Review of methods to estimate dolphin abundance from line transect surveys and presentation at Southwest Fisheries Science Center (SWFC), La Jolla, California 15-17 Oct. 2001

R. Mohn
Department of Fisheries and Oceans
Bedford Institute of Oceanography
Dartmouth, N.S., Canada

1. Executive Summary

Four documents were sent for review prior to the October 15-17, 2001 meeting. Two dealt with correction of observations at sea; the correction of distance estimates to the school and a calibration of school size once detected. These two papers were well prepared and could stand alone. The third document developed new methods for the statistical estimation of population size from the line transect data. The fourth document incorporated the corrected data and new methods into estimates of abundance for spotted and spinner dolphins.

The improvements to data and methods were reviewed and found to be sound although in many cases requiring clarification. The new estimates of abundance (Figure 1) are probably as good as the data can support. Compared to previous estimates, they are much more consistent in time, as would be expected for a marine mammal population. Also, because of the new analytical tools they are much more precise and the error distributions are more symmetric. The strategy chosen for application of the new methods/data was to pick the best model for each year's survey. While giving individually optimized estimates, a consistent approach may produce a better time series for integration into population models. Examination of the individual models (Table 1) suggests that a consistent analytical model would not depart much from the annually optimized estimates.

The calibration for group size was the major determinant of the new means of the abundance estimates (Figure 2). The calibration was not applied equally to all years as fewer of the observers in the earlier years had been calibrated. Indeed, the first two years of the survey were not corrected at all. A consistent treatment should be developed for all years. It would probably augment the estimates of the earlier years abundance and thus have implications in any reconstruction of the populations. Although the impact of the improved distance measurements was not presented, in one test run it was reported to have had a change in the AIC (Akaike information criterion) of 0.1, an insignificant amount.

The analytical methods that were developed deserve special comment. They explored several models, both parametric and non-parametric, and represented new developments for the analysis of line transect data. The computer code was developed on-site and validated using simulated data. Because the analytical tools were so new, care must be exercised that no coding errors are accidentally introduced. Until the code has chance for

further and more complete testing, simulated data should be applied routinely to assure its integrity. It would also be valuable to develop simulations that more closely represent the real data as a test bed for program development.

Because of the parallel development of data and methods, it was often difficult to assign differences in estimated stock status to individual operations. That is, if new data and analysis were both applied at one step, it was difficult to apportion the relative contribution of either. Although a complete systematic decomposition of the affects of changes to data and analysis was not possible during the review, sufficient diagnostics were provided to establish the validity of the new abundance estimates.

The authors were extremely responsive to requests for further clarification or additional work and facilitated the review process considerably.

2. Summary of Available Information

Four documents were sent for review prior to the October 15-17, 2001 meeting. Two dealt with correction of observations at sea; the correction of distance estimates to the school and a calibration of school size once detected. These two papers were well prepared and could stand alone. The third document was an analysis of the statistical estimation of population size from the newly corrected field measurements. While the analysis was well developed, the manuscript was still undergoing minor revisions during the meeting. The fourth document incorporated the corrected data and new methods into estimates of abundance for spotted and spinner dolphins. This document was completed, but required substantial clarification as many explanations of details were sometimes too terse to be followed by an external reviewer.

The first area of data improvement was Kinzey et al. (MS2001) in which the measures of distance and bearing to the dolphins were addressed. The field measurements determine bearing form angle rings and distance from reticles in the 25X binoculars. Radar was used to calibrate the observations and the precision of radial distance measurements using reticles was inversely proportional to target distance. They found that a method of calculating refraction from local air temperature and pressure information at the time of measurement provided the best fit of reticle to radar distances. When these environmental data were not available, they proposed used a regression-based, empirical correction term. This correction term was determined for each vessel and they were quite similar (0.00180 vs. 0.00171.). A smearing method to correct the horizontal angle by smearing was also discussed but not adopted.

The second area of data improvement was Gerrodette et al. (MS 2001) uses aerial photography from a helicopter to calibrate school size estimation. The analysis was based on approximately 2000 estimates of 370 dolphin schools in the eastern tropical Pacific. Estimates of school size were highly variable (Figure 1). Correction factors for each of 52 observers were developed. Also, the direct aerial estimate was used directly in 5% of the estimates. The amount and type of bias varied widely among observers, but there was a general tendency to underestimate, on average by about 12%. After calibration, there was a small tendency to overestimate school size on average 3%. The calibration procedure significantly reduced the mean square error. The method was applied retrospectively to all possible observations. The tendency to underestimate was not applied to observers who were not calibrated.

