Aquat. Living Resour. 2018, 31, 12

https://doi.org/10.1051/alr/2017048

Aquatic Living Resources
Available online at:
www.alr-journal.org

Research Article

A probabilistic procedure for estimating an optimal echo-integration threshold using the Expectation-Maximisation algorithm★

Antonio López-Serrano[1,2], Héctor Villalobos[2,*] and Manuel O. Nevárez-Martínez[3]

[1] Universidad del Mar. Ciudad Universitaria. Puerto Ángel, Oaxaca 70902, México
[2] Instituto Politécnico Nacional – CICIMAR. Av. I.P.N. s/n. Col. Playa Palo de Santa Rita, La Paz, B.C.S. 23090, México
[3] Instituto Nacional de Pesca – CRIP Unidad Guaymas, Calle Sur No. 20 Col. Cantera, Guaymas, Sonora 85400, México

Received 23 October 2017 / Accepted 11 December 2017

Handling Editor: Verena Trenkel

Abstract – To obtain reliable fish biomass estimates by acoustic methods, it is essential to filter out the signals from unwanted scatterers (e.g. zooplankton). When acoustic data are collected at more than one frequency, methods that exploit the differences in reflectivity of scatterers can be used to achieve the separation of targets. These methods cannot be applied with historical data nor recent data collected on board fishing vessels employed as scientific platforms, where only one transducer is available. Instead, a volume backscattering strength (S_v) threshold is set to separate fish from plankton, both for echogram visualisation or, more importantly, during echo-integration. While empirical methods exist for selecting a threshold, it often depends on the subjective decision of the user. A −47 dB threshold was empirically established in 2008 at the beginning of a series of surveys conducted by Mexico's National Fisheries Institute to assess the biomass of Pacific sardine in the Gulf of California. Until 2012, when a 120 kHz transducer was installed, only data collected at 38 kHz are available. Here, we propose a probabilistic procedure to estimate an optimal S_v threshold using the Expectation-Maximisation algorithm for fitting a mixture of Gaussian distributions to S_v data sampled from schools associated with small pelagic fish and their surrounding echoes. The optimal threshold is given by the Bayes decision function for classifying an S_v value in one of the two groups. The procedure was implemented in the R language environment. The optimal threshold found for 38 kHz data was −59.4 dB, more than 12 dB lower than the currently used value. This difference prompts the need to revise the acoustic biomass estimates of small pelagics in the Gulf of California.

Keywords: S_v threshold / echo-integration / EM algorithm / small pelagic fish / Gulf of California

1 Introduction

When studying aquatic organisms by acoustic methods, an essential step consists in filtering out unwanted signals received by the echosounder (Simmonds and MacLennan, 2005). Besides background noise that can be estimated and compensated (De Robertis and Higginbottom, 2007), unwanted echoes come from groups of organisms other than the target species. For example when dense plankton aggregations are encountered during acoustic surveys whose interest is on pelagic fish. Usually, an S_v threshold (volume backscattering strength, dB re 1 m^{-1}) is set to a value high enough to separate fish from plankton (Madureira et al., 1993; Fernandes et al., 2006). This threshold can be adjusted for visualisation purposes during data acquisition and, more importantly, applied during post processing in the echo-integration step. In the first case, the chief scientist's preferred visualisation S_v value affects only the echogram's appearance, but provides an idea of what is being observed during the survey and to make decisions about where a fishing haul should be carried out. This visualisation threshold is considerably higher than the one used for data acquisition, which is generally around −100 or −80 dB. In the second case, the chosen echo-integration threshold influences directly the shape of detected schools (Diner, 2001) and the nautical area scattering coefficient (s_A, m^2 nmi^{-2}) that ultimately will be transformed into fish biomass (MacLennan et al., 2002).

★ The R-code and data (sampled S_v values and echograms) used in this work are available from Villalobos et al. (2018), and can be obtained through SEANOE.

*Corresponding author: hvillalo@ipn.mx

It is accepted that no S_v threshold would be able to discern in an unmistakable manner the target species from nontarget species, and therefore the goal should be to optimally find the balance between keeping most of the echoes of target species and less of the unwanted species (Parker-Stetter et al., 2009).

In many cases, the selection of a suitable echo-integration threshold relies more on the subjective experience of the user, than on a generally established objective procedure (Eckmann, 1998). In the majority of the reviewed literature, while S_v thresholds are reported, no selection process or method is mentioned. Some authors acknowledge that the threshold was empirically selected, or based on field observations, without providing further explanations.

An empirical procedure, first described by Eckmann (1998), consists in obtaining, for a target fish species and unwanted backscatterers (e.g. a zooplankton layer) in an echogram, the s_A values resulting from a set of S_v values. The total (fish + zooplankton) and fish only s_A values are then plotted against the S_v values, and the threshold is chosen at a point where the s_A values of the target group begin to drop. A variation of this method, published by Jech and Michaels (2006), used the s_A averaged over several distance intervals (0.5 nmi) and normalized to the maximum mean s_A. These authors compared the s_A for the entire water column with that from visually chosen herring regions.

Another method involves knowing the minimum expected target strength (*TS*) for the target species. From this, the S_v of smaller targets is excluded, leading to a depth dependent S_v threshold. The minimum *TS* can be derived from in situ measurements (e.g. Gastauer et al., 2016), ex situ distributions or theoretical models. This procedure has been implemented in commercial software (see Parker-Stetter et al., 2009, for details).

Different approaches, perhaps more precise, are possible when data is collected at more than one acoustic frequency, which nowadays is becoming current practice on board Research Vessels. New methods exploiting the differences in reflectivity patterns of biological target groups, according to the frequency in which they are insonified have been developed (see for instance Fernandes et al., 2006, and references therein). Combinations of distinct acoustic frequencies enhance the differences between types of scatterers and then, though not explicitly stated, a cut off value separating the groups can be selected from the histograms of the combined S_v data from distinct frequencies.

Despite the promising results from bi- or multi-frequency methods, many existing datasets have been collected in the past using only one frequency, and therefore those methods cannot be applied. The same situation holds for data recently collected by fishing vessels that have been implemented as scientific platforms (Fässler et al., 2016; Melvin et al., 2016). Mexico's National Fisheries Institute (INAPESCA) has conducted a series of acoustic surveys to assess the biomass of Pacific sardine in the Gulf of California, and from 2008 to 2011, before another frequency was incorporated (120 kHz), only data collected at 38 kHz are available. Furthermore, the S_v echo-integration threshold currently in use was empirically established at the beginning of the series and has not been re-evaluated since.

In this work, we revisit the pertinence of this threshold value by exploring the applicability of techniques borrowed from digital image processing, in particular a probabilistic method that makes use of the Expectation-Maximisation algorithm (EM) (Dempster et al., 1977), which is an unsupervised clustering algorithm. We show that an optimal S_v threshold can be found using random samples taken from fish schools and the surrounding echoes. The proposed procedure is easy and fast, and importantly, fully implemented in free and open source software.

