned to the candidate pairs of tracks. If there for an observed track seen from one platform exits more than one candidate tracks seen from the other platform, the pair with the highest score value is defined as the duplicate track.

Change in the routine

In DIR1 the routine searched for the two surfacing being closest in time within the two duplicate candidate tracks, these was labelled those surfacing "Minimum distance pair" (MDP). Further the routine then match every preceding and succeeding surfacing in the two tracks as duplicate surfacing, assuming no missing surfacing (holes in tracks). If the MDP appears in the end of long tracks, and there exist problem with "holes in tracks" this could lead to an incorrect identification of the first observed duplicate surfacing. This may have implication on the Bernoulli trails used in estimating g(0).

Based on this observation we modified the routine to more correctly detect the first observed duplicate surfacing. For a given candidate of duplicate track; if platform A spot the whale prior or equal in time to platform B, the track seen from platform A is labelled "first track" and the track observed from platform B is called "last track". And opposite if platform B is aware of the whale first.

Let *i* be a surfacing in "first track" and *j* be surfacing in "last track". Let j=1 be the first surfacing in "last track". Further let i=1 be the first surfacing in "first track" occurring just before surfacing j=1 in "last track". We consider two candidate duplicate surfacing, called alternative I and alternative II in figure 1. Looking at j=1 in "last track", this can be matched with i=1 or i=2 in first tack. For alternative I score value S_1 is calculated, for alternative II score value S_2 is calculated, both given below. The alternative with the highest score value is defined as the duplicate surfacing between the two tracks. For a given "first track" and all candidates "last track", a score value is calculated. Among candidates the one with the highest score value is chosen.

$$\varphi(\langle x_1, x_2, x_3\rangle; \langle \sigma_1, \sigma_2, \sigma_3\rangle) = \sum_{1}^{3}\left(w_i e^{\frac{1}{2}\left(\frac{x_i}{\sigma_i}\right)^2}\right) * \prod_{1}^{3} e^{\frac{1}{2}\left(\frac{x_i}{\sigma_i}\right)^2}$$

$$\Delta_{ij}^{t} = \left|t_i - t_j\right|, \;\; \Delta_{ij}^{\theta} = \left|\theta_i - \theta_j\right|, \;\; \Delta_{ij}^{r} = \left|r_i - r_j\right|$$

$$\Delta_{ij} = \langle \Delta_{ij}^{t}, \Delta_{ij}^{\theta}, \Delta_{ij}^{r}\rangle, \qquad \sigma = \langle \sigma_t, \sigma_\theta, \sigma_r\rangle$$

$$S_1 = \varphi(\Delta_{11}; \sigma_{init}) + \frac{1}{\sqrt{(n_{first} - 1) * (n_{last} - 1)}} \sum_{i=2}^{n_{first}} \sum_{j=2}^{n_{last}} \frac{1}{2}\varphi\left(\Delta_{ij}; \sigma_{sub}\right)$$

$$S_2 = \varphi(\Delta_{21}; \sigma_{init}) + \frac{1}{\sqrt{\left(n_{first} - 2\right) * (n_{last} - 1)}} \sum_{i=3}^{n_{first}} \sum_{j=2}^{n_{last}} \frac{1}{2}\varphi\left(\Delta_{ij}; \sigma_{sub}\right)$$

$$w = \langle 4,2,1 \rangle$$

Different parameter vectors σ_{init} and σ_{sub} are used for the initial duplicate and all the possible subsequence duplicates. Parameters values for σ_{init} and σ_{sub} are given in Table 6.

Performance metrics
In this work we use three different performance metrics to evaluate the performance of the routine under different settings of the legality criteria and different model settings. First we monitor the portion of correctly identified duplicate surfacing as a percent of the total number for true duplicate surfacing. Second we monitor the number of duplicate surfacing the routine failed to detect as a portion of the total number of true duplicates. Third we monitor the number of falsely detected duplicates.

