Fisheries Research 200 (2018) 49–60

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Review

Robust clustering methodology for multi-frequency acoustic data: A review of standardization, initialization and cluster geometry

Marian Peña

Instituto Español de Oceanografía, Centre Oceanogràfic de Balears, Moll de Ponent s/n, 07015 Palma, Spain

ARTICLE INFO

Handled by Prof. George A. Rose

Keywords:
Cluster geometry
Correlation
Frequency response
Standardization
Variance
K-Means
EM clustering

ABSTRACT

Clustering is a useful unsupervised technique for the identification of acoustic groups in multi-frequency echograms based on frequency response. K-Means is the most well-known clustering technique but has significant requirements such as clusters of equal size and spherical shape. Initialization is a common problem in clustering as only local minima are usually guaranteed, and thus initialization must locate the centroids near the global minimum. Expectation-Maximization (EM) clustering also requires a good set of initial centroids but allows the identification of clusters with different statistical distributions. This work presents the comparison of these techniques applied to a case with several acoustic signatures presenting different cluster sizes and distributions. The main issues treated in this manuscript are: pre-processing of acoustic data for clustering, initialization of centroids with theoretical scattering models and the need to consider the geometry of the clusters in addition to means, including variance (spread around the mean), orientation (correlation between variables), spherical or ellipsoidal shape (difference in variance between variables) and cluster size (number of observations). EM clustering is the only technique that properly separates acoustic signatures (and noise) after using the supervised initialization presented in this study.

1. Introduction

Fisheries acoustics is a discipline that examines fishes and plankton species based on their scattering properties using the measured scattering intensity known as volume backscatter (S_v, dB re m^{-1}) (Simmonds and MacLennan, 2005). The identification of acoustic echotraces has traditionally been conducted through net sampling, known as 'ground-truthing'. However, linking acoustic and net data is complicated due to, among other things, net avoidance and acoustic shadowing of species with lower scatter. Net sampling of deep-distributed species such as mesopelagic fish often challenges the available logistics. In addition, sampling in acoustic surveys are often directed at schools/layers with higher scatter, as echotraces of lower numerical density or those that contain species with lower scatter are more difficult to spot. A priori knowledge of the location of different species or acoustic typologies in the echogram allows the proper sampling of all the desired targets (when biological information is also needed), and may be used to make commercial fishing more efficient, reducing bycatch. The identification of acoustic groups based on acoustic data without ground-truthing requires the employment of an unsupervised technique. Ideally, a quick and not very computationally demanding methodology is desired, such as clustering.

Clustering is an unsupervised machine learning technique that groups data according to similarity in the variables provided as input. As an unsupervised method, there is no training data with labels orientating the algorithm to a particular solution. Several papers have summarized the main clustering techniques (Banerjee and Davé, 2012; Xiao and Yu, 2012), which can be divided into hard-clustering, where one data point can only belong to one cluster, and fuzzy or soft clustering, where each data point may belong to several clusters through a membership function. The second group handles better overlapping clusters and is less sensitive to noise as noise influence is equally split among groups.

The most well-known clustering techniques have been designed for data without noise or outliers (Xiao and Yu, 2012). Robust variations have been posteriorly developed to adapt to real measurement data that contains noise. As shown in this paper, most clustering algorithms must also be robust for initialization (initial centroid estimation). Furthermore, the geometric characteristics of the data used is often overlooked, such as cluster size and shape. For instance, the most popular algorithm, K-Means, requires data with clusters of equal size and variance (spherical clusters). Different clustering algorithms or distance measures can lead to very different results (Jain et al., 2004). There is no single algorithm suitable for all applications and thus, data knowledge and requirement checking would reveal the most suitable. This work focuses on that analysis for fisheries acoustic data.

E-mail address: marian.pena@ba.ieo.es.

https://doi.org/10.1016/j.fishres.2017.12.013
Received 30 September 2017; Received in revised form 18 December 2017; Accepted 21 December 2017

The incorporation of several frequencies into fisheries and plankton acoustics gave birth to what it is known as multi-frequency methods, where the difference between frequencies is employed to identify acoustic groups, comparing their spectrum with theoretical scattering models. Species are categorized into three acoustic groups: gas-bearing (including a swim bladder or pneumatophore), fluid-like (with a weak acoustic signal, such as krill and copepods), and elastic shell (pteropod type) (Stanton et al., 1996). The first group presents a resonance peak at a particular frequency that depends on swim bladder size (near 18 kHz for lantern fish and around 4 kHz for small pelagic fish). The second and third groups present increasing scatter with frequency shifted in frequency with length. For vessel-borne echosounders, S_v is measured within a volume that increases with depth. Assuming only one acoustic typology is present in the volume, S_v is dependent on the scatter of one single organism (target strength, *TS*) and its numerical density ρ, following the equation $S_v = \mathrm{TS} + 10*\log_{10}(\rho)$. To remove numerical density dependence, each S_v is subtracted by the S_v of a reference frequency, usually 38 kHz for historical reasons (as it was the most common first frequency onboard research vessels). The results are known as Frequency Response $\mathrm{FR} = S_{vi} - S_{v38} = \mathrm{TS}_i - \mathrm{TS}_{38}$, which reduces the number of variables to the number of frequencies minus one (as *FR*(38) will be equal to 1 for all data points, and thus, will have little influence on the clustering; see discussion for further information). Typical working frequencies are 18, 38, 70, 120, 200 and 333 kHz but, as the usable range (depth if vertically orientated) decreases at higher frequencies, the number of frequencies that can be employed depend on the depth of the targeted species. S_v data are thus a type of curve data, like time series, where the trend (with frequency instead of time) is used to identify groups, but unlike time series, frequency is a dependent variable, while time is not (Pereira, 2013). The dependence of S_v values with frequency (serial correlation) has been modeled for the different acoustic groups. See, for example, Peña and Calise (2016) for the krill model adapted to short-length species and Peña et al. (2014) for mesopelagic fish models. As in time series, frequency shifts are bound to appear, due to length differences of organisms (reflected in the *TS* value), as well as vertical offset due to numerical density differences ($10 * log_{10}(\rho)$ term). Calculating the *FR* removes that offset and achieves some translation invariance, in a similar way that it is done for detrending in time series. The frequency shift is minimal for similar sizes, but could be the key to differentiate different species with similar *FR* tendency, but very different size, such as krill (~2–4 cm) and Mysidacea (~0.5–2.5 cm). The frequency spectrum (*FR* variation with frequency) has to be maintained in pre-processing and considered in the clustering.

In fisheries acoustics, data noise is often classified as background noise and impulse noise (Ryan et al., 2015). Background noise refers to ambient and vessel noise that affects all pings and varies in intensity and pattern with vessel speed, propeller pitch, bottom depth, number of vessels in the area, etc. (Peña, 2016). Impulse noise is usually caused by interferences with another acoustic device and affect a few pings. Several algorithms have been published to remove background and impulse noise (Ryan et al., 2015; Peña, 2016). Data with very low threshold also include white noise, a random signal having equal intensity at different frequencies. They are a sequence of serially uncorrelated random data with zero mean and finite variance. This noise needs to be accounted for when modeling acoustic data. The sample unit considered in this paper is the pixel, i.e. each data point in the 2D echogram as sampled by the echosounder. For an EK60 with 1 ms pulse duration, a pixel has a vertical length of ~19 cm. The horizontal length changes with beam width and depth due to the conical shape of the acoustic beam. For a 7° beam, the horizontal length is ~12 m at 100 m depth and ~30 m at 500 m. Each pixel represents a particular sampled volume that changes with distance to the transducer and beam angle. Differences in sampled volume between frequencies need to be accounted for when comparing pixels, particularly in cases of small echotraces.

The aim of this paper is to study the behavior of clustering techniques with multi-frequency acoustic data, very noisy data with clusters that can have very different sizes (proportion of echogram pixels). A very robust initialization procedure based on theoretical models that properly locates centroids and provides an estimation of the number of clusters is presented. The use of standardization is also analyzed. The paper is organized as follows: a short summary of clustering methods and their requisites is given, focusing on two techniques: K-Means (KM) and Expectation-Maximization (EM) clustering (also known as Gaussian Mixture Model or GMM). KM and EM clustering have already been used with acoustic data (see Section 1.3) and are both included in the top ten algorithms in data mining (Wu et al., 2008). The geometry of clusters is defined and shown with examples. A review of clustering applied to multi-frequency acoustic data is then given. The material and methods section presents the novel technique to initialize centroids. Finally, the two techniques are compared using a challenging example and the suggested initialization method.

1.1. Clustering review

Clustering techniques can be classified based on the clustering approach as center-based techniques, where one cluster is represented by its center, such as K-Means (Lloyd., 1982); density-based clustering like DBSCAN (Arlia and Coppola, 2001), where clusters are defined as areas of higher density surrounded by lower density areas; and distribution-based techniques, with clusters defined as objects belonging to the same distribution. Gaussian Mixture models fitted with an Expectation-Maximization (EM) algorithm (Krishnan and McLachlan, 1997) are included in the last category, and allow clusters to have different variances, density and size. Density-based clustering also allows the separation of clusters of different size, but requires the calculation of distances between all pair of data points, which is too computationally expensive with acoustic data.