2 Methods

2.1 Acoustic sampling and fishing hauls

All acoustic data were obtained from digital echograms recorded from May 1 to 23 2013 in the Gulf of California, Mexico, during an acoustic survey targeting small pelagic fishes on board the R/V "BIP XI" owned by INAPESCA. A Simrad EK60 scientific echosounder with two hull mounted split beam transducers (38 and 120 kHz) was used. In this work we focused only on the 38 kHz (12° circular beamwidth), given that this frequency has been used since the beginning of INAPESCA's survey series, and is the most commonly employed frequency for acoustic biomass estimation of small pelagics. The echosounder was calibrated with a 38.1 mm tungsten carbide sphere according to standard procedures (Simmonds and MacLennan, 2005). During the survey, pulse duration was set at 512 ms, while transmit power was 1000 W. The ping rate was variable, according to bottom depth (<50 m = 0.25 s; 50–100 m = 0.5 s; 100–150 m = 0.75 s; 150–200 m = 1 s; >200 m = 2 s). Either parallel or zigzag transects of variable length, depending on the continental shelf width, were sailed at eight knots (kn) from dusk to dawn (18:00 to 06:00 local time). This strategy was adopted in accordance with the behaviour of the small pelagics fishery in the Gulf of California, where the purse seine fleet operates at night (Quiñonez-Velázquez et al., 2000) during the "oscuro", a new moon centred time period lasting 22 to 26 days (Nevárez-Martínez et al., 2014). Every night, an average of three fishing hauls aimed to capture putative small pelagic fish schools and other detected echo traces were carried out using a mid-water trawl with an approximately 16 m horizontal and 12 m vertical opening, a mesh size of 38 mm at the mouth and 19 mm at the codend. On average, hauls had a 30 min duration at a speed of three kn navigated in the opposite direction to the prospection, targeting the aggregations just previously detected. The species composition and the size distribution of the more abundant species in the catch were obtained for a subsample of each haul. For this study, only fishing hauls with a total catch >15 kg and a predominance of small pelagic fish species (more than 70% as compared to other species) were considered. While this choice is arbitrary, it allowed to keep more hauls in the analysis.

2.2 Acoustic data processing

The raw acoustic data were recorded and converted to HAC (hydroacoustics file format, ICES, 2005) with the Simrad ER60 acquisition software. Before further processing, detected bottom depth in the echograms was inspected for errors and manually corrected when necessary using Movies+ software (Berger et al., 2005). From the corrected echograms for the whole survey,

echogram segments ranging from 1 to 2 nautical miles (nmi) immediately before the beginning of each one of the previously selected fishing hauls were chosen. HAC files with selected echograms were imported into the R programming language environment (R Core Team, 2017) by means of the readHAC package (Kristensen, 2017). The rest of the procedures described below were coded entirely in the R environment, and for this purpose, a series of functions were developed and integrated into the R package **echogram**, available [1] from the Comprehensive R Archive Network. After importing the acoustic data, echogram's functions were used as needed for merging data from different HAC files into a unique R data structure and trimming to particular areas of interest. Also, echogram visualisations were produced discarding in all cases the first 5 m from the transducer to avoid surface noise. In these echograms, the underlying data matrix is mapped, so every pixel represents precisely one S_v data bin having as (x, y) coordinates the ping number (and corresponding ping time) and the sample depth.

2.3 S_v data sampling

In the haul-associated echograms, visually identifiable fish schools (*sensu* Reid, 2000), presumably from small pelagic species due to their persistence when observed with a high visualization threshold (e.g. $S_v = -50$ dB) and their association to positive catches of such species, were considered for the following sampling procedure.

In each echogram, approximately 180 S_v values from randomly selected pixels were sampled, 90 from the acoustic schools and 90 from outside the schools. In the second case, care was taken to avoid sampling isolated fish echoes which may be possibly located at the periphery of the schools. These samples were labelled accordingly as "schools" or "other". The R function developed for echogram sampling allows to "point and click" in a given echogram image and returns the ping number, ping time, sample depth and S_v value of the corresponding bin in the underlying data matrix. It also allows to sample the same bin in corresponding echograms from different frequencies. The maximum depth of the samples was <100 m to minimize the influence of background noise.

Aiming to find the S_v threshold allowing to separate small pelagic schools from the generally weaker surrounding echoes, the applicability of optimal global thresholding, a segmentation technique from digital image processing, was tested.

2.4 Image segmentation

Image segmentation consists in subdividing an image into regions or objects based on pixel intensity, the S_v in this case. When the S_v values of the objects of interest (e.g. schools) in an echogram are sufficiently different from the surrounding echoes, a global threshold can be applicable to an entire echogram.

Inspection of the histogram of S_v values in an echogram may show groups with different means, and the threshold would be the S_v value that separates best the two groups. Gonzalez and Woods (2008) gave a basic global thresholding algorithm easy to implement that performs well when a clear distinction exists between modes in a histogram.

[1] https://CRAN.R-project.org/package=echogram

2.5 Optimum global thresholding

Thresholding can also be regarded as a statistical-decision problem aiming to minimize the mean error when assigning pixels to different groups. The Bayes decision rule is the known closed-form solution to this problem, and requires two parameters, namely the probability density function (PDF) of the pixel intensities of each group and the occurrence probability of each group (Gonzalez and Woods, 2008). The implementation of this solution has been considered a difficult task, giving the difficulty in estimating the PDFs, among other reasons (Gonzalez and Woods, 2008). A common simplification of the problem is to assume Gaussian distributions for the data, which in our case seems a reasonable assumption. The Bayes decision function (d) for classifying an S_v value (x) in one of two groups (ω_j, $j = 1, 2$) with Gaussian PDFs, has the form:

$$d_j(x) = p(x|\omega_j)P(\omega_j) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} P(\omega_j) \quad j = 1, 2$$

where the normal density corresponds to the probability that x comes from group ω_j, and $P(\omega_j)$ is the prior probability, equal to 0.5 if both groups are equally likely to occur. The optimal threshold will be the S_v value, denoted here as x_0, such that $d_1(x_0) = d_2(x_0)$ and $p(x_0|\omega_1) = p(x_0|\omega_2)$, which is the intersection between the PDF curves (Gonzalez and Woods, 2008).

The problem of estimating the parameters of the PDFs can be approached using the EM algorithm (Dempster et al., 1977), widely used when modelling heterogeneous data through mixture distributions. This technique constitutes a class of model-based unsupervised clustering which enables the description and distinction of subgroups, even in the absence of an indexing variable giving the identity of the observations (Benaglia et al., 2009b).

The EM algorithm aims to iteratively compute in an alternating two step procedure, maximum-likelihood estimates of the unknown parameters, which in our case, correspond to the mean and variance of the PDFs and the mixing proportions of the distributions. In the expectation step, current estimates of the parameters are used to obtain a function of the expectation of the log-likelihood, while in the maximisation step the parameters that maximise the expected log-likelihood previously found, are computed. Besides the original paper of Dempster et al. (1977), the algorithm is described in detail in the literature dealing with models for mixtures of distributions (McLachlan and Basford, 1988; McLachlan and Peel, 2000).