Legality criteria
For a surfacing seen from both platforms to be defined as a legal duplicate by the routine, three criteria have to be fulfilled. Let Δ_t, Δ_θ, Δ_r be the difference in time, bearing and radial distances, as observed form the two platforms, further let $\bar{\theta}$ and $\bar{r}$ be the average observed bearing and radial distance.

$$\bar{\theta} = (\theta_a + \theta_b)/2$$
$$\bar{r} = (r_a + r_b)/2$$

$$\Delta_t < c_t \qquad (1)$$

$$\Delta_\theta < c_\theta * 3.8 * e^{0.0117 * \min(55, |\bar{\theta}|)} \qquad (2)$$

$$\Delta_r < c_r * 1.371 * \bar{r}^{0.763} \qquad (3)$$

The functional form of equation (2) and (3) is based on the measurement error models presented in Aldrin and Tvete, (2002). Criteria 2 and 3 are shown in figure 1 and 2 for different parameter values. For two tracks to be legal duplicate, equation (1) to (3) have to be fulfilled for at least one surfacing seen form both platforms.

Simulation

The routine described in this paper is tested on simulated data. All generated surfacing are based on real data collected from VHF-tagged animals in the period from 1989 to 2008. All spotted surfacing are based on real covariates as observed in the survey in the period from 2002 to 2007, and realised by the hazard probability model estimated in Bøthun et al., (2009). A summary of vital parameters in the error model is given in table 5.

We tested the routine for all estimated Neymann-Scott parameters estimated in Skaug (2009). The estimated Neymann-Scott parameters for the areas NON, BAW1 and LOC are used as low, median and high density area based on the estimated kappa value.

The duplicate identification routine is tested on different combinations of model parameters and Neyman-Scott parameters. Further the routine is tested with different setting of the legality criteria. For every simulation performance statistics is calculated.

The routine is tested with different setting of the legality criteria. In this work the criterion used in Bøthun et al., (2008) is referred to as a "liberal" criteria. Based on comments from IWC, (2008) and initial simulation have we defined a new "strict" criterion. In this work we tested different setting of the legality criterion around the "strict" accept region by increasing the parameters by 10%, 20% and 50% and by decreasing the parameters by 10% and 20%. All parameter settings of legality criterion is given in table 7.

RESULTS

Simulated data

The results of the simulation and testing of the routine described in this paper is summarised in table 1 to table 4.

Simulations 1 to 6, 7 to 12 and 13 to 18 in table 1 are mimicking respectively a low density (block NON), median density (block BAW1) and high density area (block LOC). For each of these densities settings; going from a strict criterion increased by 50%, labelled S*1.5 to a strict criterion decreased by 20%, labelled S*0.8, the numbers of duplicated the routine wasn't able to detect increased. The number of falsely identified duplicates, independent of the availability of a true duplicate, decrease as the accept region decreased. Further we notice that the number of correctly identified duplicates decreased with an decreasing accept region.

Simulation 19 to 24 are all based on a median density area. Further the stochastic part of the measurement error in radial distance is removed. Compared to simulation 7 to 12 which include stochastic error in radial distance, the numbers of correctly identified duplicate increases. The numbers of duplicates the routine didn't identify decreased. The numbers of falsely identified duplicates, given that there exist a true duplicate observation, also decreased. But numbers of falsely identified duplicate, given that there didn't exit a true duplicate observation stayed the same.

Simulation 25 to 27 is based on "perfect" data, i.e. without measurement error, without missing values and after a observer spot an animal, he or her will track the animal without losing it until the end of the track. Under such a model the routine correctly identify 99.9% of the true duplicate independently of any criterion setting.

In simulation 29, removing the possibility of missing value didn't affect the performance of the routine.

Simulation 30 is similar to simulation no. 10 expect that in simulation no. 30 the observers are assumed to track the whale without losing the animal after initial observation until past abeam. Without probability of "losing track" the numbers of falsely identified duplicate, given no common observed surfacing between the two observes, drop to zero.

Simulation 32 to 48 represent a broad range of density settings with the liberal legality criterion. The same set of density setting are used in simulation 49 to 66, which test the strict legality setting. The overall picture, comparing the liberal and strict legality criterion for a given density setting in table 3 and table 4 is that the numbers of duplicates the routine didn't identify are a bit higher for strict criterion. The number of falsely identified duplicate, given no common surfacing are much higher for the liberal criterion. Further we notice that for the liberal criterion, most of the rejection are due to the theta part of the criterion. For the strict criterion the numbers of rejection from the time and theta part of the criterion are more balanced.