Two of the critical aspects of clustering techniques are the pre-allocation of number of clusters and initialization of the centroids. Preselecting the number of clusters K is still a very challenging problem in clustering. The available techniques to estimate K are based on comparing different runs of the algorithm, which make them cumbersome. Even though several cluster validity indices (CVIs) exist, they are inefficient when clusters widely differ in density or size (Zalik, 2010). They are usually based on maximizing compactness and minimizing overlap among clusters, but in the presence of noise, overlapping is prone to appear. Distances between centroids do not take into account the cluster shape and dispersion; points from two neighboring but not dispersed clusters can be more separated than two spread clusters that overlap, despite the distance between the centroids being large. Using only centroid information (such as with the Davies-Bouldin measure (DB) (Davies and Bouldin, 1979), the Hartigan index (Ha) (Hartigan, 1975) or the Krzanowski-Lai index (KL) (Krzanowski and Lai, 1988)) is not sufficient to interpret the geometrical structure of the data, and therefore not sufficient for the separation between clusters. The elbow method, one of the most common CVIs based on the variance curve, was found to be unsuitable for several datasets in Milligan and Cooper (1985) and, as seen in Santos and Embrechts (2014) with 30 benchmark datasets, no cluster validation index is perfect.

In general, clustering algorithms guarantee convergence to the closest local minima, so the initial location of the centroids must ensure that this minimum is also the global minimum. MacQueen (1967) suggested choosing K random observations as initialized centroids, but different initialization runs may generate rather different clusters and more dense clusters have a higher probability to attract one or two centroids.

Center-based techniques assume all clusters are spherical (equal variance-covariance). Often standardization/normalization (centering each variable to 0 and scaling by its standard deviation or range) is used to equal the variance of all variables. Multifrequency echograms often

present differences in variance, as each frequency has a different sensitivity to noise and directivity among other things, but as shown below, standardization alters the *FR* spectrum. In addition, standardization works globally on the dataset, but even after applying it, clusters may present different distributions. The cluster size (or prior probability) is often required to be equal, i.e. each cluster has a roughly equal number of observations. This is often not true with acoustic data, as in the example presented.

1.1.1. K-Means (KM)

K-Means or Lloyd's algorithm (Lloyd., 1982) is the most popular clustering algorithm. The steps of K-Means are:

1. Randomly choose K items and make them initial centroids.
2. For each point, find the nearest centroid and assign the point to that cluster.
3. Update the centroid of each cluster as the mean average of the observations in that cluster.
4. Repeat steps 2 and 3 until no point switches clusters or the maximum number of iterations is reached.

KM uses hard membership, i.e. each data point is assigned to exactly one cluster.

1.1.2. Expectation-Maximization (EM) clustering

Expectation-Maximization (EM) clustering (Krishnan and McLachlan, 1997) is a distribution-based clustering method where clusters are defined based on how likely the objects included are to belong to the same distribution. Overfitting is overcome by constraining the algorithm with a specific number of Gaussian distributions (Gaussian mixture models).

EM clustering is a soft-membership clustering technique that allows clusters to have different sizes and statistical distributions. Instead of maximizing the differences in means between clusters, the EM clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions. The objective of this clustering algorithm is to maximize the overall probability or likelihood of the data, given the resulting clusters. KM is a variant of EM clustering, with the assumption that clusters are spherical (with identical variance-covariance matrices for each cluster, assuming Gaussian distribution).

Each iteration includes two steps (Expectation and Maximization) and the algorithm finishes when the distribution parameters converge or reach the maximum number of iterations.

E-Step: In the E-step, data are estimated given the observed data and current estimates of model parameters. This step estimates the probability of each element **x** belonging to each cluster C_k.

$$P(x|C_k) = \frac{1}{(2\pi)^{d/2}|\sum_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)} \quad (1)$$

where $P(x|C_k)$ are mixture components, d indicates dimension, t transpose and μ_k and Σ_k are the mean and covariance matrices of cluster C_k. The "membership weights" are calculated as follows:

$$w_{ik} = P(z_{ik} = 1|C_k) = \frac{P_k(x_i|z_k, C_k)\alpha_k}{\Sigma_{m=1}^{K} P_m(x_i|z_m, C_m)\alpha_m} \quad (2)$$

$P(z_{ij} = 1|C_j)$ is a vector of K binary indicator variables that are mutually exclusive and exhaustive (i.e. one and only one of the z_k's is equal to 1, and the others are 0). z is a random variable representing the identity of the mixture component that generated x. The α_k are the mixture weights, representing the probability that a randomly selected x was generated by component k, where $\Sigma_{k=1}^{K}\alpha_k = 1$).

M-Step:

The M-step estimates the parameters of the probability distribution of each class for the next step with $\alpha_k = \frac{N_k}{N}$ where N_k is the number of elements assigned to component k and N the total number of elements. The class K means μ_k and the covariance Σ_k are calculated as

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N} w_{ik}x_i \quad \text{and} \quad \Sigma_k = \frac{1}{N_k}\sum_{i=1}^{N} w_{ik}(x_i - \mu_k)(x_i - \mu_k)^t \quad (3)$$

1.2. Cluster geometry

Clustering can also be classified based on the covariance structure considered (Erar, 2011). The full covariance matrix of each cluster is $\Sigma_k = \lambda_k D_k A_k D_k^T$ where λ_k are the eigenvalues that specify the cluster size, the eigenvectors D_k indicates the orientation and D_k^T its transpose, and A_k is the cluster shape. With spherical clusters, all clusters have an equal shape with a diagonal covariance matrix (no correlation between variables), where the cluster size may be equal for all clusters ($\Sigma_k = \lambda I$) or different ($\Sigma_k = \lambda_k I$). Diagonal clusters present different variances per variable, with fixed cluster size and shape ($\Sigma_k = \lambda B$), varying shape but fixed cluster size ($\Sigma_k = \lambda B_k$) or both varying ($\Sigma_k = \lambda_k B_k$). In the latter case, the clusters are elliptical but parallel to the axes. With $\Sigma_k = \lambda D A D^T$ the clusters are elliptical, but the same covariance structure applies to all clusters. General models with full covariance do not constrain the covariance matrix to being diagonal, allowing correlation between variables ($\Sigma_k = \lambda_k D_k A_k D_k^T$). Fig. 1 shows three different cluster geometries and the corresponding covariance matrix. The left case presents a spherical cluster with variance equal to one for both variables (diagonal values). The middle figure includes a non-spherical cluster (difference in variance for the x and y axis) but no correlation (parallel to one of the axes). The right plot presents a non-spherical cluster with correlation between variables (non-diagonal terms of the covariance matrix are non-zero). Changing the correlation sign would draw a cluster orientated on the opposite direction (top left to bottom right).

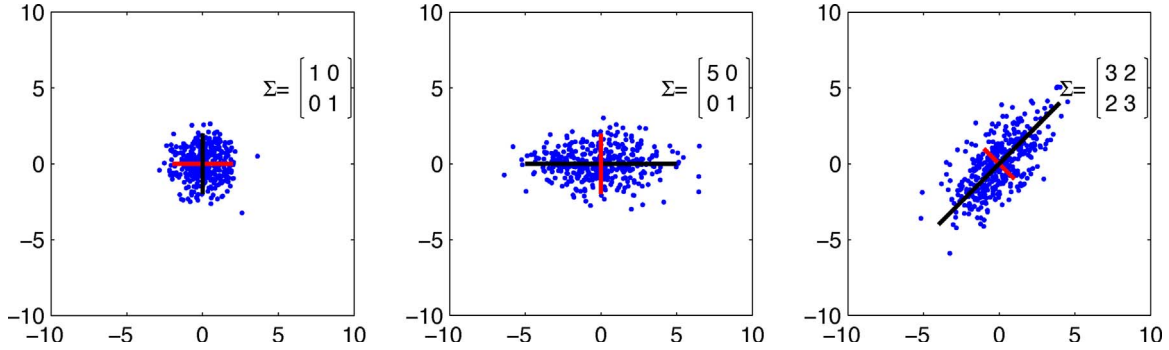

Fig. 1. Representation of three cluster geometries. The left case presents a spherical cluster with variance equal to one for both variables (diagonal values). The middle figure includes a non-spherical cluster (difference in variance for the *x* and *y* axis) but no correlation (parallel to one of the axes). The right plot presents a non-spherical cluster with correlation between variables (non-diagonal terms of the covariance matrix are non-zero). Changing the correlation sign would draw a cluster orientated on the opposite direction (top left to bottom right).