In this work, we used the implementation of the EM algorithm for mixtures of normal distributions in the mixtools R package (Benaglia et al., 2009b), which requires only the vector of S_v data, from which the PDFs parameters (means and standard deviations), mixing proportions and posterior probabilities are estimated.

The EM algorithm was applied to three sets of raw S_v values from an echogram with conspicuous schools at the surface: (1) all data excluding sea bottom echoes; (2) data from the surface to 100 m depth; (3) an area (400 pings × 40 m depth) enclosing the surface schools. The algorithm was also applied to similar data sets after averaging the S_v data using a 5 × 5 kernel by means of functions in the **mmand** R package (Clayden, 2017). Additionally, the sampled raw

Table 1. Catch composition (%), date and geographic location of the selected fishing hauls for which associated echograms were analysed. The category "Other spp." includes fish species from families Triglidae, Carangidae, Ariidae, Scombridae and Myctophidae, among others, occasionally present in the hauls.

Haul no.	Date (local time)	Longitude	Latitude	Total catch (kg)	*Engraulis mordax*	*Etrumeus teres*	*Opisthonema spp.*	*Sardinops sagax*	*Scomber japonicus*	Other spp.
4	04/05/2013 01:31	−112.52	28.82	350	100	–	–	–	–	–
8	07/05/2013 20:46	−113.42	29.51	34.8	71.9	2.0	–	0.1	–	25.9
11	08/05/2013 19:38	−113.11	29.00	36.2	5.5	6.1	–	55.8	32.6	–
12	09/05/2013 20:45	−113.53	28.95	18.8	–	22.3	8.5	56.4	1.6	11.2
14	10/05/2013 03:33	−113.00	28.79	25.1	99.6	–	–	–	–	0.4
15	10/05/2013 23:26	−112.93	28.47	120	100	–	–	–	–	–
17	11/05/2013 22:25	−112.78	28.02	350	–	–	–	100	–	–

S_v values from the seven selected echograms were pooled into a single sample regardless of their category (i.e. "schools" or "other"), the histogram was plotted and the EM algorithm was also applied.

In all previous cases, a two-component normal-mixture model was fitted. Where applicable, the optimal global threshold was established as the S_v value at the intersection of the fitted PDFs. As noted before, this value represents the point where posterior probabilities are equal (0.5).

The performance of the estimated threshold was tested by comparing the s_A values obtained using this threshold, and the empirical value of −47 dB currently used for small pelagic fishes in the Gulf of California by INAPESCA's fisheries research station at Guaymas. For this purpose, a portion of the survey approximately 100 nmi long was selected and echo-integrated by layer (5–150 m depth) and 1 nmi elementary sampling units (ESU). The difference in percentage among the estimated and empirical threshold was computed.

3 Results

3.1 Biological data

From the total number of fishing hauls carried out during the survey, seven dominated by small pelagic fish species were selected for this study (total catches between 18 and 350 kg). The northern anchovy (*Engraulis mordax*) and Pacific sardine (*Sardinops sagax*) were the species best represented. Three hauls were monospecific for northern anchovy and one for Pacific sardine. Northern anchovy represented also almost 72% in another trawl, while Pacific sardine accounted to more than 55% in two others. Chub mackerel (*Scomber japonicus*), round herring (*Etrumeus teres*) and thread herring (*Ophistonema spp.*), all small pelagic species, were also represented in the catches (Tab. 1). The category "Other spp." in Table 1 regroups fish species from the families Triglidae, Carangidae, Ariidae, Scombridae and Myctophidae, among others, that were occasionally present in the hauls.

3.2 Acoustic data

Figure 1 shows the echogram for which the EM algorithm was applied to the three different sets of data depicted in the left panel (a): the whole echogram (excluding echoes below the sea bottom); the first 100 m from the surface; and a rectangular area enclosing the surface schools. The locations of the random samples taken from the schools and their surroundings in this particular echogram are represented as well. The same echogram after averaging is shown in the right panel (b), for which the EM algorithm was also applied to the same sets as before.

In Figure 2, the histograms for each set of raw (left panels) and averaged (right panels) S_v data are shown. In the histograms of the two larger sets (Fig. 2a, b, d, e) a dominant mode is visible at around −75 dB and another barely perceptible close to −50 dB. In these four cases the EM algorithm was not able to detect the smaller distribution. In contrast, for the subsets corresponding to the rectangular area enclosing the schools (Fig. 2c, f), the second mode became conspicuous and was detected by the EM algorithm. While the PDFs seemed to fit reasonable well, data overlap was important for the raw S_v (Fig. 2c); it persisted when the two modes became more separated in the averaged data (Fig. 2f). A closer inspection of the PDFs fitted to the histograms indicated that the mean of the stronger echoes should probably be larger than estimated, and hence, the threshold should be higher. For the raw S_v data the threshold found was −54.2 dB, while for the averaged S_v data it was −60.4 dB.

These results showed that limiting the extent of analysed data to an area of interest increased the representativeness of the S_v values of the schools which improved the performance of the EM algorithm. Also, echogram smoothing improved the separation of the modes and, by reducing data variability the distributions became leptokurtic, probably producing departures from normality.

3.3 Optimal threshold from raw S_v samples

Concerning the sampling procedure, a total of 1209 S_v values were taken from the echograms associated to the seven selected fishing hauls, 634 for category "other" and 575 for category "schools". In one echogram only 47 samples were taken from the schools due to their small size.

In Figure 3a, histograms of the combined S_v samples taken from the seven echograms are represented by category. While two modes are clearly visible, a non-negligible overlap is evident as a result of the spread of both distributions.

