Automatic classification of acoustically detected krill aggregations: a case study from Southern Ocean

Ignazio Fontana[a], Marco Barra[b,*], Angelo Bonanno[a], Giovanni Giacalone[a], Riccardo Rizzo[c], Olga Mangoni[d], Simona Genovese[a], Gualtiero Basilone[a], Rosalia Ferreri[a], Salvatore Mazzola[a], Giosué Lo Bosco[e,**] and Salvatore Aronica[a,**]

[a]*Institute of Anthropic Impacts and Sustainability in the marine environment - National Research Council (IAS-CNR), Campobello di Mazara, TP, Italy*
[b]*Institute of Marine Sciences - National Research Council (ISMAR-CNR), Napoli, Italy*
[c]*Institute of High Performance Computing And Networking - National Research Council (ICAR-CNR), Palermo, Italy*
[d]*Department of Biology, University of Naples Federico II, Napoli, Italy*
[e]*Department of Mathematics and Computer Science, University of Palermo, Palermo, Italy*

ARTICLE INFO

Keywords:
Hierarchical clustering
k-means
Krill
Ross Sea
Internal validation indices
Acoustic

ABSTRACT

Acoustic surveys represent the standard methodology to assess the spatial distribution and abundance of pelagic organisms characterized by aggregative behaviour. The species identification of acoustically observed aggregations is usually performed by taking into account the biological sampling and according to expert-based knowledge. The precision of survey estimates, such as total abundance and spatial distribution, strongly depends on the efficiency of acoustic and biological sampling as well as on the species identification. In this context, the automatic identification of specific groups based on energetic and morphological features could improve the species identification process, allowing to improve the precision of survey estimates or to overcome problems related to biases in biological sampling. In the present study, we test the use of well-known unsupervised clustering methods focusing on two important krill species namely *Euphausia superba* and *Euphausia crystallorophias*. In order to obtain a reference classification, the observed echoes were first classified according to specific criteria based on two parameters accounting for the acoustic response at 38 kHz and 120 kHz. Different clustering methods combined with three distance metrics were then tested working on a wider set of parameters, accounting for the depth of insonified aggregation as well as for energetic and morphological features. The clustering performances were then evaluated by comparing the reference classification to the one obtained by clustering. Obtained results showed that the k-means performs better than the considered hierarchical methods. Our findings also evidenced that working on a specific set of variables rather than on all available ones highly impact k-means performances.

1. Introduction

In the last decades, marine fishery science widely used acoustic-based techniques to obtain information about the spatial distribution and abundance of economically and ecologically important pelagic (i.e. inhabiting the water column) organisms ([2], [7], [8], [6], [25], [38], [52]). Several specific characteristics make acoustic methods an effective tool in monitoring pelagic populations. Some pelagic organisms are characterized by aggregative behaviour, thus forming large aggregations of individuals, called schools or swarms, that are easily identified by means of acoustic methods. The use of a scientific echo-sounder allows recording data on wide areas in a relatively small amount of time, leading to a synoptic and spatially detailed view of the status of aquatic resources. Furthermore, using different frequencies (typically 38 kHz, 120 kHz and 200 kHz) may allow researchers to identify the insonified species and/or classes of organisms (zooplankton, gelatinous etc.). Briefly, during acoustic surveys, an acoustic pulse is transmitted by a hull-mounted transducer at regular time intervals; when the acoustic wave, vertically travelling along the water column, encounter objects characterized by a different density, part of the energy is backscattered and recorded, thus allowing to map the encountered objects on the so-called echogram (Fig. 1).

In the context of acoustic surveys carried out to monitor pelagic organisms, the "objects" of interest are the aggregations of living organisms inhabiting the water column (Fig. 1). In most cases, it is not possible to uniquely identify species based only on acoustic information since different species with similar acoustic responses occur in the surveyed area. Thus, biological samples are needed to characterize the observed echoes and perform biological measurements. The biological sampling effort depends on several factors: the environmental heterogeneity of the study area, the number of species characterizing the considered ecosystem, the spatial