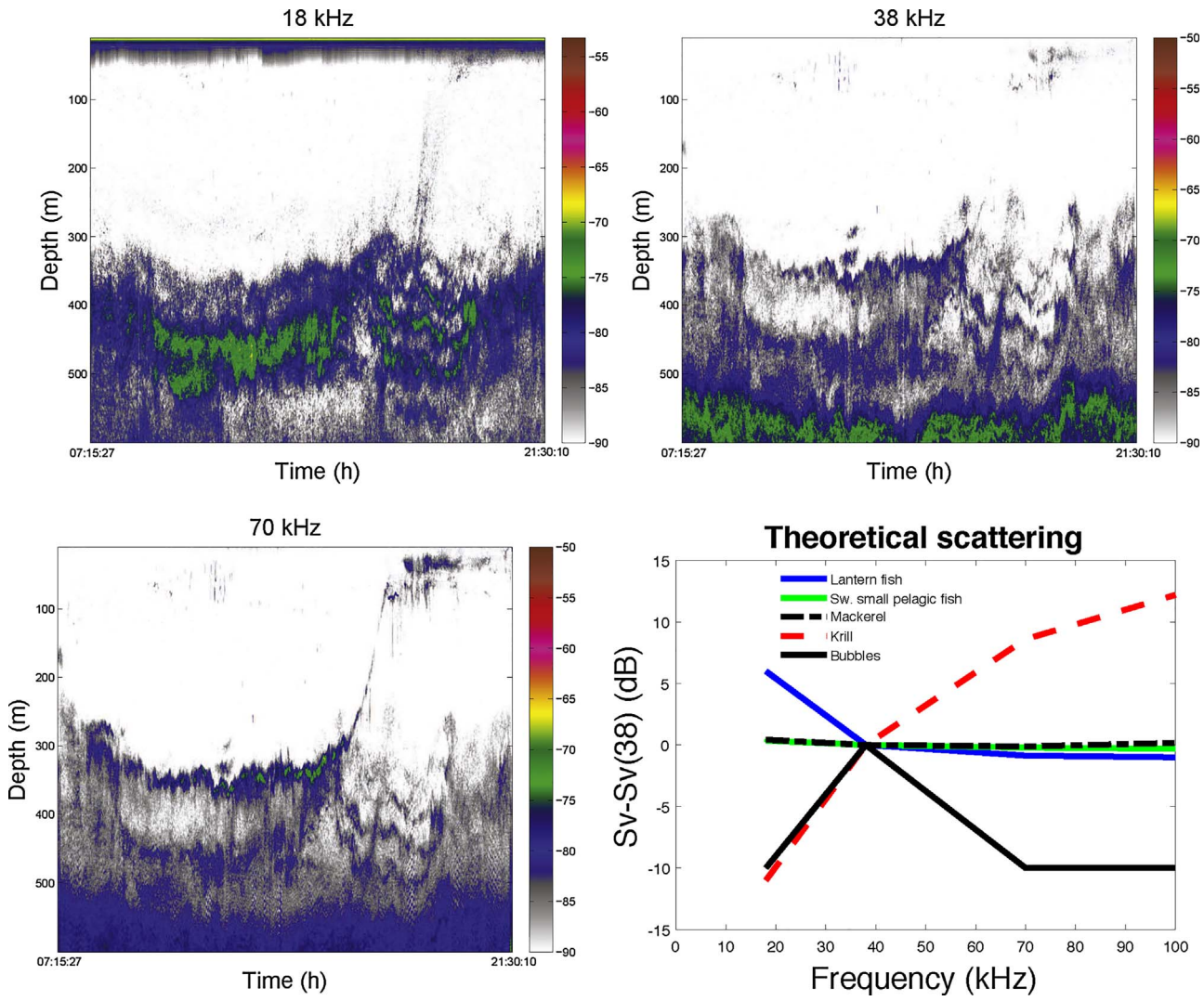

Fig. 2. Original echograms from the SCAPA survey at three frequencies after denoising with the algorithm in Peña (2016). Colors correspond to scattering values (S_v, dB). Note that the ringing noise layer at the surface of the 18 kHz echogram is not used in clustering as it is not seen in the other echograms. Bottom right: Theoretical frequency responses (*FR*) used as centroids. (For interpretation of the references to color in the figure legend, the reader is referred to the web version of this article.)

1.3. Acoustic data clustering background

This paper is focused on the application of robust partitional (non-hierarchical) clustering techniques to fisheries and plankton multi-frequency acoustic data at the pixel level for identification of acoustic groups exclusively based on their *FR*. Previous works in this area include Anderson et al. (2007), Woillez et al. (2012) and Ross et al. (2013). Ross et al. (2013) applied KM to broadband data (71 frequencies) comparing the use of absolute S_v, *FR* and RGB. *FR* data were calculated as the subtraction of the maximum S_v at each observation. Although they named this pre-processing as normalization, note that it was not applied to columns but to rows, and was thus a decentering technique that removes the numerical density term. The RGB data were created by shrinking the number of variables (71 frequencies) into a three-dimensional color-based space that represents the general tendency of the spectrum. They used random initialization and the elbow method (< 10% variation of the variance curve as the number of clusters estimation). Anderson et al. (2007) employed the EM clustering algorithm with acoustic data, but using S_v values as variables, and initializing centroids with the clusters found by a KM pre-processing. They employed a version of the Bayesian Information Criterion (BIC) that considers the sum of the probability of all points belonging to their allocated cluster to estimate the number of clusters. Woillez et al. (2012) combined unsupervised and supervised learning by joining training of labeled data with clustering of unlabelled data; *FR* data were employed. The unsupervised portion used EM clustering initialized with KM (with no mention of how KM was initialized). The BIC method was used to estimate the number of clusters.

A similar application with monofrequency acoustic data, using S_v at different depths as variables, was presented in Behagle et al. (2016) and Boersch-Supan et al. (2017). Behagle et al. (2016) estimated the number of clusters with the Calinski criterion, which considers the within- and between-group dispersion. Boersch-Supan et al. (2017) clustered vertical profiles of mesopelagic acoustic data using K-medoids (equivalent to KM with the L1 distance). They employed the silhouette technique to estimate the number of clusters. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation), calculated with any distance metric, such as the L2 distance. Thus, only means are considered.

Similar works have been applied to a mixture of external variables (temperature, salinity, etc.) and aggregated acoustic data, usually at the school level (Campanella and Taylor, 2016; Cox et al., 2010; Fablet et al., 2009), but also to data averaged in 'nodes' defining larger aggregations (Buelens et al., 2009) or even to fish acoustic tracks (Rakowitz et al., 2012). Clustering employing the 'kernel trick' was used in Buelens et al. (2009). Although the kernel trick allows the non-

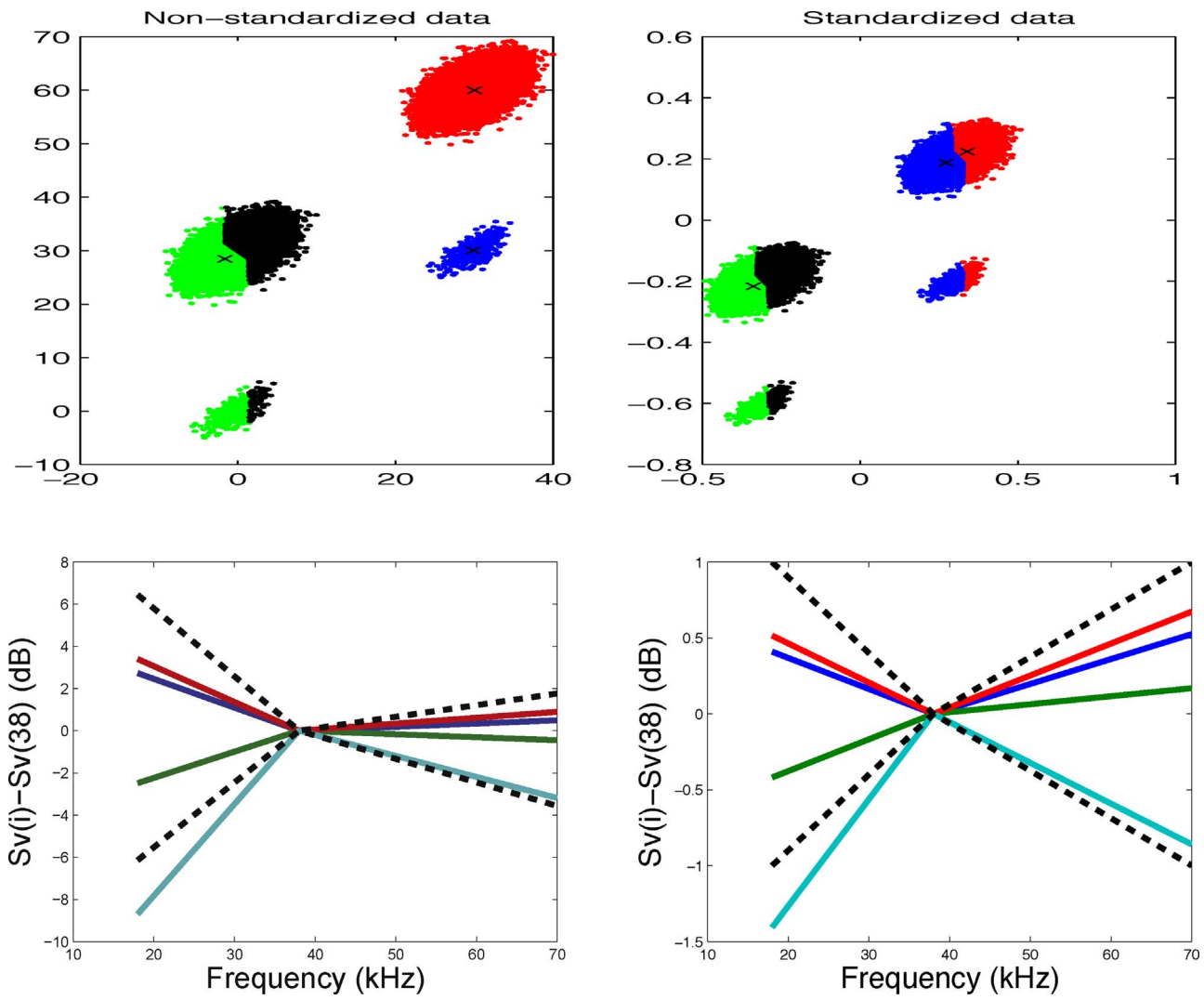

Fig. 3. Standardization effects. Upper figures: KM clustering of simulated clusters with different sizes (number of points) before and after standardization. Points belonging to the dame cluster present the same color. Location of centroids are marked with an *x*. Bottom figures: frequency response before and after standardization. Four different tendencies are distinguished with colors. Dotted lines mark standard deviations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

linearly projection of data into a subspace where clusters can be linearly separated, it requires the calculation of distances between all pairs of points, which makes them impractical for large datasets, such as acoustic data at a pixel level. In this publication, the echogram was pre-clustered into nodes according to different smoothing and averaging techniques, which is not comparable with the current study. Hierarchical clustering is not considered in this review, although it has been applied to a mixture of environmental and aggregated acoustic data (Bertrand et al., 1999; Domokos, 2009; Doray et al., 2009).