The area of analytical improvement was Forcada (MS 2001). This work contained a number or approaches, regression models, parametric and non-parametric covariate models and simulation studies. A comparison of traditional and robust regression techniques showed the former tends to underestimate mean group size. However, the robust method was dropped from subsequent analysis. The author concluded "Parametric and non-parametric multivariate models of the detection function incorporated into the line transect abundance estimator provide substantial improvement in estimating mean group size and abundance." Non-parametric methods, however, tend to be more biased as the number of covariates increases, and there is a need for an objective covariate selection to improve abundance estimates. Parametric methods provide the best trade-off of bias and precision as shown by simulations, and objective model selection allows for more reliable estimates of abundance and its variability. In the simulation, all four methods (traditional univariate regression, robust regression, parametric and non-parametric multivariate) tended to underestimate the mean group size. The simulation also gave a validation of the software developed for this analysis.

The integration of the re-calibrated data and newly developed methods was presented in Gerrodette and Forcada (MS2001), which produced new estimates of abundance for the two dolphin species under consideration. The input data were stratified line-transect surveys, which were carried out, in 12 different years within the period 1979 and 2000. A parametric covariate model was chosen of the detection process and group size and optimized for each survey. Variances and confidence intervals were estimated by bootstrap analysis. Table 1 below summarizes the models used each year and Figure 1 shows the resultant abundance estimates and their confidence intervals.

3. Review of Information and Results.

This review is presented in four sections answering the questions: 1) What/is the quality of the estimates of dolphin abundance, and are the estimates as accurate/precise as possible?; 2) Can we be sure?; 3) How could we be more sure?; 4) What are the implications for population modeling? The fourth question was not included in the Terms of Reference, but rather was posed by SWFC staff at the beginning of the review.

4.a Quality of estimates.

A plot of the new versus old estimates of abundance (Figure 1) show much tighter error bars and a general trend toward the long term average. The confidence intervals are also more symmetric. The means of each year's estimate are closer to the long-term mean for each species. Results that minimize year-to-year variation are consistent with the slow growth rate of dolphins adding credence to the new results.

Three parametric models were tried to fit the detection probability, half-normal, exponential power series and a hazard rate model. An example of their fits to the detection function data are shown in Figure 3. A non-parametric kernel fit was also tested and it is also shown. As well as testing models, various cofactors were tested. Table 2 shows the AIC for various covariates for Eastern spinner dolphins in 1998.

4.b Reported tests to ascertain quality of estimates.

The preparedness of supplied documents was generally good. While the presentations had additional slides, still more were needed for thorough review, especially for Forcada (MS2001) and Gerrodette & Forcada (MS2001), which contained more technical and operational details. Much detail was included, but more was needed to reflect the quantity and depth of the work carried out since the last review. For example, covariance matrices and AICs for covariates were supplied at the meeting in response to reviewer's requests.

Effect of unidentified sightings was reported to be small, but was not quantified. This issue was a source of controversy in earlier material (Anon 1999b) and the magnitude should be quantified. A plot of the stock mixing is shown to give an example of the degree of intermixing (Figure 4). It was reported that spatial analysis of these data has been initiated.

Software reliability is an issue when the analytical programs are developed specifically for project. The suit of programs continuously evolves and often there is not enough time for thorough testing at each step. Some guarantees of reliability were reported. Comparisons were made to other products (DISTANCE, St. Andrew's package and MARK from Colorado State). Tests using simulated data were done and reported, but should be carried out routinely to catch any accidentally introduced errors. Also standard diagnostics of fits, condition, and convergence criteria should be reported.

Confluence of results from several sources of development (group size calibration, reticle correction and analytical models) made it very difficult to define how to apportion the impact of each treatment. Testing and revision were still being performed during the visit to La Jolla. The ability to isolate cause/effects is important to assure that both critical treatments can get closer scrutiny and to assure that the effects are consistent (in sign and magnitude) with expert opinions and experience.

4.c Further tests for quality of estimates.

Three areas are indicated for further testing and assurance of reliability; an extension of the sensitively analysis mention in 4b, more comprehensive simulation studies and survey design.