*Corresponding author
**Equal contribution
ignazio-fontana@cnr.it (I. Fontana); marco.barra@cnr.it (M. Barra); angelo.bonanno@cnr.it (A. Bonanno); giovanni.giacalone@cnr.it (G. Giacalone); riccardo.rizzo@cnr.it (R. Rizzo); olga.mangoni@unina.it (O. Mangoni); simona.genovese@cnr.it (S. Genovese); gualtiero.basilone@cnr.it (G. Basilone); rosalia.ferreri@cnr.it (R. Ferreri); salvatore.mazzola@cnr.it (S. Mazzola); giosue.lobosco@unipa.it (G. Lo Bosco); salvatore.aronica@cnr.it (S. Aronica)
ORCID(s): 0000-0002-0527-9827 (I. Fontana); 0000-0002-8621-504X (M. Barra); 0000-0002-3868-751X (A. Bonanno); 0000-0001-8172-4710 (G. Giacalone); 0000-0001-5007-6925 (R. Rizzo); 0000-0001-7789-0820 (O. Mangoni); 0000-0001-9570-178X (S. Genovese); 0000-0002-5732-5055 (G. Basilone); 0000-0001-7339-4942 (R. Ferreri); 0000-0002-4622-8129 (S. Mazzola); 0000-0002-1602-0693 (G. Lo Bosco); 0000-0003-3489-1473 (S. Aronica)

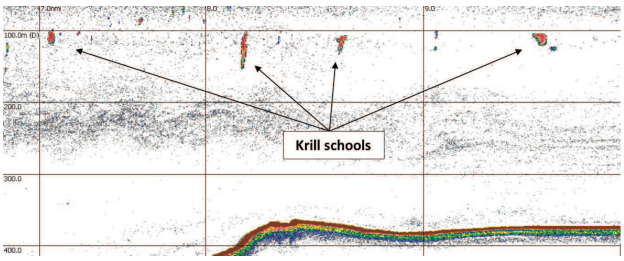

Figure 1: Example of echogram at 120 kHz. The coloured patches represents aggregations of krill organisms.

overlap among species as well as on the available vessel time and the operative scenarios that could make biological sampling particularly difficult. It is important to single out that the precision of survey estimates (e.g. total abundance and organisms' spatial distribution) strongly depends on the efficiency of both acoustic and biological sampling. Nonetheless, misclassification of echoes due to a bias in the species identification procedure may represent an important source of error. In this context, several authors investigated the possibility to perform species identification through semi-automatic classification procedures ([23], [21], [22], [12], [1]). In the work [22] Fallon et al. used a random forest approach to discriminate among mackerel icefish, krill and a mixed-species group in the Southern Ocean, obtaining an error rate of 5.45%, 3.73% and 8.06% respectively. The authors evidenced that the most important variables permitting to discriminate the three groups were the minimum S_v value at 120 kHz (Sv_{120}) and to a lesser extent the geographical position, the school depth and time of day. Working on small pelagic fish and in a different environment, D'Elia et al. (2014) [21] adopted a random forest algorithm to classify schools of small pelagic fish species in the central Mediterranean Sea by considering both energetic and morphological variables (i.e. parameters characterizing some aspect of school shape) as well as the depth of insonified schools. In particular, the study focused on three distinct pelagic fish species (anchovy, sardine and horse mackerel) and a fourth group considering different small pelagic fish species (OPS: Other Pelagic Species) that were less abundant in the study area. The authors reported a total successful classification rate of 76% and highlighted the poor discriminant power of morphological parameters and the importance of the school depth in the classification performances. Working on a similar dataset, and considering the same groups, Aronica et al (2019) [1] adopted an artificial neural network approach, obtaining a successful classification rate for the considered group of about 95%. In this case, the authors evidenced the importance of considering environmental variables (along with morphological and energetic ones) as they introduce in the classification procedure important information about the environmental preferences of considered species.