2. Material and methods

2.1. Example data

Simulated data including four clusters were firstly employed to separately determine the influence of the different geometrical parameters of clusters in KM. Only the most informative examples are shown. Then, multi-frequency acoustic data recorded during the SCAPA surveys, four seasonally distributed (February, April, July and November 2015) research surveys carried out to study the structure and carbon pathways of the planktonic foodweb, were employed with KM and EM clustering. Data from SCAPA surveys were collected off Cape Peñas (bathymetry from 30 to 4600 m), in the Central Cantabrian Sea. The *FR* spectrum variation with standardization of four selected pixels is firstly presented. A particularly challenging scenario of acoustic data with different cluster sizes recorded during the November SCAPA survey down to 2000 m depth was considered. Acoustic backscatter was recorded with a sphere calibrated (Demer et al., 2015) Simrad EK60 echosounder, using the split-beam transducers to record at multiple frequencies simultaneously (pulse duration of 1 ms). In order to reach 600 m depth, the three lowest frequencies (18, 38 and 70 kHz) were used within the clustering. A beam width of 7° was common for all frequencies except 18 kHz, with an 11° beam. The difference in sampled volume or the slight shift in transducer position was not considered in this work as the layers scrutinized were continuous and homogeneous, particularly after the denoising process (see below). All data were processed in Matlab. Only pixels having volume backscatter values (S_{vi}, dB re m^{-1}) above −90 dB at all three frequencies were used in the clustering processes. The denoising technique in Peña (2016) was applied to remove background noise. This method smooths data based on local variance: the higher the variance the lower the smoothing. This ensures that only samples within the same school or layer are averaged, while edges with higher local variance are kept unchanged. This technique also increments data quality (signal-to-noise ratio) and takes into account transmission and absorption losses. Previously, impulse noise has to be removed. White noise data were accounted for modelling an

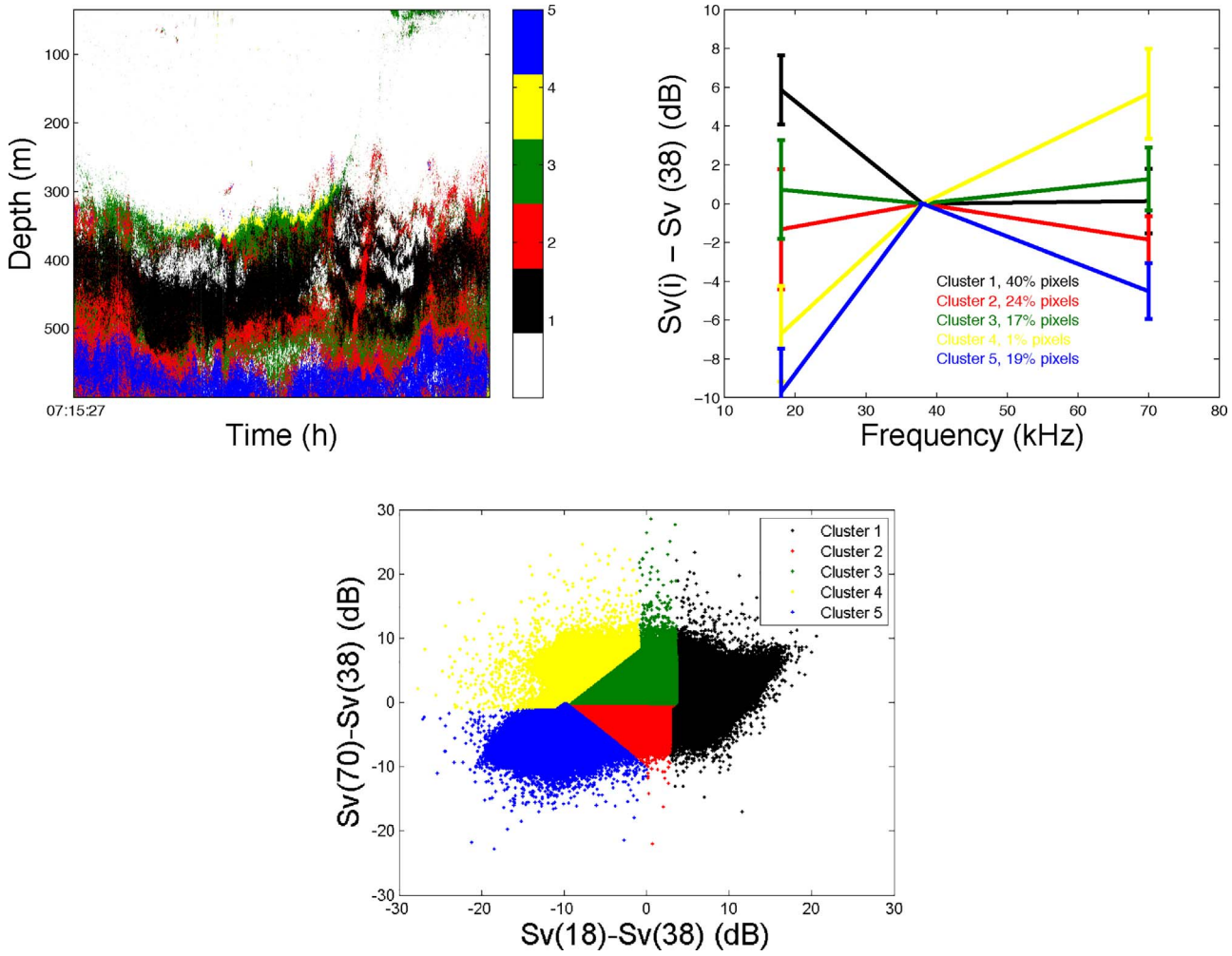

Fig. 4. Initial clustering using the theoretical centroids in Fig. 2. Upper left: location of clusters in the echogram. White pixels indicate unconsidered pixels below the S_v threshold. Upper right: clusters' *FR* with standard deviations as bars. The legend includes the cluster size in percentage. Bottom: Scatterplot of S_v differences. Clusters two and three have similar *FR* and are joined before using these clusters for initialization.

extra cluster for noise data (Banerjee and Davé, 2012).

2.2. Clustering options

The data supplied to the algorithms have echogram pixels as observations and *FR* (S_{vi}-S_{v38}, $i = 18, 70$) values at different frequencies as variables. The use of two variables allows us to show the 2D scatterplots of the resulting clusters. The selection of a similarity measure in time-series (or sequential data) is of crucial importance in clustering. The L1 (or cityblock) distance is used in this work, due to its robustness against noise, for the center-based algorithms (such as KM). The maximum number of iterations is set to 500, but the algorithm is previously stopped if the total error (sum of distances between data points and their cluster centroid) stops decreasing. A minimum of five iterations is imposed. The fast Matlab implementation in Chen (2016) was employed for the EM clustering. The final log-likelihood is provided as a summary of how likely are points to belong to the cluster they are in.

2.3. Clustering initialization

A good initialization is essential to properly locate the centroids near the global minimum. A new methodology is presented here based on theoretical scattering models. Knowledge of the plausible species in the area is most relevant for this purpose. The theoretical *FR* of the general acoustic groups in the area are calculated and employed as initial centroids. The employed data include epipelagic and mesopelagic data echograms and thus, the expected spectrum according to theoretical models (Peña et al., 2014; Peña and Calise,2016) are (see bottom right plot in Fig. 2): (i) mesopelagic fish with swim bladders such as myctophids or *Maurolicus muelleri* with a similar descending spectrum from 18 kHz. The lantern fish may present S_v ranges of 5–7 dB (Fujino et al., 2009). (ii) Krill found in Spanish waters have a 3 cm mode in length and thus, present in general an increase of *FR* with frequency. (iii) Layers of unidentified organisms that scatter most strongly at 38 kHz are always present along the water column in the Bay of Biscay, which could be due to larvae or phytoplankton gas inclusions. Additionally, the deep scattering layer (DSL) starting around 600 m depth produced by *Cyclothon spp* resonance (Peña et al., 2014) has a similar *FR*. (iv) The epipelagic zone (0–200 m depth) may include small pelagic fish with swim bladders, with a similar response at the working frequencies. This *FR* can also be due to the afore mentioned white noise, but contrary to small pelagic fish, there is no school or shoal indicating aggregating behavior. In the mesopelagic realm (200–1000 m depth), this *FR* is due to background noise. (v) Mackerel, which presents ascending S_v plots in the absence of a swim bladder but an initial decline in S_v values from 18 to 38 kHz (Korneliussen, 2010). Centroid *FR* spectrum (with cluster size information) and location of the different groups in the echogram are shown for each technique. Scatterplots of the two variables colored by cluster are presented to show the cluster shape.