A comprehensive sensitivity analysis should be performed. This would apply to the data corrections and models. The sensitivity would be in terms of the mean abundance estimates, their variance and where applicable an AIC.

The simulation study reported in Forcada (MS2001) should be enhanced. Although it performed well as a test of software integrity, other benefits are possible. Also, the description and results were too brief; the description was less than a page, no figures were produced and only one table of results. For example, how did the simulated data compare with real observations (Figures 3 and 5)? The simulation had one covariate, cue, which was not chosen as being the most important covariate for any year (Table 1) and had a worse AIC was of less importance in terms of AIC as is shownin Table 2. A simulation with a realistic covariate model description (say group size and/or distance) might be a better vehicle for model selection. Such a simulation could also act as test bed for kernel methods with respect to bandwidth to help chose between the two methods for choosing bandwidth posed. It might also address the problem of over-fitting reported for the non-parametric methods.

Of importance for both data analysis and simulation testing is the impact of rare large events. One example of this is in the application of histogram truncation in the probability of detection as a function of distance. It was anecdotally reported that in one year there was a very large school spotted just at the edge of detection and its inclusion/exclusion had an effect of increasing the abundance estimate by about 100,000 compared to an estimated of about 500, 000. Although ignoring such rare events improves the variance estimates in bootstrapping, some consideration must be given to their existence.

The description of the survey design was reported in Gerrodette & Forcada (MS2001, p3):

> Within each stratum, transect lines were randomly but not uniformly spaced, given the logistical constraints of ships range and speed. Ships moved at night, which contributed to some independence among daily transects.

It is not obvious from this description how much subjectivity is imposed and the degree to which the design is compromised. Smith (1999) raised issues about survey design and made recommendations on optimal and adaptive designs. Assessing the statistical implications of the current design should be addressed

A scatter plot was produced during the meeting of the mean square error (MSE) of calibrated observations of group size versus the number of sightings. After about 600 sightings, the MSE was constant at about 20% even though the data extended beyond 1200 sightings. This suggests that after a certain amount of experience performance is unchanged. If the 20% is not an irreducible error inherent in field observations, perhaps detailed feedback using aerial photography could enhance performance.

Linear correlations among the covariates were presented during the review. Although the signal to noise seems quite high, non-linear or process based models of the inter-relationship among the covariates should be considered.

4.d Implications for population modeling.

The strategy chosen for application of the new methods/data was to pick the best model for each year's survey. While giving individually optimized estimates, a consistent approach may produce a better time series for integration into population models. Examination of the individual models (See Table 1 below.) suggests that a consistent analytical model would not depart much from the annually optimized estimates. Even if the consistent treatment meant that the estimates were used as indices as opposed to absolute estimates, population models could estimate the scaling coefficient (efficiency) and a fairly restrictive prior could be imposed. An index of quality could also be imposed as a weighting to reflect the amount of calibration applied to each years data.

It may be possible to use the aerial data as a source of ancillary data for population modeling. These data may provide size or stage (pups, juveniles, mature animals) information at least at a gross level. Structured population data and models could improve estimates of survivorship and recruitment, allowing a closer examination of the intrinsic growth rate, which has been used to describe recovery.

The Tuna Vessel Observer Data (TVOD) have been collected for over 20 years but were removed from consideration for population modeling. I could not find clear documentation for the reasons for the removal, but at least part of the rationale seems to have been the divergence between the TVOD and the a trend suggested by the 1998 survey point. See Anon (1999b, Appendix 2.). The reworking of the 1998 data and the two subsequent years greatly changes this perception. The exclusion of the TVOD should be re-evaluated. Models with dynamic or time-dependent weighting may allow their inclusion into a population reconstruction.

4. **Review of Advice N/A**

5. **Recommendations (See Section 4)**

6. **Implications (See Section 4)**

7. References

Material provided (before and during La Jolla meeting)

Anon. 1999a. Report of the invited experts on the research vessel estimates of dolphin abundance in the ETP, 1998 and 1999 surveys

Anon 1999b. Report to Congress on the initial finding, required under the Marine Mammal Protection act of 1972… 25 March 1999.

Forcada, J. MS2001. Multivariate methods for size-dependent detection in conventional line transect sampling.

Gerrodette, T., W. Perryman, J. Barlow. MS2001. Calibrating group size estimates of dolphins in the eastern tropical Pacific Ocean.