Real data

In the data from the survey period 2002 – 2007 the new routine with the liberal legality criterion (50, 10, 5.5) for time, radial distances and bearing respectively, found 171 duplicates. With the strict criterion (12, 5, 3.2) the routine found 149 duplicates. We manually inspected the 22 possibly duplicate rejected by the strict criterion and only found evidence for three of those being duplicates. The number of possible duplicates rejected by the strict criterion was (24,1, 14) and (11,0,21) for the liberal criterion. As seen in the simulated data, introducing a more strict criterion results in a shift from rejection by the bearing criterion towards rejection by the time criterion.

DISCUSSION

The routine presented in this paper utilize more of the available information in judging duplicate observation than the old routine. The objective of the new routine is to more correctly match the first duplicate observation, in contrast to the old routine that focusing on finding the best match within the hole series of surfacing. The old routine assume no holes in tracks, the new routine is still vulnerable to holes in tracks in the first part of the tracks, but the objectives is to more correctly match the initial sightings. This can have some effect on the Bernoulli trails.

Both the new and old routine have been manually checked on real data. We found some cases where the old routine wasn't able to detect true duplicates (Skaug et al. 2004). With the new routine we only found a few cases. But there are still cases that are impossible to judge whether it is a false positive or a false negative duplicate.

The core of this new duplicate identification routine is the score function and the legality criterion. The aim of the score function is to assign a objective value to duplicate candidates and rank them according to this value. An earlier version of the score function did not contain the product part in the functional form. In cases where only one of the time, bearing or radial distance differences was small, this functional form still gave a high score value. For observation made at the same bearing but with large differences in time and radial distances the routine tend to mach such candidates. This led to the inclusion of the product part in the score function. Removing the sum part of the score function may not alter the performance, but this is not tested. In simulation 19 to 24 we removed the stochastic error in radial distance, this resulted in an improved performance. Based on this observation, removing radial distance difference from the score function may be beneficial.

The objective of this work has been to test different settings for the legality criterion. Different functional forms of the legality criterion has not been tested. The optimal parameter setting depends on several properties of the dataset (density, amount of noise, missing values etc). Therefore the parameter setting we recommend is a compromise, with a good overall performance. The

simulation indicate that a setting of 12 seconds as a maximum of difference in time within observation of a common surfacing is close to optimum. The gradient of missing true duplicate is much steeper going below 12 seconds than the gradient of false duplicate going above 12 seconds. In the “liberal” legality setting most of the rejection by the criterion was made by the bearing part of the criterion. We note that with the “strict” legality setting the number of rejection made by the radial distance part and the bearing part are almost equal. We note that the testing of the different parts of the criterion is done sequential in the order time, distance and bearing. Further the time is the information recorded with the highest precision. Based on this it is our assumptions that most of the rejection should be made based on the time part of the criterion. This can be done in two different ways. Either by using a more strict time criterion or by using a more liberal bearing criteria. Finding the optimal balance within the different part of the legality criteria can be seen at as a three dimensional optimization problem, but ensuring a good performance across a broad range of scenarios may complicate this a bit, but we recommend further work to be done to investigate if a different parameter setting of the legality criterion may increase the overall performance over what we has achieved.