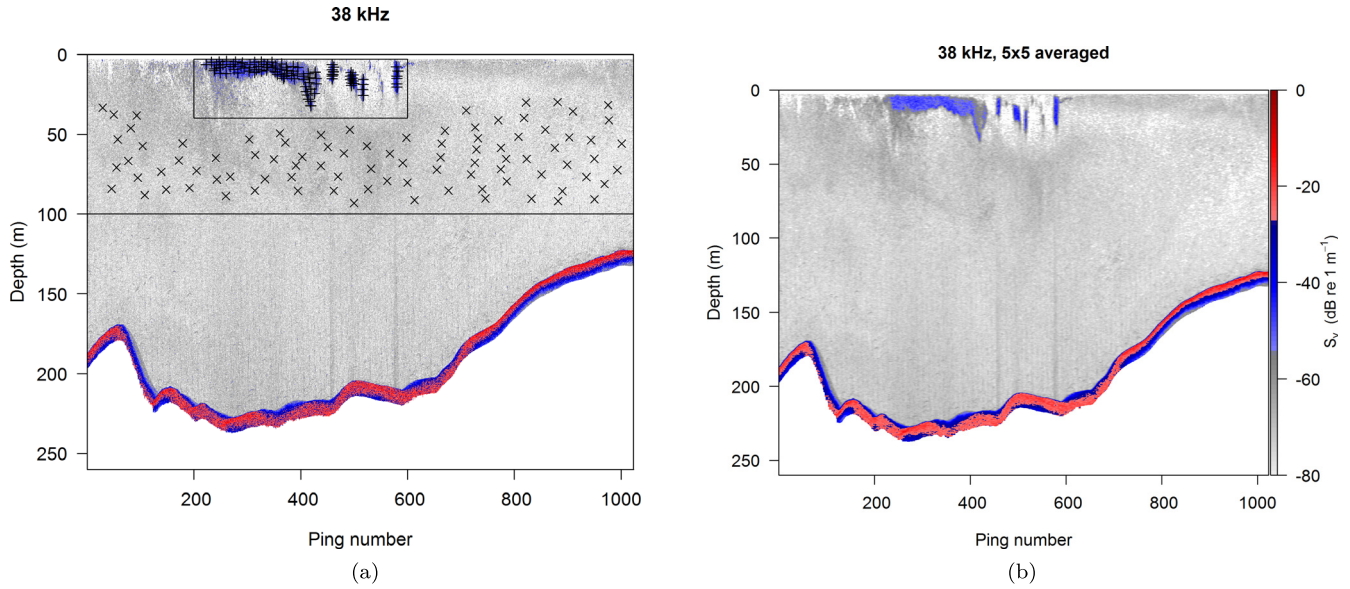

Fig. 1. (a) Raw S_v echogram associated to haul no. 4 for which the EM algorithm was applied to three sets of data: the whole echogram excluding echoes below the sea bottom; the first 100 m from the surface, delimited by the horizontal line; and the rectangular area enclosing the schools. The locations of the random samples taken from the schools (+) and their surroundings (×) are also represented. (b) Averaged echogram using a 5 × 5 kernel for which the EM algorithm was also applied to the same sets of data as in (a). The colour scale is the same for both echograms.

The histogram in Figure 3b represents the pooled S_v samples regardless of their category, along with the Gaussian PDF curves fitted by the EM algorithm for each component identified. The optimal threshold, −59.4 dB, is the value where the posterior probabilities are equal. This value is represented by the black vertical line and corresponds to the intersection of the curves.

The estimated parameters for both distributions as well as the final mixing proportions are shown in Table 2. The diffrence between means was almost 30 dB, with a mean of −76.44 dB for category "other", and −47.05 for the category "schools". The standard deviations for each category were 9.48 and 7.04 dB, respectively.

The posterior probabilities for the two components identified by the EM algorithm in relation to S_v values are plotted in Figure 2. This plot illustrates the complementary nature of the posterior probabilities. Besides the optimal threshold, vertical dotted lines were drawn at the 0.99, 0.95 and 0.90 probabilities. From this plot and Table 3, values of $S_v < -68.9$ had a 0.99 probability or more of being in the category "other", while at −59.4 this figure droped to 0.5. Conversely, the "schools" probability increased from 0.5 to 0.9 between −59.4 and −54.1 dB.

Examples of echograms using the optimal global threshold of −59.4 dB are shown in Figure 4b, d, f alongside with their raw S_v values (c and e). In Figure 4a, a threshold of −60.4 was used which was estimated from the averaged S_v data shown in Figure 2f.

The comparison of s_A values obtained by echo-integration using the −59.4 dB optimal threshold, and the −47 dB empirical value currently used by INAPESCA, revealed that applying the higher threshold value this coefficient might be underestimated by approximately 10% for large s_A values (>4000 m^2 nmi^{-2}). For the more frequent small s_A values the underestimation may be 80% or more.

4 Discussion

In this work, the applicability of optimal global thresholding using the EM algorithm to find an echo-integration S_v threshold was demonstrated. The approach is suitable but not limited to situations when only one acoustic frequency is available. We consider that the best estimate was achieved with the pooled raw S_v values sampled from the schools and their surroundings for all analyed echograms.

While the method could be applied to a whole echogram, the large proportion of weak echoes masks the stronger school values, which in turns makes it harder for the EM algorithm. By focusing on a smaller echogram area enclosing the schools, this problem is partially solved. However, given that the echoes are potentially produced by scatterers with different acoustic properties, and that the observed response is a continuous variable, the overlap of signals remains important.

Smoothing the echogram by averaging the S_v values helped by reducing the natural variability (Fernandes et al., 2006; Sato et al., 2015), and a valley separating the modes became conspicuous in the histograms. Nevertheless, a side effect of echogram smoothing is that the distribution of averaged S_v data becomes leptokurtic, possibly affecting the fitting of Gaussian PDFs by the EM algorithm. The threshold found with the averaged S_v for the rectangular area enclosing the schools seemed correct, and is indeed only one dB lower than the optimal threshold, but from Figure 2f, it appears that the mode of the schools and consequently the threshold, should be higher, as was mentioned in the results. Moreover, it can be argued that a threshold obtained from a small area in an echogram could not represent an entire survey unless some kind of averaging of different estimates from several echograms was done.