Gerrodette, T., J. Forcada. MS2001. Estimates of abundance of dolphin stocks affected by the tuna purse-seine fishery in the eastern tropical Pacific Ocean

Kinzey, D., T. Gerrodette, D. Fink. MS2001. Accuracy and precision of perpendicular distance measurements in shipboard line-transect sighting surveys.

Smith, S.J. 1999. Review of the Scientific program and information relevant to the International Dolphin Conservation Program Act (IDCPA) held at the Southwest Fisheries Science Center, La Jolla, California, 8 to 11 March, 1999. Submitted to UM Independent System for Peer Reviews.

Appendices, Tables and Figures

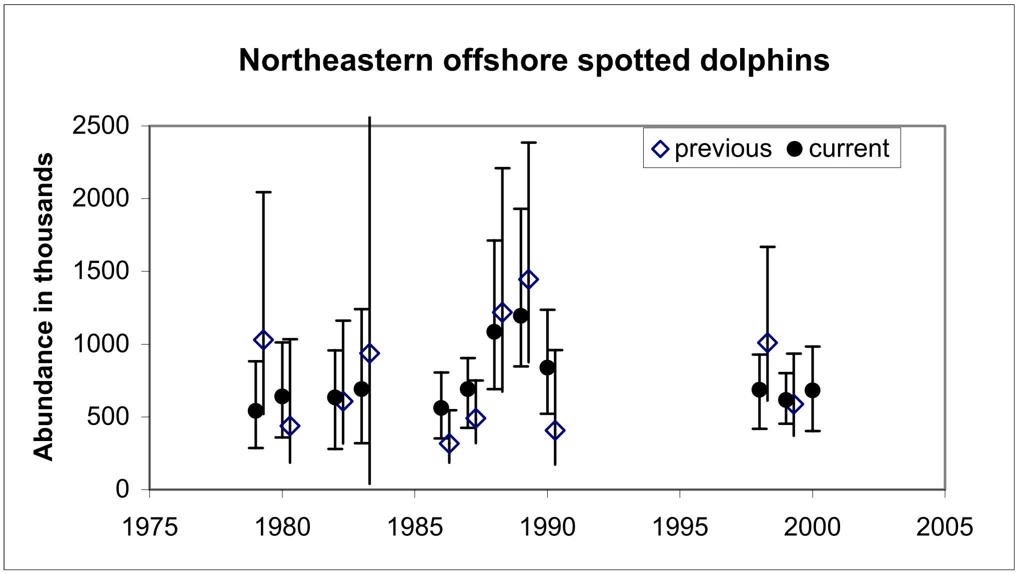
Table 1. Table of yearly models for analysis supplied by authors during meeting,

Year	Eastern spinner dolphin		Offshore spotted dolphin	
	Model	Covariates*	Model	Covariates*
1979	Half-normal	pd+gs pd	Half-normal	pd+gs pd
1980	Half-normal	pd+gs pd	Half-normal	pd+gs pd
1982	Half-normal	pd+gs pd	Half-normal	pd+gs pd
1983	Half-normal	pd+gs pd	Half-normal	pd+gs pd
1986	Half-normal	pd+gs	Half-normal	pd+gs+ship
1987	Half-normal	pd+gs	Half-normal	pd+gs+birds
1988	Half-normal	pd+gs	Half-normal	pd+gs+ship
1989	Half-normal	pd+gs	Half-normal	pd+gs+ship
1990	Half-normal	pd+gs	Half-normal	pd+gs
1998	Half-normal	pd+gs	Half-normal Hazard-rate	pd+gs pd+gs
1999	Half-normal/Hazard-rate Half-normal/Hazard-rate	pd+gs pd+gs+time	Half-normal Half-normal	pd+gs pd+gs+time
2000	Half-normal Half-normal	pd/pd+gs pd+bf	Half-normal Hazard-rate	pd+gs pd+gs+ship

*pd – perpendicular distance
gs – group size
ship – ship effect
time – time of day
bf – Beaufort scale

Table 2. Akaike information criterion (AIC) for various covariates using the half-normal model of detection probability. This is an abridged table of information supplied during review. Lower values are better and a difference in AIC of 2 or more indicates significant improvement.