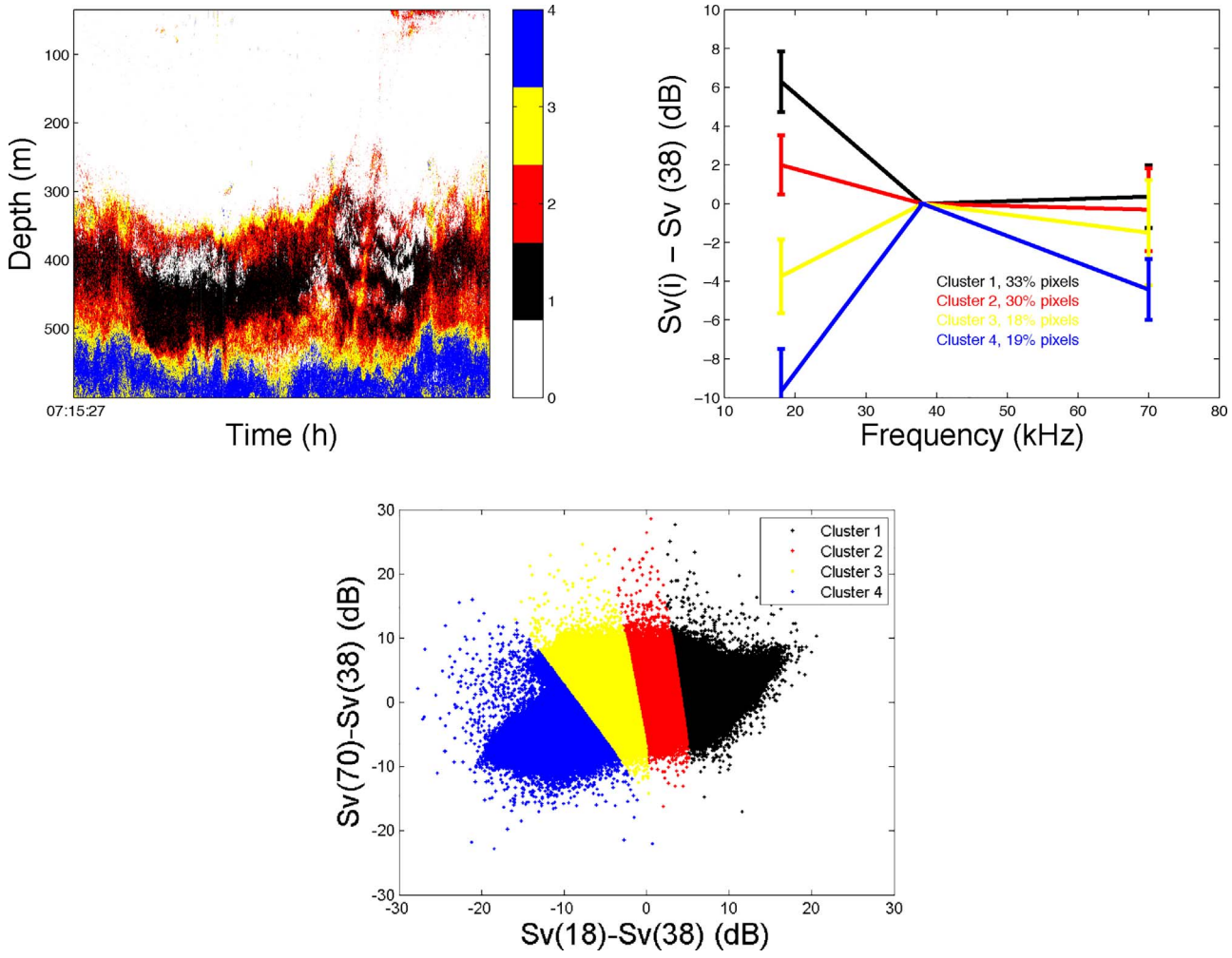

Fig. 5. KM clustering using the results in Fig. 4 as an initialization. Upper left: location of clusters in the echogram. White pixels indicate unconsidered pixels below the S_v threshold. Upper right: clusters' *FR* with standard deviations as bars. The legend includes the cluster size in percentage. Bottom: Scatterplot of S_v differences.

3. Results

The influence of cluster geometry on clustering was analyzed with simulated data and K-Means. The results show that the most influential aspect is clearly cluster size, which greatly conditions the random initialization process. Cluster variance and orientation are also relevant. The upper plots in Fig. 3 show the resulting clustering of four clusters of different cluster size, with and without standardization. The random initialization always locates more than one centroid in one of the more abundant clusters, resulting in poor clustering. Standardization makes the global variances equal but cluster geometry remains non-spherical. The lower plots in Fig. 3 show the influence of standardization in *FR* of four randomly selected echogram pixels. Standardization makes variance at 18 and 70 kHz equal (dotted lines in the right figure) but this also changes the relative difference between frequencies. Converting S_v to *FR* (necessary to remove numerical density) varies the total variance per variable as the variance of a subtraction is the sum of the original variances minus two times the covariance between them. This means that the resulting variances will change depending on the selected frequency of reference. Standardization also modifies covariances, as it is a linear combination $a + b * x$ for each frequency (where b is minus the mean divided by the standard deviation at that frequency) and the resulting covariance will be the original covariance multiplied by b of each frequency. If the variance range increases, values above the mean will move upwards and vice versa. In this case, values at 70 kHz have a greater shift upwards and downwards than 18, altering the spectrum trend. The S_v values present a negative covariance between 18 and 38 kHz of 1.72, whereas the covariance is positive between 70 and 38 kHz (11.97). This produces a lower variance in *FR* at 70 kHz (remember that $Var(A - B) = Var(A) + Var(B) - 2 * Cov(AB)$), and higher modifications in the resulting tendency after standardization at this frequency. These values will change with the species present in the echogram and thus *FR* without standardization (to maintain the frequency spectrum) and an algorithm that can deal with non-spherical clusters should be used.

The original denoised echograms at the three available frequencies of the real acoustic dataset are included in Fig. 2. The visualizing threshold (-90 to −50 dB) allows three basic layers to be seen scattering more at 18 kHz (400–500 m depth), 38 kHz (500–600 m depth) and 70 kHz (300–350 m depth). We can visually note the difference in the number of pixels in each layer, with the latter being more scarce. The bottom right figure depicts the theoretical *FR* used in the initialization process. Fig. 4 shows the initial pixel grouping according to the theoretical centroids by calculating the lowest L1 distance from each pixel to those theoretical spectra. When the proportion of clusters is very uneven, the probability of finding less abundant groups, such as krill in

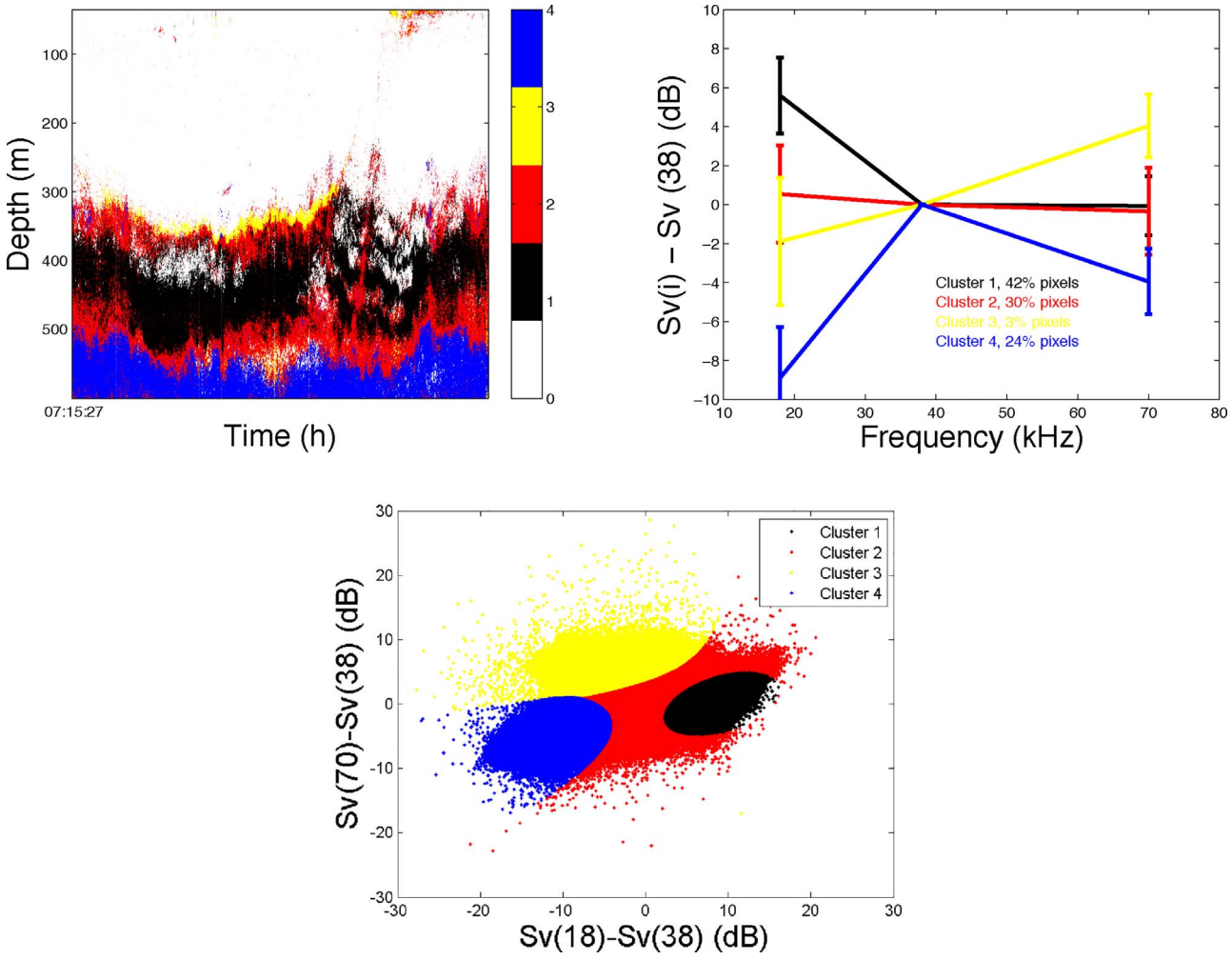

Fig. 6. EM clustering using the results in Fig. 4 as an initialization. Upper left: location of clusters in the echogram. White pixels indicate unconsidered pixels below the S_v threshold. Upper right: clusters' *FR* with standard deviations as bars. The legend includes the cluster size in percentage. Bottom: scatterplot of S_v differences. Log-likelihood = −5.1517.