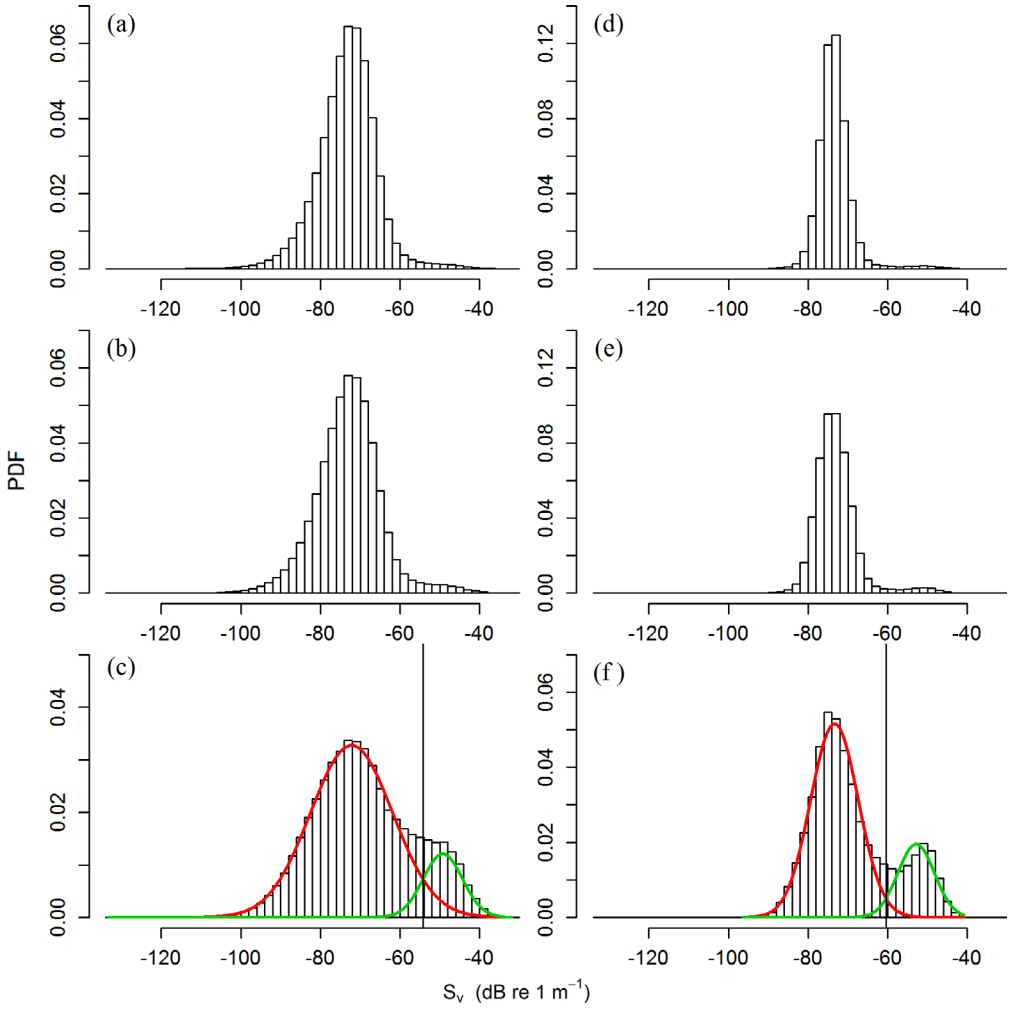

Fig. 2. Histograms for the different sets of data associated to haul no. 4 explained in Figure 1. The left panel histograms correspond to (a) the raw S_v data for the whole echogram (excluding sea bottom); (b) the first 100 m from the surface; and (c) the rectangular area enclosing the schools. The histograms of the averaged S_v data for the same sets are represented in the right panels (d, e and f). The red and green lines in the bottom histograms are the Gaussian PDFs fitted by the EM algorithm to the weak and strong (likely the schools) echoes. The black vertical lines in (c) and (d) represent the optimal thresholds found for both sets, −54.2 and −60.4 dB, respectively.

Random sampling focused on several schools was helpful to circumvent the two situations mentioned above. On the one hand, by sampling only the structures of interest and, assuming that the echoes in each category were homogeneous, or at least more alike as when compared between categories, data variability was reduced. On the other hand, by taking the same number of samples from both structures, the histogram became more balanced, with modes about the same size, helping the PDF fitting. Besides, by pooling the samples from many structures of the same general type found in a survey, the histograms and therefore the estimated threshold, are representative of the survey.

Gonzalez and Woods (2008) mentioned that for this probabilistic approach it can be difficult to estimate the PDFs for the categories analysed. The EM algorithm implementation in R (Benaglia et al., 2009b) proved to be useful for this purpose, in particular the model based on mixtures of normal distributions. Even when the sampled raw S_v data for both categories showed departures from normality, we decided parsimoniously to use this model instead of a non-parametric algorithm (Benaglia et al., 2009a) because those departures were not important, as suggested by the graphical fit of the estimated PDFs (Fig. 3b).

Besides returning estimates of the means and standard deviations for "schools" and "other" categories, as well as their mixing proportions, the EM algorithm provided the vectors of posterior probabilities from which the optimal threshold was found. Posterior probabilities arise from the Bayes decision function and represent the probability of membership to either category, not that a given S_v value is assigned to one group or the other (Anderson et al., 2007). This function is optimal because it minimises the classification error (Gonzalez and Woods, 2008).

The EM algorithm has been extensively used as a statistical tool in unsupervised classification problems since the late 1970s (McLachlan and Peel, 2000). More recently, there have been applications to fisheries acoustics multi-frequency data, for classifying fish and invertebrates