Covariates	**AICc**
Univariate	279.64
Group size	273.93
Time	281.51
Visual cue	283.34
Group size + time	275.90
Group size + cue	290.03
Group size + time + cue	298.45



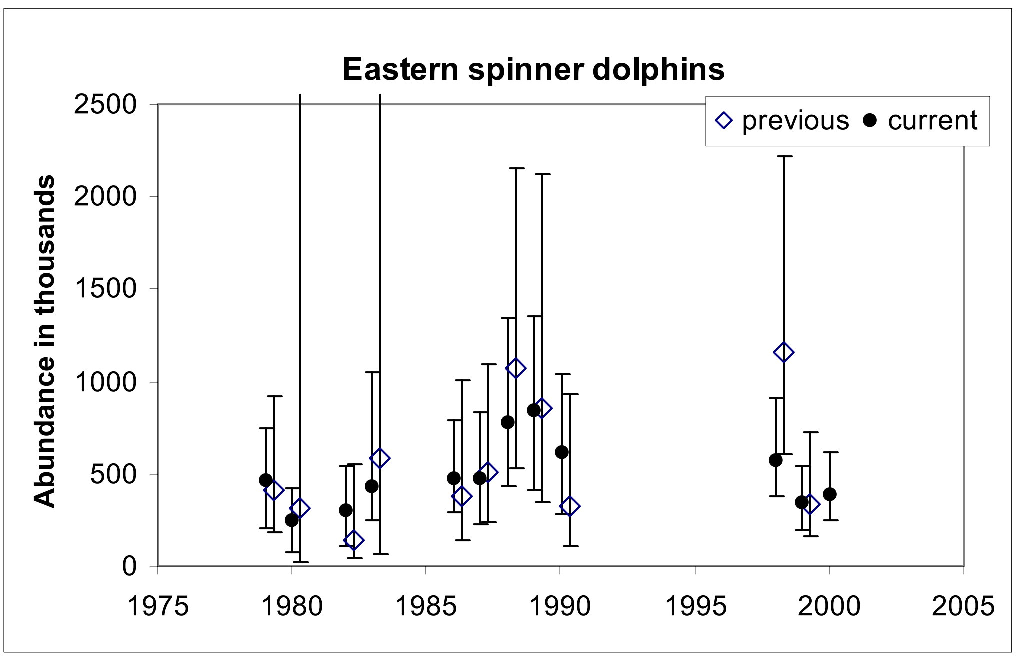


Figure 1 Supplied by SFSC staff during meeting. Comparison of abundance estimates using new data and analysis (current) with previous estimates.