In this work, we tested the use of unsupervised clustering algorithms to identify the echoes recorded during a multi-purpose survey carried out in the Ross Sea (Southern Ocean) during 2016/2017 austral summer under the umbrella of the Italian National Antarctic Research Program (Project P-ROSE, "PNRA16 00239"). In particular, the analysis focused two important krill species inhabiting the Ross Sea, namely *Euphausia crystallorophias* and *Euphausia superba* ([2], [36], [37], [38]). The Ross Sea is characterized by the presence of different sub-systems following alternative pathways for primary production ([41], [5]) and for the transfer of energy toward upper trophic levels trought krill species. Krill species in the Southern Ocean ecosystem represent an important prey item for many species such as penguins and whales and play a key role in the energy transfer between the lower and upper trophic levels, also impacting the carbon sink [13]. Due to their importance, a number of studies focused on their spatial distribution and abundance by means of acoustic methods ([39], [9], [2], [37], [17]). In order to acoustically identify echoes belonging to different krill species, several authors analyzed the acoustic properties of krill species inhabiting the Southern Ocean at different frequencies (mainly 38 kHz and 120 kHz), focusing on the analysis of volume backscattering strength (Sv dB re $1m^{-1}$) values [40], [9], [36], [37], [38]. In this context, Madureira et al. (1993) [40] evaluated for the first time the possibility to discriminate among three different krill species (*Eupausia frigida*, *Euphausia superba* and *Themisto gaudichaudii*) by analyzing the response at different frequencies. In particular, by regressing Sv values at 38 kHz and 120 kHz (Sv_{38}, Sv_{120}), specific regression coefficients were proposed for each species. Brierley et al. (1998) [9], following the work in [40] and analyzing a larger dataset, proposed specific regression coefficients along with their coefficients intervals for six species namely *Eupausia frigida*, *Euphausia superba*, *Rhincalanus gigas*, *Thysanoessa macrura*, *Themisto gaudichaudii* and *Antarctomysis maxima*. Anyway, the authors argued that considering a higher number of species, none of them was uniquely identified based on Sv_{38} and Sv_{120} values, even if most of the *E. superba* echoes could be distinguished from all the others adopting specific threshold during data analysis. Recently, La et al. [37], investigated the acoustic properties of *E. crystallorophias*, providing the regression equation coefficients and confidence intervals for this species. From a practical point of view, it is was observed that for *E. superba* the differences between Sv_{120} and Sv_{38}, was in the range 2 - 12 dB, while for *E. crystallorophias* such difference ranged between 12 - 18 dB ([37]). It is important to single out that, Fallon et al (2016) [22], working on trawl-verified acoustic aggregations, observed that in the case of *E. superba* the above-mentioned rule about the difference between Sv_{120} and Sv_{38} was verified only for 61% of the observed aggregations. In the present work, due to the lack of proper biological sampling, the observed aggregations related to krill organisms were partitioned between *E. crystallorophias* and *E. superba* by means of the regression equations based on Sv_{120} and Sv_{38} values. All the aggregations falling outside the criteria of *E. crystallorophias* and *E. superba* were classified as "unknown" as it was not possible to clearly associate them to a particular species. Based on the obtained reference clas-

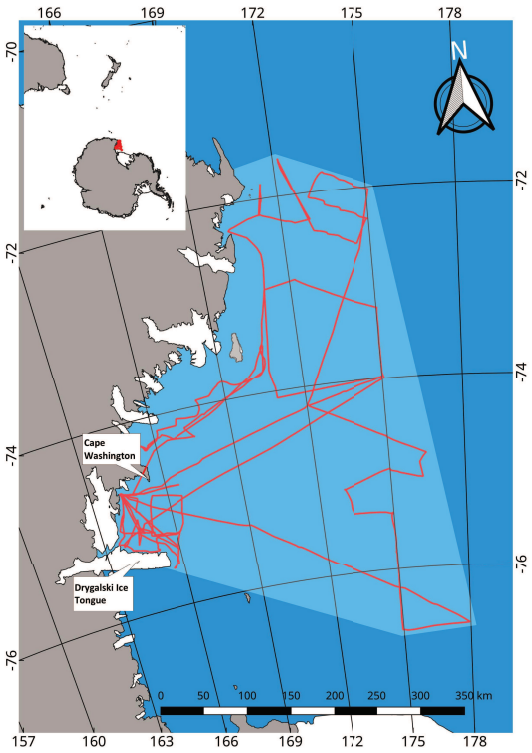

Figure 2: Study area and acoustic traks.

sification, we tested different unsupervised clustering methods and distance metrics working on a set of energetic and morphological parameters related to the insonified aggregations to evaluate the performance of the different unsupervised clustering methods in resembling the reference classification.

2. Material and Methods

2.1. *Acoustic data: acquisition and processing*

An acoustic survey was carried out in the period 05-11/01/2017 during the XXXII Antarctic expedition on-board the R/V "Italica" under the Italian National Antarctic Research Program and in the framework of the P-ROSE project (Plankton biodiversity and functioning of the Ross Sea ecosystems in a changing southern ocean). In particular, acoustic data were collected utilizing an EK60 scientific echosounder at three different frequencies (38 kHz, 120 kHz and 200 kHz) and calibrated following standard techniques [24]. The echo-sounder was configured to ping simultaneously at each frequency with a pulse duration of 1024ms. Acoustic sampling followed an opportunistic strategy (Fig. 2), recording data during all transfers among the sampling stations for a total of about 2200 nmi.