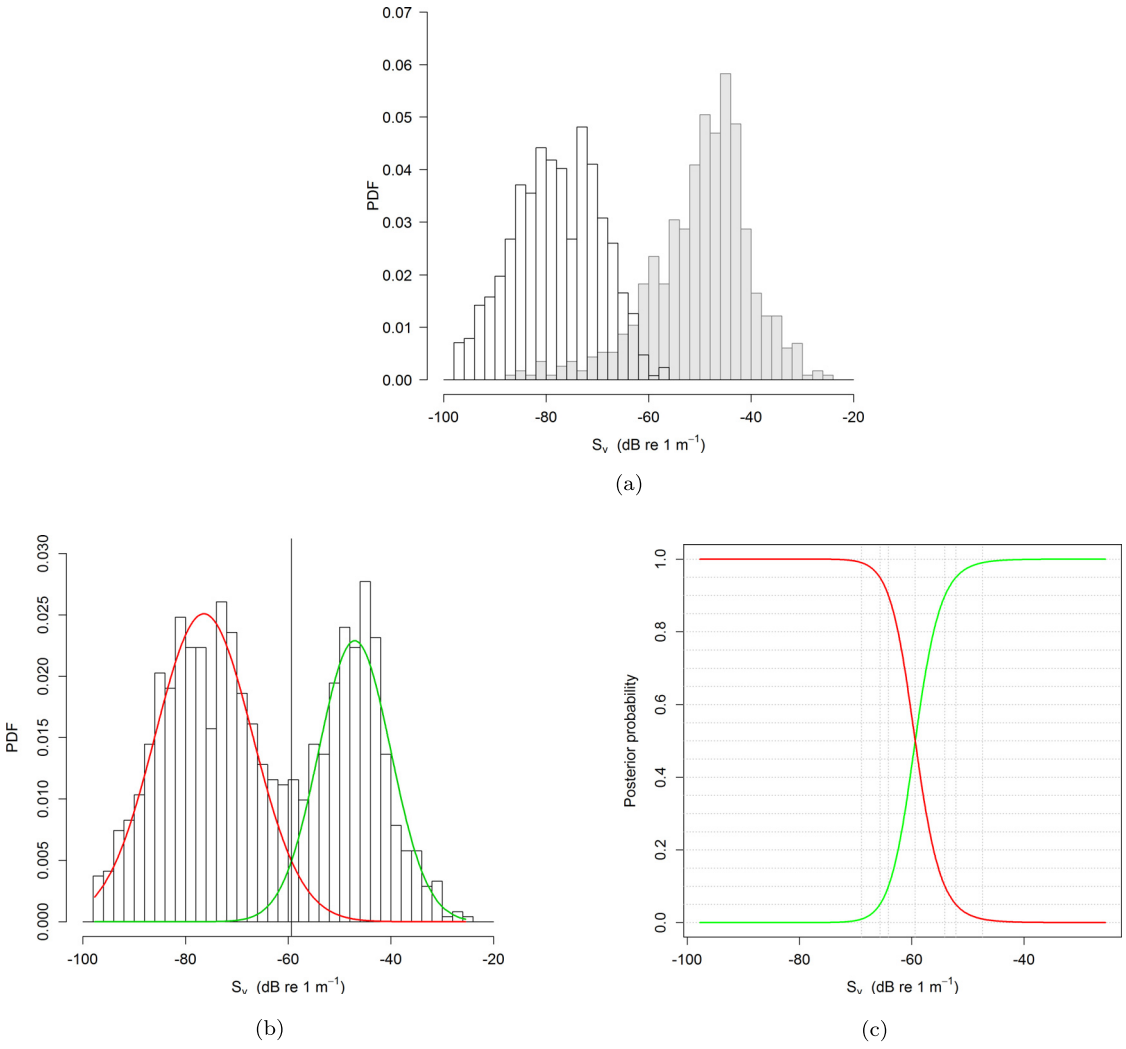

Fig. 3. (a) Histograms of sampled raw S_v values from schools (grey bars) and from category "other" (empty bars) for seven echograms associated to fishing hauls with small pelagic fish catches. (b) Histogram of pooled S_v data regardless of category with the Gaussian PDFs fitted by the EM algorithm. The vertical black line at the intersection of the curves represents the optimal −59.4 dB threshold. (c) Posterior probabilities of assigning a given S_v value to each group. The red and green lines in (b) and (c) represent, respectively, the "other" and "schools" categories.

Table 2. Estimated parameters of the probability density functions of S_v values for components "schools" and "other", and the final mixing proportions (lambda).

	"Schools"	"Other"
Mean	−47.05	−76.44
Std. dev.	7.04	9.48
Lambda	0.404	0.596

Table 3. Complementary posterior probabilities for the categories "schools" and "other", and the corresponding S_v value.

"Schools"	"Other"	S_v
0.01	0.99	−68.93
0.05	0.95	−65.65
0.10	0.90	−64.14
0.25	0.75	−61.80
0.50	**0.50**	**−59.40**
0.75	0.25	−56.78
0.90	0.10	−54.11
0.95	0.05	−52.13
0.99	0.01	−47.41

(Anderson et al., 2007;Woillez et al., 2012), and discriminating juvenile and adult fish school clusters (Fablet et al., 2012). In this work, we demonstrate the feasibility of using this tool to estimate a probabilistic threshold with single frequency data. Moreover, the EM algorithm can be applied when more than two groups are present in the data and for multivariate classifications (Benaglia et al., 2009b).

As to the identity of the sampled school structures, while it is not possible to individually identify each observed aggregation, on the basis of depth and catch composition of

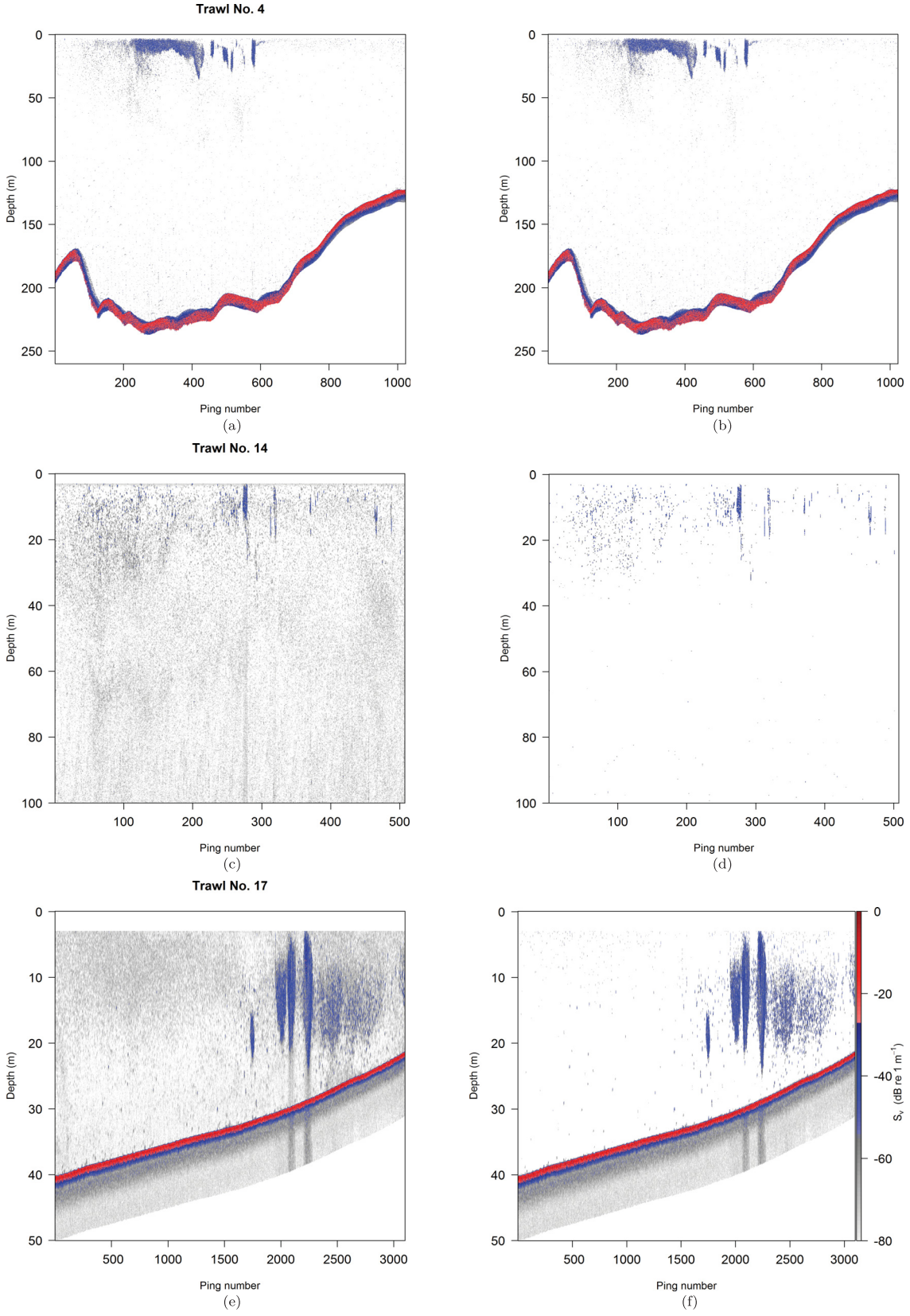

Fig. 4. Examples of echograms associated to fishing hauls with small pelagic fish catches. (a) and (b) are, respectively, the thresholded versions of the raw echogram in Figure 1a using −60.4 (from Fig. 2f) and the optimal −59.4 dB estimated from the raw S_v samples. (d) and (f) are the thresholded versions of (c) and (e) using a −59.4 dB threshold. The colour scale is the same for all echograms.

the fishing hauls, we assume that they were representative of the small pelagic fish community of the Gulf of California. The echograms analysed in this work were associated to fishing hauls located nearby the "Grandes Islas" region, an area well known for high primary productivity (Álvarez-Borrego, 2010) and also as an important fishing ground for the small pelagic fishery fleet of the Gulf of California (Nevárez-Martínez et al., 2014). This was consistent with the catch composition of the hauls, where northern anchovy and Pacific sardine were the most important species. No attempt was made to assess the effect on the threshold of differences in species dominance in the catch, instead, we focused on estimating a threshold applicable to the entire survey.