Tables

Simulation	Criterion	Block (density)	No. of common surfacing	Performance metrics				No. rejection by legality test		
				% correct duplicates	% missing duplicates	% wrong duplicate given common surfacing exist	% wrong duplicate given no common surfacing exist	Time	R	theta
1	S*1.5	NON	2450	93.0	2.8	4.2	7.5	155	21	294
2	S*1.2	NON	2450	92.7	3.5	3.8	4.3	223	38	275
3	S*1.1	NON	2450	91.8	4.4	3.8	3.8	224	59	271
4	S	NON	2450	90.8	5.6	3.6	3.2	247	89	270
5	S*0.9	NON	2450	88.3	8.4	3.2	2.7	294	123	265
6	S*0.8	NON	2450	84.4	12.8	2.8	2.0	339	191	264
7	S*1.5	BAW1	2554	91.5	3.3	5.2	7.6	542	28	848
8	S*1.2	BAW1	2554	90.9	4.2	5.0	4.7	647	72	809
9	S*1.1	BAW1	2554	89.9	5.4	4.7	4.0	686	102	788
10	S	BAW1	2554	88.1	7.3	4.6	3.3	761	137	768
11	S*0.9	BAW1	2554	85.9	9.8	4.3	2.7	822	172	754
12	S*0.8	BAW1	2554	82.8	13.4	3.8	2.2	915	224	712
13	S*1.5	LOC	1770	88.9	5.5	5.6	7.5	485	22	828
14	S*1.2	LOC	1770	88.5	6.7	4.9	4.7	583	78	795
15	S*1.1	LOC	1770	88.4	6.8	4.8	4.1	622	94	781
16	S	LOC	1770	87.5	7.9	4.6	3.7	660	106	778
17	S*0.9	LOC	1770	85.5	10.1	4.4	3.1	735	143	744
18	S*0.8	LOC	1770	81.6	14.5	3.9	2.5	789	185	730

Table 1 Simulation 1-6 low density, 7-12 median density, 13-18 high density.

Simulation	Criterion	Block (density)	No. of common surfacing	Performance metrics				No. rejection by legality test		
				% correct duplicates	% missing duplicates	% wrong duplicate given common surfacing exist	% wrong duplicate given no common surfacing exist	Time	R	theta
19	L	BAW1	2554	96.4	1.0	2.7	16.3	192	2	1398
20	S*1.5	BAW1	2554	96.4	1.0	2.7	7.7	514	11	851
21	S*1.2	BAW1	2554	96.3	1.1	2.7	4.9	600	23	816
22	S*1.1	BAW1	2554	96.1	1.3	2.7	4.2	633	33	795
23	S	BAW1	2554	95.6	1.8	2.7	3.5	700	45	779
23	S*0.9	BAW1	2554	94.7	2.7	2.6	2.8	749	42	766
24	S*0.8	BAW1	2554	93.5	4.0	2.5	2.2	834	44	720
25	L	BAW1	3266	99.9	0.1	0.0	0.0	279	17	1566
26	S*1.5	BAW1	3266	99.9	0.1	0.0	0.0	696	28	1063
27	S*0.8	BAW1	3266	99.9	0.1	0.0	0.0	997	96	964
28	S	BAW1	3266	86.7	7.9	5.4	0.0	1074	175	983
29	S	BAW1	2608	87.8	7.2	4.9	3.4	786	149	793
30	S	BAW1	3216	86.8	8.1	5.1	0.0	1034	158	957
31	S	BAW1	2554	88.1	7.3	4.6	3.3	761	137	768

Table 2 Simulation 19-24 without stochastic error in R, 25-27 perfect observations, 28 no missing value & no track lose, 29 no missing value, 30 no track lose, 31 full error model, equal to simulation 10 & 49.

Simulation	Criterion	Block (density)	No. of common surfacing	Performance metrics				No. rejection by legality test		
				% correct duplicates	% missing duplicates	% wrong duplicate given common surfacing exist	% wrong duplicate given no common surfacing exist	Time	R	theta
32	L	BAW1	2912	91.3	3.2	5.5	15.7	209	30	1615
33	L	BAE	2420	92.7	2.4	5.0	17.9	76	17	560
34	L	BJ	4916	92.3	2.9	4.8	15.5	273	47	1986
35	L	FI1-02	1966	90.1	4.8	5.1	16.5	238	24	1434
36	L	GA	4775	90.6	3.9	5.5	16.3	647	81	3480
37	L	JMC	2369	85.4	7.4	7.2	15.3	735	51	2845
38	L	KO	2005	90.4	4.4	5.2	16.5	167	24	950
39	L	LOC	2024	88.8	5.3	5.9	15.6	220	37	1551
40	L	NC1-04	2425	91.2	3.7	5.1	17.4	161	21	991
41	L	NC1-06	1890	89.4	5.4	5.2	16.0	85	2	320
42	L	NC2	2415	91.0	3.7	5.3	15.7	133	17	1017
42	L	NON	2790	93.2	2.2	4.6	17.5	58	10	554
43	L	NOS-06	2088	91.5	2.9	5.7	16.7	99	17	870
44	L	NOS-02	2345	91.0	3.6	5.5	17.7	150	27	1199
45	L	NS	1049	92.2	3.1	4.7	18.4	40	3	233
46	L	NVN	1916	91.5	3.5	5.0	16.2	95	22	731
47	L	NVS	1624	93.5	1.9	4.6	14.8	22	2	194
48	L	SV	3328	90.3	4.3	5.3	16.0	453	54	2666