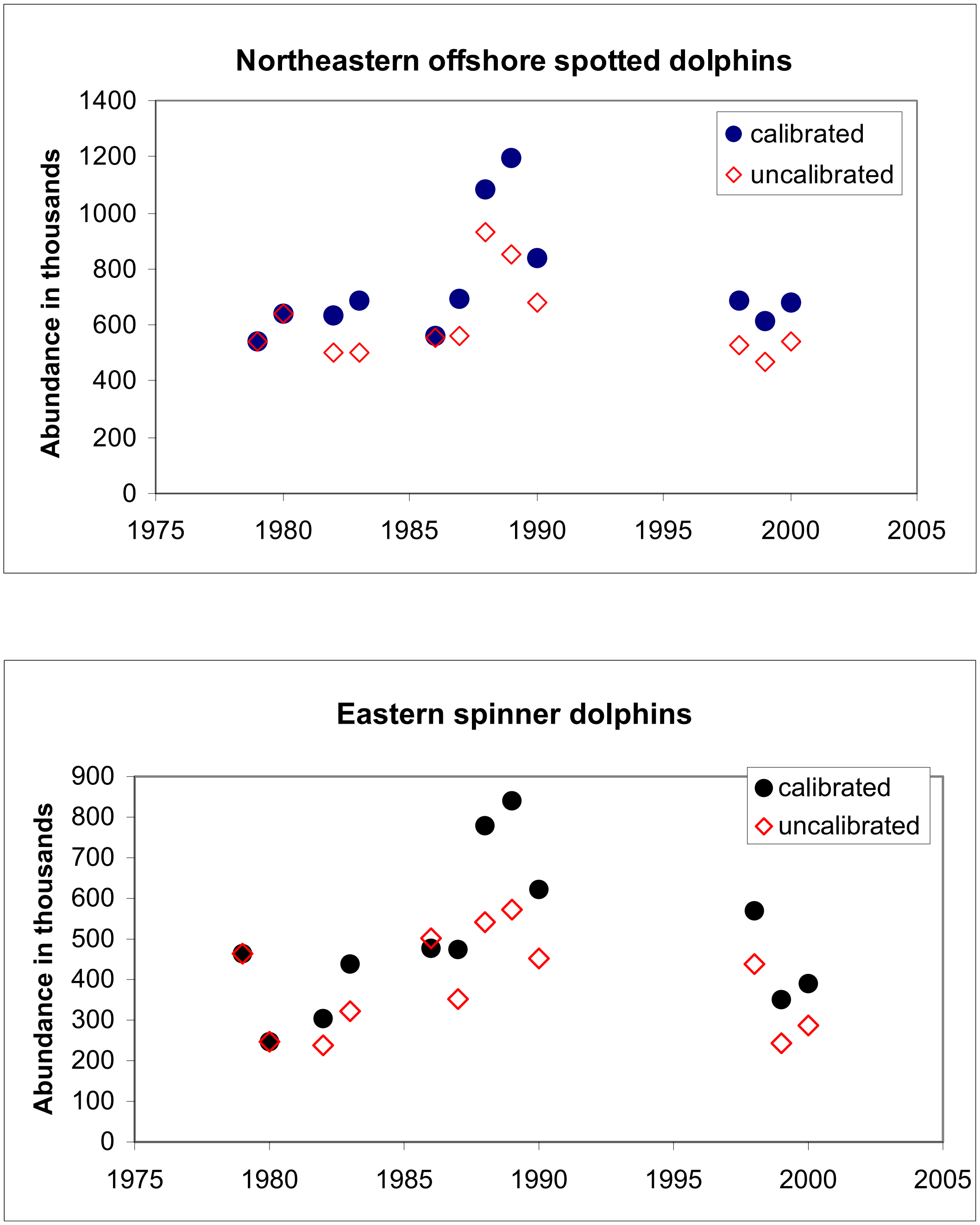


Figure 2 Supplied by SFSC staff during meeting. Effect of group size calibration on estimates of dolphin abundance with previous estimates.

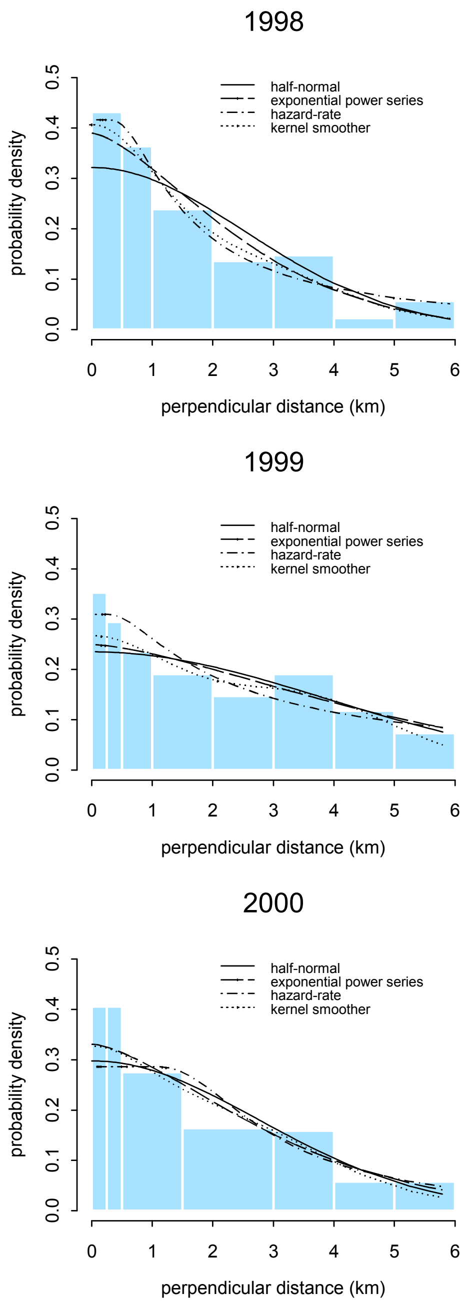


Figure 3. A comparison of models to fit of detection probability histograms. Presented by Forcada during the presentation of his document.