Table 1
General recomendations and tips for clustering fisheries acoustic data.

Recomendations and tips
Use clustering techniques modelling non-spherical clusters such as EM clustering.
* EM clustering works well with ficheries acoustic data.
* K-Means and its derivatives are not suitable for fisheries acoustic data.

Preprocessing
Use Sv differences as variables.
Do not standardize the Sv differences.

Initialization
If you know the species, use the most accurate modelling.
If you do not know the species, use as many general FR as possible.
Always add a noise cluster (or two) with a flat FR.
Clustering will always favour the most abundant groups.
Adding groups increments the chance to find less abundant species but increments time.

Results
Uneven cluster percentages is to be expected in most cases.
Well-defined continuos layers for each group are to be expected.
Scatterplots should show well defined elipsoidal clusters.
Clusters aligned along one dimension in the scatterplot indicates bad clustering.
A variance reduction in the FR plots from initialization is to be expected but not compulsory.
The noise cluster should cover most of the scatterplot and overlap the other clusters.
If an extra FR is used, two clusters will show the same or complementary distributions.
FR slopes away from theory may indicate mixing.
Increasing the number of iterations may help to reduce mixing.

this example, is very low. Employing theoretical centroids in this way allows the inclusion of all possible tendencies and identifies less abundant groups. The krill layer is clearly identified in cluster four with only 1% of the pixels. The lantern fish are included in cluster one, with scatter descending with frequency. The layer with higher scatter at 38 kHz in cluster five is the deep scatter layer (DSL). Cluster two and three clearly present the same *FR* although split by positive and negative values of S_{v70}-S_{v38} (the histogram, not shown, depicts their complementary distribution). These two clusters are thus joined together for the initialization process resulting in four clusters. Therefore, this initial process also allows the selection of the number of clusters for the clustering process. All clustering results are shown assuming four clusters in the data and those initial centroids and pixels pre-clustering.

Employing the results in Fig. 4 as initial centroids in KM (Fig. 5) shows that, even with better initialization, KM clustering is not optimal when the cluster sizes are very different. This technique is influenced by variance differences between attributes as is noticeable in the scatterplot (Fig. 5): the S_{v18}-S_{v38} has more variance than the vertical variable S_{v70}-S_{v38} and the resulting clusters are split along the former. The acoustic spectra do not include any group with higher responses at 70 kHz (krill type) as it is mixed with noise within cluster 3. Clusters one and two present similar *FR* (lantern fish type) with different slopes. Cluster four includes the deepest layer with a higher response at 38 kHz. Forty-three percent of the echogram pixels have values above the lower threshold in all frequencies that are mostly equally split into four clusters. The similar pixel percentages show that KM tries to

equalize the cluster sizes as much as possible and, in order to do that, splits the most abundant groups into several groups. Cluster variances are also similar, located at different means on the S_{v18}-S_{v38} axis. Using KM and random initialization (results not shown), we obtain similar results with spherical clusters split based on the S_{v18}-S_{v38} variable, although cluster sizes are less uneven (13, 10, 9 and 13%) indicating more mixed and equally sized clusters.

Finally the EM clustering results with theoretical initialization (Fig. 6) present the four acoustic groups (lantern fish, krill, noise and DSL) with no mixing. Cluster sizes are uneven (42, 30, 3 and 24%) and present different distributions with full covariance. The orientation of the clusters, non-parallel to the axis, indicates correlation between variables. We can clearly see that missing data are taken into account with this model, as each cluster represents an incomplete ellipsoid. White noise surrounding all groups is identified and split into cluster two. The krill layer found presents a lower range in *FR* than the theoretical S_v difference values estimated with the model for *Meganyctiphanes norvegica* in Peña and Calise (2016). According to this model, the krill S_v range is reduced when the sizes are larger, as the flattening of the spectrum is reached at a lower frequency; orientations other than horizontal would also reduce the spectrum range (Calise and Skaret., 2011). Further results using the EM clustering with different combinations of FR at the initialization step are included in Annex.

4. Discussion

This work evaluates the clustering performances of two algorithms (KM and EM clustering) with acoustic data, focusing on pre-processing, cluster geometry and initialization. The use of S_v as variables produces clusters that depend on numerical density, as in Anderson et al. (2007) with clusters named 'low S_v' and 'high S_v'. The use of absolute S_v data in Ross et al. (2013) was equivalent to clusters found using only one frequency, proving the clustering based on S_v intensity. Table 7 in Campanella and Taylor (2016) presented species mixing in several clusters with some species being part of several groups. This study did not consider the *TS* variation with depth of mesopelagic fishes, as the example does not include migration time and they are all occupying similar depths. This could, however, be important in other cases. Behagle et al. (2016) and Boersch-Supan et al. (2017) employed mono-frequency acoustic data and the comparison was made between pixels located at the same depth, removing this problem, but still using S_v values as proxies of numerical density, when in fact an increase in S_v values in a particular zone could be due to larger or stronger scatterers (different species, lengths, etc.), or to an increase in numerical density. The use of *FR* removes numerical density but implies that *FR* (38) is equal to 1 for all observations. This means that this variable has 0 range and thus, no weight on center-based clustering techniques. In fact, one way to change the weight to a variable is to change its scale: "expressing a variable in smaller units leads to a larger range for that variable, which will then have a large effect on the resulting structure" (Kaufman and Rousseeuw, 2005). However, as the EM clustering also considers correlation between variables, results are slightly different when using the *FR* (38) information. Including *FR* (38) with the SCAPA data in the EM clustering entailed a 2% of pixels change: 3% of the pixels categorized as myctophids using two variables moved to the noise group when employing three variables and 2% of the pixels categorized as noise with two variables moved to the krill and +38 clusters adding the *FR* (38) column. It is only a slight modification of the pixels considered as noise in this case, but could be more relevant in other scenarios and should be considered.

We found that standardization should be approached with caution in clustering as it discards information that could be relevant, such as the acoustic spectrum (correlation between variables). The same would hold if including depth, latitude and longitude information; by standardizing these data, the spatial information matching a particular depth with a geographical area would be lost. This could also mean transforming spherical clusters into ellipsoids, for example in the case of rounded clusters aligned along a variable (such as depth, with vertically separated species). The use of S_v profiles as in Behagle et al. (2016) and Boersch-Supan et al. (2017) is prone to the same geometrical issues, with differences in variance among variables (some intermediate depths present low scatter day and night while others, like the surface layer, vary greatly) and correlation between variables (S_v at different depths) modifying the cluster shape. This correlation could even be more complex due to the acoustic beam conical shape, that increments the overlap of sampled volumes with depth. The correlation between variables should thus present an increasing intensity with depth. A -90 dB threshold is used in this work. This is an important parameter, as it changes the cluster size, which is particularly relevant for less abundant groups. In general as many pixels should be maintained as possible, and the higher frequency usually limits the number of pixels, due to higher acoustic attenuation and absorption with distance from the echosounder. Ross et al. (2013) noted several times that KM requires clusters of equal size, which may result in the wrong clusters. Probabilistic methods without these requirements are mentioned but not used due to 'underlying assumptions that we consider potentially inappropriate for our data'. No further explanation of those assumptions is given.

One of the main findings of this work was the application of a solid centroid initialization that directs the algorithms to the correct minimum and helps in the estimation of the number of clusters. Even after properly considering cluster geometry and removal of numerical density, initialization is key to separating sound clusters. Clustering *FR* data in Ross et al. (2013) was not successful due to the use of random initialization and the KM technique. Initializing EM clustering with KM as in Woillez et al. (2012) is considered to improve the results, but still depends on KM finding all the acoustic groups. Their resulting mixed clusters reflect a bad initialization in the number of clusters or cluster centroid. Campanella and Taylor (2016) and Boersch-Supan et al. (2017) employ the silhouette technique to estimate the number of clusters with variables of different variance. Particularly in the former study, which included external variables, this technique was prone to poor results as it is based on distances from the observed points to the cluster centers, overlooking cluster overlap and ellipsoidal shapes. The same holds for other techniques, such as the Calinski measure employed in Behagle et al. (2016). The theoretical initialization technique presented in this study is a much faster option than running KM to initialize centroids or running the algorithm several times with different numbers of clusters to find the most optimal. Knowledge of the local species and their theoretical *FR* is necessary to create the initial centroids and to interpret the results. It can also be easily extended to include external variables, for example in cases of two species presenting the same *FR* but different mean salinity or temperature values. However, these values must be confidently known for all the groups. The use of external variables implies larger variance differences between variables. For instance, for the 600 m upper waters presented in this study, depth, temperature, salinity and dissolved oxygen present variances of 29,950, 4.54, $5 * 10^{-4}$, and 0.09, respectively. The need to employ clustering techniques considering variances when mixing these variables with acoustic data is evident. The general recomendations and tips for clustering fisheries acoustic data derived from this study are summarized in Table 1.