With regard to the category "other", it is clear that it may comprise a wide variety of scatterers with different morphologies and reflective patterns, and some of them, like siphonophores, presenting gas filled structures (Warren et al., 2001) which give them a high reflective capacity (Stanton et al., 1994, 1996) and the ability to generate S_v values comparable to fishes (Trevorrow et al., 2005). Siphonophores are a common group in the Gulf of California, for which 24 species have been reported (Gasca and Suárez, 1991). So we cannot exclude the possibility of high S_v values being sampled outside the schools, however the general trend should not be greatly affected. Also, isolated fishes could as well have been present, specially at night, when schools tend to disperse and mix with the zooplankton scattering layers to feed (Helfman, 1986). All this may account to a certain degree for the overlap of S_v values observed for the pooled histogram.

The −59.4 dB threshold found in this work is more than 12 dB lower than the empirical −47 dB threshold currently used by INAPESCA for their spring Pacific sardine acoustic biomass assessments in the Gulf of California. According to the comparison of s_A values resulting from both thresholds, current estimates might be underestimated by 10% in ESU's where schools are present. In the same region, Domínguez-Contreras et al. (2012) used a −50 dB threshold for biomass estimation of small pelagic fish, though for 120 kHz. Citing previous work on the Pacific coast of Baja California Sur (Robinson et al., 2007), these authors mentioned that the threshold was selected in accordance with "the lowest S_v value recorded from Pacific sardine schools during previous surveys", but no further details were given. Our threshold falls within the −60.7 dB to −41.5 dB range associated by Gregg and Horne (2009) with sardine and anchovy aggregations in Monterey Bay, California using 120 kHz.

At 38 kHz and for Atlantic herring (*Clupea harengus*), another small pelagic species, a similar value has been used (−60 dB) in the Baltic Sea (Peltonen and Balk, 2005), but Jech and Michaels (2006) used −66 dB on Georges Bank. S_v thresholds of −65 dB and −70 dB have also been used to differentiate schools of small pelagic fishes off the coast of South Africa (38 kHz, Lawson et al., 2001) and Florida (208 kHz, Churnside et al., 2003). Other values reported in the literature, used in algorithms for school detection of several species at 38 kHz, are between −65 and −60 dB (Petitgas et al., 1998; Reid, 2000; Burgos and Horne, 2007).

We consider that the probabilistic selection of an optimal S_v threshold suitable for echo-integration as described here has advantages over the graphical method based on the plot of s_A vs. S_v (Eckmann, 1998; Jech and Michaels, 2006), where the threshold choice still depends on a subjective decision. With respect to the procedure proposed by Parker-Stetter et al. (2009), while it might be more adequate given that it considers the *TS* for the species of interest, it requires additional knowledge (*TS* vs. length relationship) and data (actual *in situ TS* values, for instance) to be correctly applied. The practical implementation of the procedure described here is straightforward and can be accomplished in the R language environment, which offers the additional advantage of providing access to other statistical procedures. Finally, while the optimal threshold is clearly defined when the posterior probabilities area equal (0.5), both Figure 3c and Table 3, could be used as decision tools for choosing an echo-integration threshold based on statistically informed criteria.

Acknowledgements. We thank the crew members of the R/V "BIP XI", technicians and research staff of INAPESCA's *Centro Regional de Investigación Pesquera* (CRIP *Unidad Guaymas*) who participated in the survey. We also thank François Gerlotto for his useful comments to an earlier version of the manuscript, and two anonymous reviewers who contributed to improve our work. HV was supported by COFAA-IPN and EDI-IPN.

References

Álvarez-Borrego S. Physical, chemical, and biological oceanography of the Gulf of California, in: C.R. Brusca (Ed.), The Gulf of California: biodiversity and conservation, University of Arizona Press, Tucson, AZ, 2010, pp. 24–48.

Anderson CIH, Horne JK, Boyle J. 2007. Classifying multi-frequency fisheries acoustic data using a robust probabilistic classification technique. *JASA Express Lett* 121: EL230–E L237.

Benaglia T, Chauveau D, Hunter DR. 2009a. An em-like algorithm for semi-and nonparametric estimation in multivariate mixtures. *J Comput Graphi Stat* 18: 505–526.

Benaglia T, Chauveau D, Hunter DR, Young D. 2009b. mixtools: an R package for analyzing mixture models. *J Stat Softw* 32: 1–29.

Berger L, Durand C, Marchalot C, Diner N. 2005. Movies + user manual version 4.3, Tech. Rep. DNIS/ESI/DLE/DTI/00-051, IFREMER.

Burgos JM, Horne JK. 2007. Sensitivity analysis and parameter selection for detecting aggregations in acoustic data. *ICES J Mar Sci* 64: 160–168.

Churnside JH, Demer DA, Mahmoudi B. 2003. A comparison of lidar and echosounder measurements of fish schools in the Gulf of Mexico. *ICES J Mar Sci* 60: 147–154.

Clayden J. 2017. mmand: mathematical morphology in any number of dimensions. R package version 1.5.0. URL https://CRAN.R-project.org/package=mmand

De Robertis A, Higginbottom I. 2007. A post-processing technique to estimate the signal-to-noise ratio and remove echosounder background noise. *ICES J Mar Sci* 64: 1282–1291.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc Ser B* (Methodol.) 39: 1–38.

Diner N. 2001. Correction on school geometry and density: approach based on acoustic image simulation. *Aquat Living Resour* 14: 211–222.

Domínguez-Contreras JF, Robinson CJ, Gómez-Gutierrez J. 2012. Hydroacoustical survey of near-surface distribution, abundance and biomass of small pelagic fish in the Gulf of California. *Pac Sci* 66: 311–326.

Eckmann R. 1998. Allocation of echo integrator output to small larval insect (*Chaoborus sp.*) and medium-sized (juvenile fish) targets. *Fish Res* 35: 107–113.

Fablet R, Gay P, Peraltilla S, Peña C, Castillo R, Bertrand A. 2012. Bags-of-features for fish school cluster characterization in pelagic ecosystems: application to the discrimination of juvenile and adult anchovy (*Engraulis ringens*) clusters off Peru. *Can J Fish Aquat Sci* 69: 1329–1339.

Fässler SM, Brunel T, Gastauer S, Burggraaf D. 2016. Acoustic data collected on pelagic fishing vessels throughout an annual cycle: operational framework, interpretation of observations, and future perspectives. *Fish Res* 178: 39–46.