Table 3 Summary of simulation, liberal settings

Simulation	Criterion	Block (density)	No. of common surfacing	Performance metrics				No. rejection by legality test		
				% correct duplicates	% missing duplicates	% wrong duplicate given common surfacing exist	% wrong duplicate given no common surfacing exist	Time	R	theta
49	S	BAW1	2554	88.1	7.3	4.6	3.3	761	137	768
50	S	BAE	2130	89.1	7.2	3.7	3.4	257	97	256
51	S	BJ	4276	89.8	6.7	3.4	2.7	844	204	914
52	S	FI1-02	1734	87.5	8.8	3.6	3.4	687	120	685
53	S	GA	4186	88.4	7.2	4.3	3.0	1635	293	1610
54	S	JMC	2098	83.7	10.7	5.5	4.4	1639	221	1471
55	S	KO	1760	88.4	7.7	3.9	3.4	558	96	540
56	S	LOC	1770	87.5	7.9	4.6	3.7	660	106	778
57	S	NC1-04	2111	87.7	8.4	3.9	3.2	476	133	502
58	S	NC1-06	1656	86.8	8.6	4.6	3.0	333	83	209
59	S	NC2	2102	87.7	7.8	4.5	3.4	456	88	428
60	S	NON	2450	90.8	5.6	3.6	3.2	247	89	270
61	S	NOS-06	1842	88.3	7.2	4.5	2.8	384	86	421
62	S	NOS-02	2064	88.8	7.0	4.2	3.0	495	100	520
63	S	NS	911	89.8	6.4	3.8	2.6	111	32	117
64	S	NVN	1667	88.1	7.6	4.3	2.0	316	71	330
65	S	NVS	1426	89.7	6.2	4.1	2.4	89	57	85
66	S	SV	2936	88.7	6.8	4.5	2.8	1188	177	1127

Table 4 summary of simulation, strict setting

Probabilities of losing end of track	Probabilities of missing initial value			Probabilities of missing subsequence values		
	R	Theta	Phi	R	Theta	Phi
19 %	0.0109	0.0085	0.143	0.192	0.129	0.223

Table 5 Error model parameters

	σ_t [seconds]	σ_Θ [degree]	σ_r [meters]
Initial duplicate, σ_{init}	15	30	300
Subsequence duplicate, σ_{sub}	5	10	100

Table 6 Parameters in duplicate matching score function

	Recorded values		
	c_{time}	c_{θ}	c_r
Liberal	50	10.0	5.5
Strict * 1.50	18	7.5	4.8
Strict * 1.20	14	6	3.8
Strict * 1.10	13	5.5	3.5
Strict	12	5.0	3.2
Strict * 0.90	11	4.5	2.9
Strict * 0.80	10	4	2.6