5. Conclusions

Selecting the correct clustering algorithm for your data based on requirements is key for an accurate result. Non-spherical clusters and correlation between variables in multi-frequency acoustic data need to be considered for clustering. A good initialization is also essential to locate the centroids near the global minima for all techniques; a new methodology based on theoretical scatter models is presented. Differences of S_v (FR) are necessary to remove dependence from numerical density and standardization should no be employed in order to keep relevant correlation information. External variables can be added as long as their mean value is well known for all the considered groups.

Acknowledgments

We thank the collaboration of all scientists and crew involved in the SCAPA project (CTM2013-45089-R).

Annex

This section includes EM clustering results for the same data, using a different combination of theoretical FR for initialization. When using four initial groups including two noise (flat) FR, one for lantern fish, one for +38 kHz without a krill group (Fig. A.1), the resulting clustering is quite similar to Fig. 6, as one of the noise groups has evolved into a krill group. Although the log-likelihood is slightly lower in this case, its FR is not as steep due to some remaining noise mixing, also visible in the scatterplot, where some yellow dots are separated from the rest in the lower right corner. Using four groups including two noise groups, one lantern fish and one krill gives similar results (results not shown), with one of the noise groups evolving into the missing +38 kHz group. When only using three initial groups with one noise FR, one krill and one +38 kHz group without considering lantern fish (Fig. A.2), the resulting clustering finds three clusters but switching the krill FR into a lantern fish tendency. This is due to the much superior abundance of fish in this echogram. These results indicate a very solid clustering as longs as all the present FR are considered at initialization. The use of several noise groups may help finding unknown groups. If not enough groups are used, the most abundant species will show up.

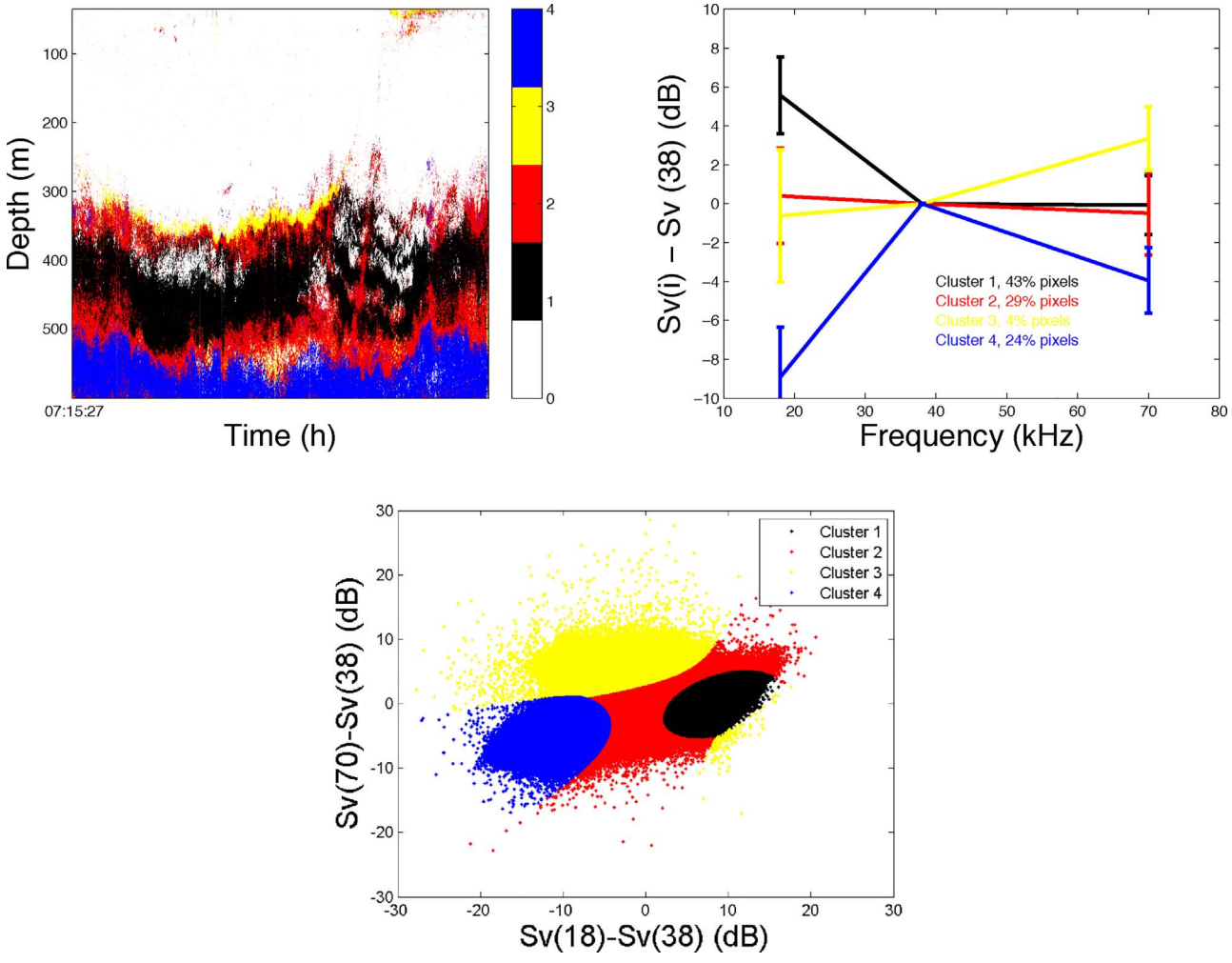

Fig. A.1. EM clustering using two noise groups, one +38 kHz and one lantern fish group as an initialization. Upper left: location of clusters in the echogram. White pixels indicate unconsidered pixels below the S_v threshold. Upper right: clusters' FR with standard deviations as bars. The legend includes the cluster size in percentage. Bottom: scatterplot of S_v differences. Log-likelihood = −5.1516. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

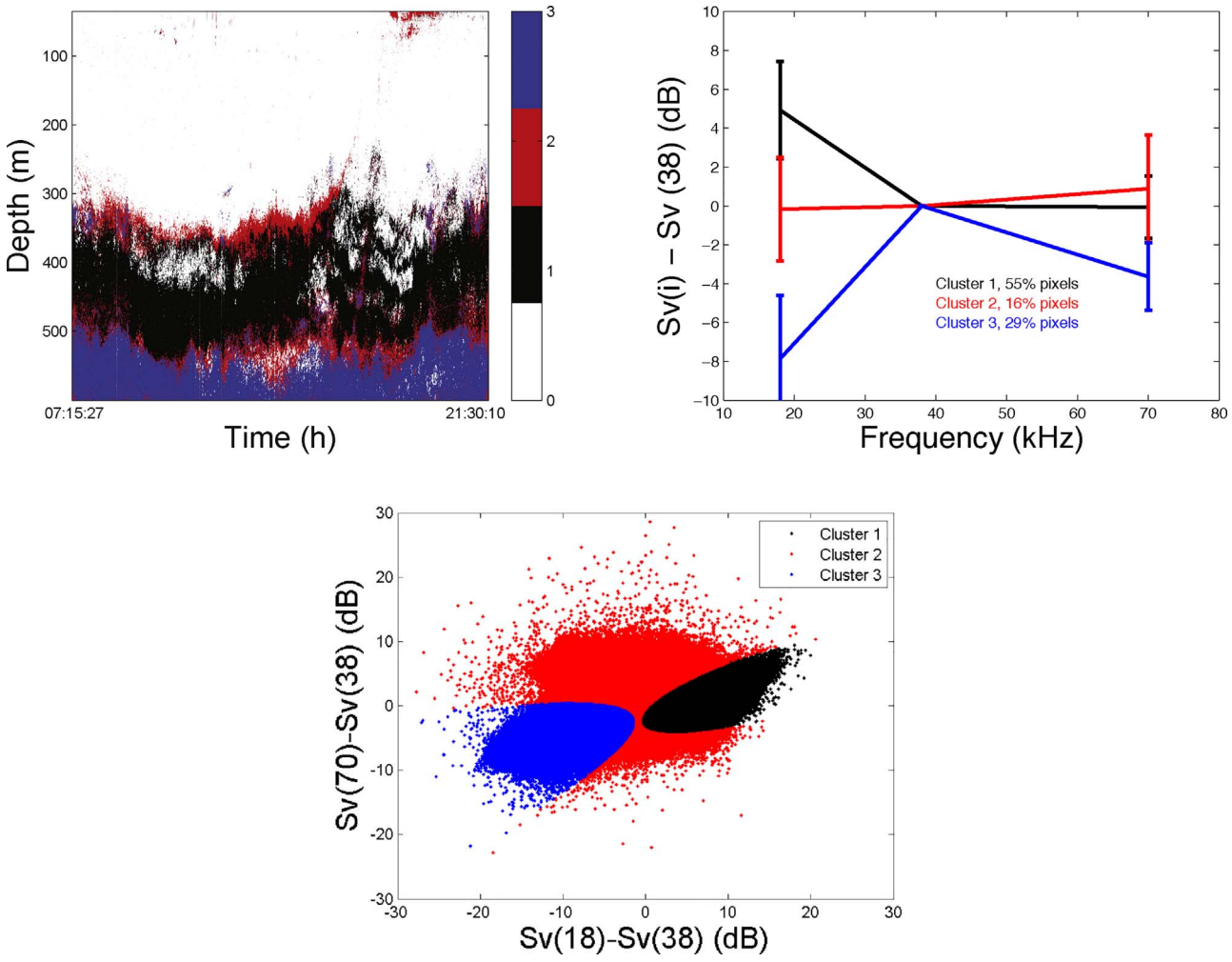

Fig. A.2. EM clustering using one noise group, one +38 kHz and one krill group as an initialization. Upper left: location of clusters in the echogram. White pixels indicate unconsidered pixels below the S_v threshold. Upper right: clusters' *FR* with standard deviations as bars. The legend includes the cluster size in percentage. Bottom: scatterplot of S_v differences. Log-likelihood = −5.1667.