Fernandes P, Korneliussen R, Lebourges-Dhaussy A, Massé J, Iglesias M, Diner N, Ona E, Knutsen T, Gajate J, Ponce R. 2006. The SIMFAMI project: species identification methods from acoustic multi-frequency information, Tech. rep., Final Report to the EC no. Q5RS- 2001-02054.

Gasca R, Suárez E. 1991. Nota sobre los sifonóforos (Cnidaria: Siphonophora) del Golfo de California (agosto-septiembre, 1977). *Cienc Pesq Mex* 8: 119–125.

Gastauer S, Scoulding B, Fässler SMM, Benden DPLD, Parsons M. 2016. Target strength estimates of red emperor (*Lutjanus sebae*) with Bayesian parameter calibration. *Aquat Living Resour* 29: 301.

Gonzalez RC, Woods RE. Digital image processing, 3rd Edition, Pearson Prentice Hall, Upper Saddle River, NJ, 2008.

Gregg MC, Horne JK. 2009. Turbulence, acoustic backscatter, and pelagic nekton in Monterey Bay. *J Phys Oceanogr* 39: 1097–1114.

Helfman GS. 1986. Fish behaviour by day, night and twilight, in: T.J. Pitcher (Ed.), The behaviour of Teleost fishes, 1st Edition, Croom Helm Ltd., London, pp. 366–387.

ICES. 2005. Description of the ICES *hac* standard data exchange format, version 1.60, Tech. Rep. 278, ICES Cooperative Research Report.

Jech JM, Michaels WL. 2006. A multifrequency method to classify and evaluate fisheries acoustics data. *Can J Fish Aquat Sci* 63: 2225–2235.

Kristensen K. 2017. readHAC: read Acoustic HAC Format, R package version 1.0. URL https://CRAN.R-project.org/package=readHAC

Lawson GL, Barange M, Fréon P. 2001. Species identification of pelagic fish schools on the South African continental shelf using acoustic descriptors and ancillary information. *ICES J Mar Sci* 58: 275–287.

MacLennan DN, Fernandes PG, Dalen J. 2002. A consistent approach to definitions and symbols in fisheries acoustics. *ICES J Mar Sci* 59: 365–369.

Madureira LS, Everson I, Murphy EJ. 1993. Interpretation of acoustic data at two frequencies to discriminate between antarctic krill (*Euphausia superba* Dana) and other scatterers. *J Plankton Res* 15: 787–802.

McLachlan GJ, Basford KE. Mixture models: inference and applications to clustering, Vol. 84 of Statistics: textbooks and monographs, Marcel Dekker, New York, 1988.

McLachlan GJ, Peel D. Finite mixture models, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, 2000.

Melvin GD, Gerlotto F, Lang C, Trillo P. 2016. Fishing vessels as scientific platforms: an introduction. *Fish Res* 178: 1–3.

Nevárez-Martínez MO, Martínez-Zavala M, Jacob-Cervantes ML, Cotero-Altamirano CE, Santos-Molina JP, Valdez-Pelayo A. Peces pelágicos menores, in: L.F.J. Beléndez-Moreno, E. Espino-Barr, G. Galindo-Cortes, M.T. Gaspar-Dillanes, L. Huidobro-Campos, E. Morales-Bojórquez (Eds.), Sustentabilidad y Pesca Responsable en México, Evaluación y Manejo, 1st Edition, SAGARPA – Instituto Nacional de Pesca, Mexico City, 2014, pp. 87–139.

Parker-Stetter SL, Rudstam L, Sullivan P, Warner D. Standard operating procedures for fisheries acoustic surveys in the Great Lakes, Great Lakes Fisheries Commission Special Publication, Ann Arbor, MI, 2009.

Peltonen H, Balk H. 2005. The acoustic target strength of herring (*Clupea harengus* l.) in the northern Baltic Sea. *ICES J Mar Sci* 62: 803–808.

Petitgas P, Diner N, Georgakarakos S, Reid D, Aukland R, Massé J, Scalabrin C, Iglesias M, Muiño R, Carrera-López P. 1998. Sensitivity analysis of school parameters to compare schools from different surveys: a review of the standardisation task of the EC-FAIR programme CLUSTER. *ICES Documents CM 1998/J*: 23.

Quiñonez-Velázquez C, Nevárez-Martínez MO, Gluyas-Millán MG. 2000. Growth and hatching dates of juvenile Pacific sardine *Sardinops caeruleus* in the Gulf of California. *Fish Res* 48: 99–106.

R Core Team. R: a language and environment for statistical computing, R foundation for statistical computing, Vienna, Austria, 2017. URL https://www.R-project.org/

Reid DG. Report on echo trace classification, ICES cooperative research report 238, ICES, Copenhagen, Denmark, 2000.

Robinson CJ, Gómez-Aguirre S, Gómez-Gutiérrez J. 2007. Pacific sardine behaviour related to tidal current dynamics in Bahía Magdalena, México. *J Fish Biol* 71: 200–218.

Sato M, Horne JK, Parker-Stetter SL, Keister JE. 2015. Acoustic classification of coexisting taxa in a coastal ecosystem. *Fish Res* 172: 130–136.

Simmonds J, MacLennan D. Fisheries acoustics: theory and practice, Fish and aquatic resources series, 2nd Edition, Blackwell Science Ltd., Ames, Iowa, 2005.

Stanton TK, Wiebe PH, Chu D, Benfield MC, Scanlon L, Martin L, Eastwood RL. 1994. On acoustic estimates of zooplankton biomass. *ICES J Mar Sci* 51: 505–512.

Stanton TK, Chu D, Wiebe PH. 1996. Acoustic scattering characteristics of several zooplankton groups. *ICES J Mar Sci* 53: 289–295.

Trevorrow MV, Mackas DL, Benfield MC. 2005. Comparison of multifrequency acoustic and *in situ* measurements of zooplankton abundances in Knight Inlet, British Columbia. *J Acoust Soc Am* 117: 3574–3588.

Villalobos H, López-Serrano A, Nevárez-Martínez MO. 2018. Volume backscattering strength samples and echograms (38 kHz) associated to small pelagic fish schools in the Gulf of California, SEANOE, Mexico. http://doi.org/10.17882/53034.

Warren J, Stanton T, Benfield M, Wiebe P, Chu D, Sutor M. 2001. *In situ* measurements of acoustic target strengths of gas-bearing siphonophores. *ICES J Mar Sci* 58: 740–749.

Woillez M, Ressler PH, Wilson CD, Horne JK. 2012. Multifrequency species classification of acoustic-trawl survey data using semi-supervised learning with class discovery. *J Acoust Soc Am* 131: EL184–E L190.

Cite this article as: López-Serrano A, Villalobos H, Nevárez-Martínez MO. 2018. A probabilistic procedure for estimating an optimal echo-integration threshold using the Expectation-Maximisation algorithm. *Aquat. Living Resour.* 31: 12