Table 7 Parameters in legality criterion

Figures

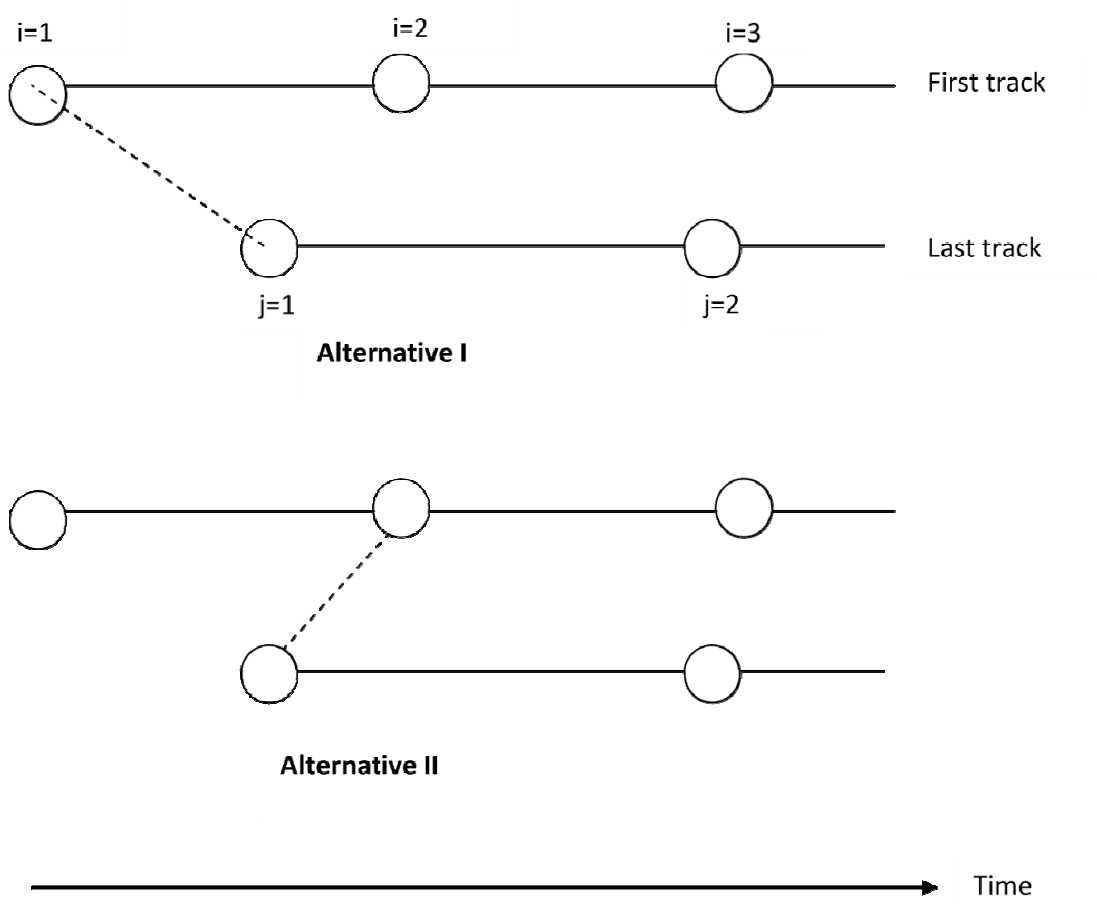

Figure 1 observed surfacing i=1n or i=2 are chosen as duplicate observation of observed surfacing j=1 based on score value.