References

Anderson, C.I.H., Horne, J.K., Boyle, J., 2007. Classifying multi-frequency fisheries acoustic data using a robust probabilistic classification technique. J. Acoust. Soc. Am. 121 (6), EL230–EL237.

Arlia, D., Coppola, M., 2001. 2001 experiments in parallel clustering with DBSCAN. In: Euro-Par 2001: Parallel Processing: 7th International Euro-Par Conference Manchester. UK August 28–31, Proceedings. Springer Berlin.

Banerjee, A., Davé, R.N., 2012. Robust clustering. WIREs Data Min. Knowl. Discov. 2, 29–59.

Behagle, N., Cotte, C., Ryan, T.E., Gauthier, O., Roudaut, G., Brehmer, P., Josse, E., Cherel, Y., 2016. Acoustic micronektonic distribution is structured by macroscale oceanographic processes across 20–50° S latitudes in the South-Western Indian Ocean. Deep Sea Res. Part I: Oceanogr. Res. Pap. 110, 20–32.

Bertrand, A., Le Borgne, R., Josse, E., 1999. Acoustic characterisation of micronekton distribution in French Polynesia. Mar. Ecol. Prog. Ser. 191, 127–140.

Buelens, B., Pauly, T., Williams, R., Sale, A., 2009. Kernel methods for the detection and classification of fish schools in single-beam and multibeam acoustic data. ICES J. Mar. Sci. 66, 1130–1135.

Boersch-Supan, P.H., Rogers, A.R., Brierley, A.S., 2017. The distribution of pelagic sound scattering layers across the southwest Indian Ocean. Deep Sea Res. Part II: Top. Stud. Oceanogr. 136, 108–121.

Calise, L., Skaret, G., 2011. Sensitivity investigation of the SDWBA Antarctic krill target strength model to fatness, material contrast and orientation. CCAMLR Sci. 18, 97–122.

Campanella, F., Taylor, J.C., 2016. Investigating acoustic diversity of fish aggregations in coral reef ecosystems from multi-frequency fishery sonar surveys. Fish. Res. 181, 63–76.

Chen, M., 2016. Matlab File Exchange. (accessed 30.09.16). https://es:mathworks:com/matlabcentral/fileexchange/26184-em-algorithm-for-gaussian-mixture-model-em-gmm-.

Cox, M.J., Warren, J.D., Demer, D.A., Cutter, G.R., Brierley, A.S., 2010. Three-dimensional observations of swarms of Antarctic krill (Euphausia superba) made using a multi-beam echosounder. Deep Sea Res. Part II: Top. Stud. Oceanogr. 57 (7), 508–518.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. (2), 224–227.

Demer, D.A., Berger, L., Bernasconi, M., Bethke, E., Boswell, K., Chu, D., Domokos, R., et al., 2015. Calibration of Acoustic Instruments. Tech. Rep., ICES Coop. Res. Rep. No. 326.

Domokos, R., 2009. Environmental effects on forage and longline fishery performance for albacore (Thunnus alalunga) in the American Samoa Exclusive Economic Zone. Fish. Oceanogr. 18, 419–438.

Doray, M., Petitgas, P., Nelson, L., Mahévas, S., Josse, E., Reynal, L., 2009. The influence of the environment on the variability of monthly tuna biomass around a moored, fish-aggregating device. ICES J. Mar. Sci. 66, 1410–1416.

Erar, B., 2011. Mixture Model Cluster Analysis Under Different Covariance Structures Using Information Complexity. University of Tennessee, Knoxvill Master Thesis.

Fablet, R., Lefort, R., Karoui, I., Berger, L., Massé, J., Scalabrin, C., Boucher, J.M., 2009. Classifying fish schools and estimating their species proportions in fishery-acoustic surveys. ICES J. Mar. Sci. 66 (6), 1136–1142.

Fujino, T., Sadayasu, K., Abe, K., Kikodoro, H., Tian, Y., Yasuma, H., Miyashita, K., 2009. Swimbladder morphology and target strength of a mesopelagic fish, Maurolicus japonicus. J. Mar. Acoust. Soc. Jpn. 36 (4), 241–249.

Hartigan, J.A., 1975. Clustering Algorithms, 99th Edition. John Wiley & Sons, Inc., New York, NY, USA.

Jain, A.K., Topchy, A., Law, M.H., Buhmann, J.M., 2004. Landscape of clustering algorithms. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, vol. 1. IEEE. pp. 260–263.

Kaufman, L., Rousseeuw, P.J., 2005. Finding Groups in Data: An Introduction to Cluster Analysis, vol. 344 John Wiley and Sons.

Korneliussen, R.J., 2010. The acoustic identification of Atlantic mackerel. ICES J. Mar. Sci.: J. Cons. 67 (8), 1749–1758.

Krzanowski, W.J., Lai, Y., 1988. A criterion for determining the number of groups in a data set using sum-of-squares clustering. Biometrics 23–34.
Lloyd, S.P., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28 (2), 129–137.
Krishnan, T., McLachlan, G., 1997. The EM Algorithm and Extensions, vol. 1 (997. Wiley, pp. 58–60.
MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 281–297.
Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50 (2), 159–179.
Peña, M., Olivar, M., Balbin, R., López-Jurado, J.L., Iglesias, M., Miquel, J., 2014. Acoustic detection of mesopelagic fishes in scattering layers of the Balearic sea (western Mediterranean). Can. J. Fish. Aquat. Sci. 71 (8), 1186–1197.
Peña, M., 2016. Incrementing the data quality of multi-frequency echograms using the Adaptive Wiener Filter (AWF) denoising algorithm. Deep Sea Res. Part I: Oceanogr. Res. Pap. 116, 14–21.
Peña, M., Calise, L., 2016. Use of SDWBA predictions for acoustic volume backscattering and the Self-Organizing Map to discern frequencies identifying Meganyctiphanes norvegica from mesopelagic fish species. Deep Sea Res. Part I: Oceanogr. Res. Pap. 110, 50–64.
Pereira, C.M.M., de Mello, R.F., 2013. Common dissimilarity measures are inappropriate for time series clustering. Rev. Inform. Teórica Apl. 20 (1), 25–48.
Rakowitz, G., Tuser, M., Riha, M., Juza, T., Balk, H., Kubecka, J., 2012. Use of high-frequency imaging sonar (DIDSON) to observe fish behaviour towards a surface trawl. Fish. Res. 123, 37–48.
Ross, T., Keister, J.E., Lara-Lopez, A., 2013. On the use of high-frequency broadband sonar to classify biological scattering layers from a cabled observatory in Saanich Inlet, British Columbia. Methods Oceanogr. 5, 19–38.
Ryan, T.E., Downie, R.A., Kloser, R.J., Keith, G., 2015. Reducing bias due to noise and attenuation in open-ocean echo integration data. ICES J. Mar. Sci. 72 (8), 2482–2493.
Santos, J.M., Embrechts, M., 2014. A family of two-dimensional benchmark data sets and its application to comparing different cluster validation indices. Pattern Recognition. Springer, pp. 41–50.
Simmonds, E.J., MacLennan, D.N., 2005. Fisheries Acoustics: Theory and Practice, 2nd ed. Blackwell Science, London, pp. 437.
Stanton, T.K., Chu, D., Wiebe, P.H., 1996. Acoustic scattering characteristics of several zooplankton groups. ICES J. Mar. Sci. 53, 289–295.
Woillez, M., Ressler, P.H., Wilson, C.D., Horne, J.K., 2012. Multifrequency species classification of acoustic-trawl survey data using semi-supervised learning with class discovery. J. Acoust. Soc. Am. 131 (2), EL184–EL190.
Wu, X., Kumar, V., Ross Quinlan, J., et al., 2008. Knowl. Inf. Syst. 14, 1–37.
Xiao, Y., Yu, J., 2012. Partitive clustering (K-means family). WIREs Data Min. Knowl. Discov. 2, 209–225.
Zalik, K.R., 2010. Cluster validity index for estimation of fuzzy clusters of different sizes and densities. Pattern Recognit. 43, 3374–3390.