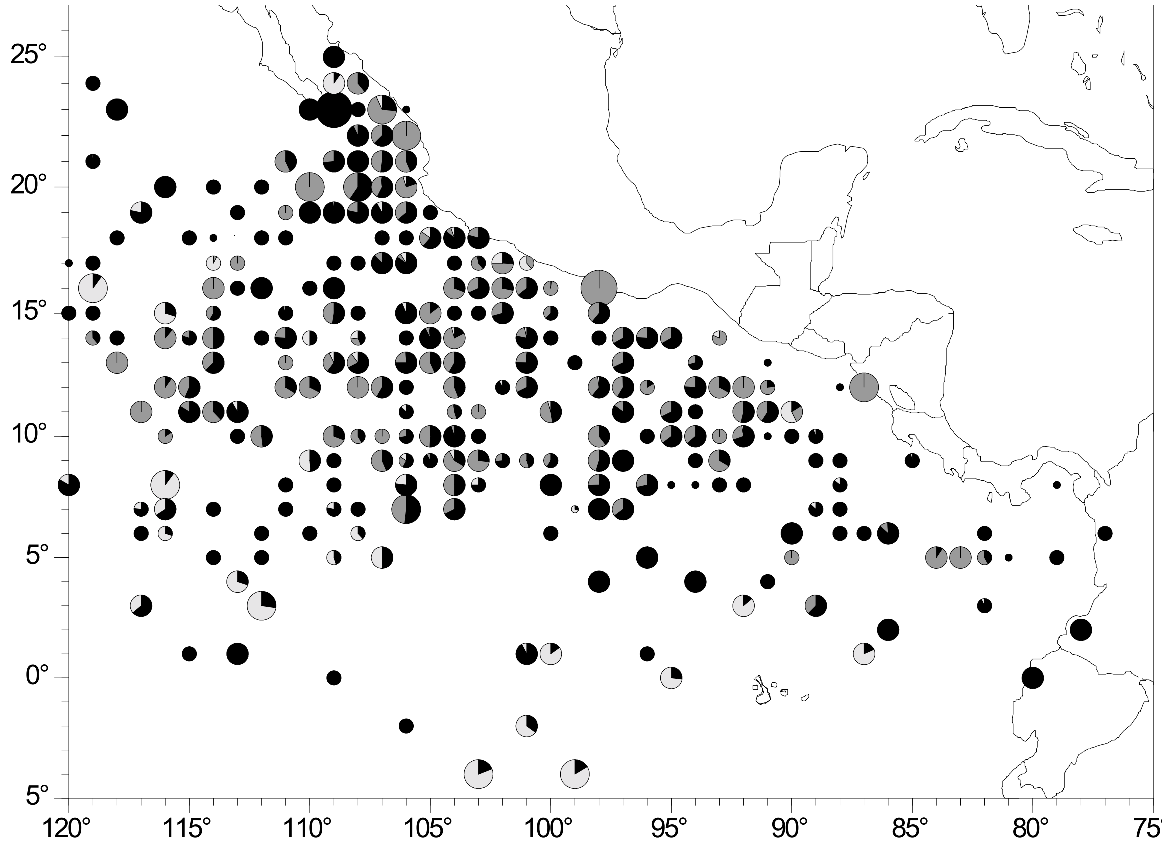


Figure 4. Expanding symbol plot of group size averaged over 1998-2000 surveys and aggregated to 1-minute squares. Each symbol is partitioned into spinner, (black), spotted (dark gray) and other (light gray) components. Based on data supplied by SFSC staff during meeting.