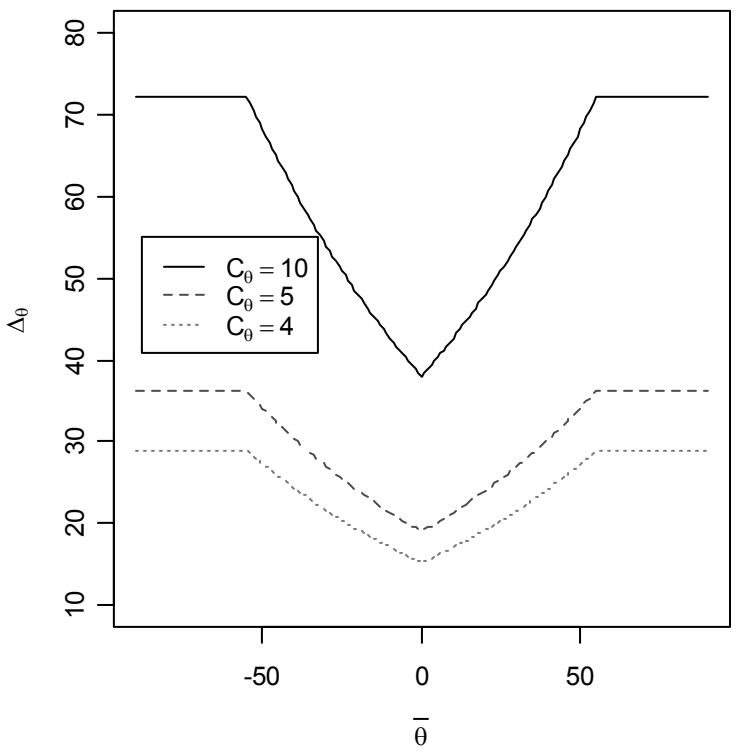

Figure 2 Legality criterion for bearing given different parameter settings.

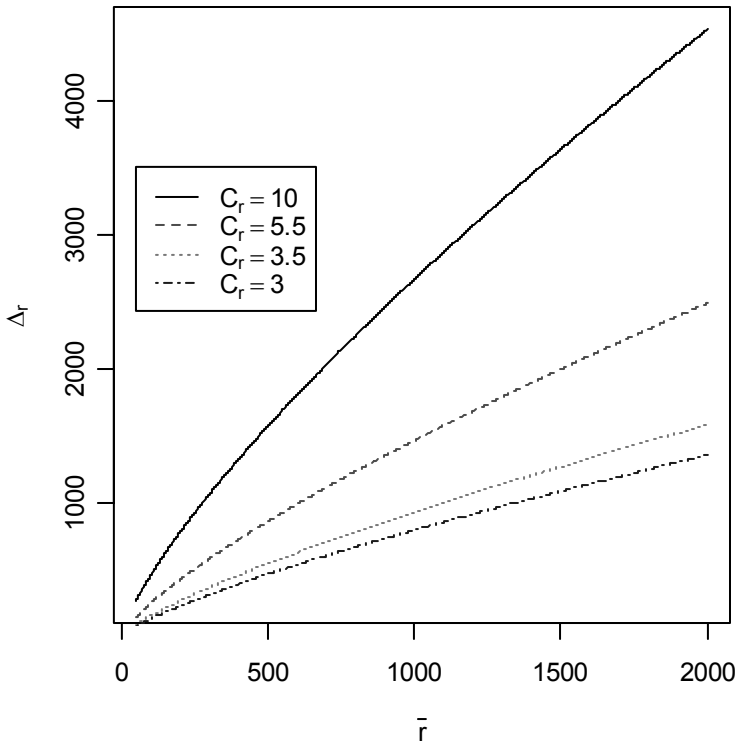

Figure 3 Legality criterion for radial distance given different parameter settings.

Reference List

Aldrin, M., Tvete, F.I. 2002. Measurment Error Models for the Norwegian minke Whale Surveys 1996-2001, IWC/SC/54/RMP3.

Bøthun, G., Skaug, H.J., Øien, N. 2007. Towards an estimate of minek whale abundance in the Norteast Atlantic based on survey data collected over the period 2002-2007, IWC/SC/PFI4.

Bøthun, G., Skaug, H.J., Øien, N., 2009. Abundance of minke whales in the Northeast Atlantic based on survey data collected over the period 2002-2007. IWC/SC/61/RMP2.

International Whaling Commission. 2008. Report of the Scientific Committee. Annex D, Appendix 7. North Atlantic Minke Whales RMP Implementation Review.

Schweder, T, Skaug, H.J., Dimakos, X.K., Langaas, M., Øien, N. 1997. Abundance of Northeastern Atlantic Minke Whales, Estimates for 1989 and 1995. J Cetacean Res Manag , Spec Issue.

Skaug, H.J., Øien, N, Bøthun, G, Schweder, T. 2002. Abundance of northeastern Atlantic minke whales for the survey period 1996-2001, SC/54/RMP5. J Cetacean Res Manag , Spec Issue (53).

Skaug, H.J., Oien, N., Schweder, T., Bothun, G., 2004. 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.

Skaug, H.J., 2009. A method for estimating spatial cluster structure from line transect data. IWC/SC/61/RMP5.