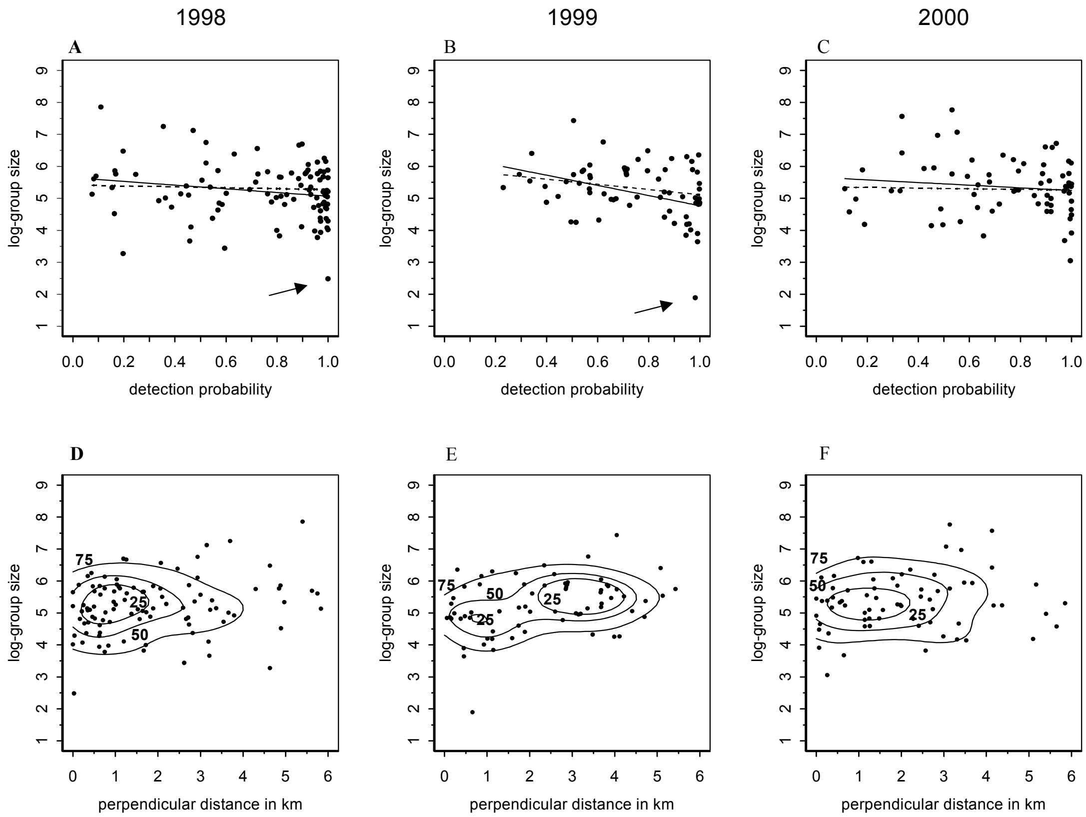


Figure 5. Data showing the interrelationship of group size, detection probability and distance. (From Forcada MS 2001, Figure 3)

Appendix 1.– Terms of reference

General

The topic of the review is the abundance of several species of tropical pelagic dolphins that associate with tuna and are killed in the eastern tropical Pacific purse-seine tuna fishery. In 1997, the US Congress proposed changing the definition of "Dolphin-Safe" tuna, but it made the change in definition contingent on the results of studies of the impact of the tuna fishery on depleted dolphin populations. Estimates of dolphin abundance based on cruises carried out in 1998-2000 form a central part of these studies. The tuna-dolphin issue is a controversial issue among NMFS, US tuna industry, foreign tuna industry, and environmental groups.

The main task of the consultant is to review the methods used to estimate abundance from line-transect data, including covariate detection models. The fact that these dolphins occur in a wide range of school sizes presents unique problems for the estimation of expected group size, so considerable effort has been devoted to this analysis. The expertise of the consultant should include knowledge of statistics and methods of population estimation, especially distance sampling (line-transect) methods.

Documents supplied to the reviewers will include draft manuscripts describing the covariate analysis, simulations to test the performance of several estimators, calibration of school size estimates, and assignment of partially identified sightings. Background papers will include previous relevant publications and reports. The raw data and software used in the analysis will be available to the reviewers if they wish.

Specific

The consultant's duties shall not exceed a maximum total of 2 weeks- several days to read all background documents, attend a three-day meeting with scientists at the NMFS La Jolla Laboratory, in San Diego, California, and several days to produce a written report of the findings. It is expected that the individual contribution of the consultant shall reflect the consultant's area of expertise; therefore, no consensus opinion (or report) will be accepted. Specific tasks and timings are itemized below:

1. Read and become familiar with the relevant documents provided in advance to the consultant;

2. Discuss background documents with scientists at the NMFS La Jolla Laboratory, in San Diego, CA, for 3 days, from October 15-17, 2001;

3. No later than November 16, 2001, submit a written report of findings, analysis, and conclusions. The report should be addressed to the "UM Independent System for Peer Reviews, " and sent to David Die, UM/RSMAS, 4600 Rickenbacker Causeway, Miami, FL