Acoustic row data were then processed using Echoview© software [32] in order to extract all the echoes related to aggregations of krill organisms. In the first step, the depth range for the analysis was defined. In particular, the region between 0 and 8.5 meters depth was excluded thus avoiding artefacts due to the beam formation distance and the noise due to cavitation and waves. Similarly, the echogram region related to a depth higher than 350 meters was excluded from the analysis due to the strong attenuation of the signals at 120 kHz and 200 kHz. In a second step, a 3x3 convolution filter was applied [22] and the background noise was removed by using the algorithm proposed by De Robertis and Higginbottom (2007) [18]. Other echogram regions affected by instrumental or environmental (i.e. waves and ice) noise were visually identified and removed manually. Finally, to extract only echoes related to krill organisms, a -80db threshold was applied on 120 kHz frequency according to Choi et. al (2018) [15]. Working on the 120kHz frequency (the reference frequency for krill species; [38]), all the aggregations were thus identified through the school detection module in Echoview. At the end of the processing, a total of 1334 aggregations were identified; for each aggregation, several parameters related to the energetic and geometric characteristics as well as the average depth of each aggregation were extracted (Table 1).

In addition, two more parameters were computed, namely: the frequency response at 120 and 200 kHz, respectively computed as

$$RF_{120} = 10 * log_{10}(\frac{10^{Sv.mean_{120}}}{10^{Sv.mean_{38}}}) \quad (1)$$

and

$$RF_{200} = 10 * log_{10}(\frac{10^{Sv.mean_{200}}}{10^{Sv.mean_{38}}}) \quad (2)$$

2.2. *Data preparation and exploratory data analysis*

During the survey, the bad weather conditions lead to a reduction of the time available for biological sampling that was limited to a low number of sampling stations. Under these conditions, it was not possible to properly characterize the observed acoustic aggregations based on the information obtained by the biological sampling. Consequently, the observed aggregations were classified according to regression equations developed on the energetic values at two different frequencies (38 kHz and 120 kHz) and available in the literature ([9], [37]). In particular *E. Superba* was identified on the basis of the following equation [9]:

$$Sv_{120} = 16.8 + 1.13 * Sv_{38} \quad (3)$$

taking into account for 95% confidence intervals of intercept and slope (13.6 - 20.54 and 1.08 - 1.18 respectively) . In a similar way, aggregations related to *E. crystallorophias* were identified by [37]:

$$Sv_{120} = 5.50 + 0.89 * Sv_{38} \quad (4)$$

Also in this case 95% confidence intervals of intercept and slope were considered (3.96 - 7.05 and 0.88 - 0.91 respectively). All the aggregations falling outside the classification criteria for *E. crystallorophias* and *E. Superba* were classified as *unknown* . Once the reference classification was

Table 1
Energetic and geometric parameters extracted for each aggregation identified by means of school detection module. The $*$ symbol indicate that the variable was extracted for each of the frequencies.

Parameter	Units	Description
Height.mean	m	Average school height
Depth.mean	m	Average school depth
Length	m	Length of the identified school
Thickness	m	Thickness of the identified school
Perimeter	m	Perimeter of the identified school
Area	m^2	Area of the identified school
Image.compactness		Ratio between perimeter and area
Sv.mean $*$	dB re $1m^{-1}$	Average recorded Sv value
Sv.max $*$	dB re $1m^{-1}$	Maximum recorded Sv value
Sv.min $*$	dB re $1m^{-1}$	Minimum recorded Sv value
SD $*$		Standard deviation of Sv values
Skew $*$	dB re $1m^{-1}$	Skewness of Sv values
H.rough $*$	dB re $1m^2/m^3$	Horizontal dispersion of acoustic energy within the school
V.rough $*$	dB re $1m^2/m^3$	Vertical dispersion of acoustic energy within the school
RF_{120}		Frequency ratio between 120 kHz and 38 kHz
RF_{200}		Frequency ratio between 200 kHz and 38 kHz

performed, an exploratory analysis was carried out. In a first step, the presence of significant differences among the identified groups was assessed for each variable using the Kruskal-Wallis Anova test followed by Mann-Withney posthoc test. Finally, since clustering methods could be negatively affected by extreme values, the frequency distribution and the skewness index of each variable was inspected in order to evaluate the possibility to apply specific transformations thus minimizing the influence of the tails. All the statistical analyses were carried out using the R statistical environment [46].

2.3. *Clustering methods*

The performances of different clustering methods and distances in resembling the reference classification were tested. In particular, the 5 hierarchical variants *Complete linkage, Single linkage, Average linkage, Median and Centroid* were tested using three different distance metrics: *Euclidean, Manhattan and Minkowski* (with p=3). The *K-means* with Euclidean and Manhattan distance was also tested (Minkowski distance was not available for k-means). In all the considered cases (i.e. the combination of clustering method and distances), the data were standardized to avoid scale problems. The classification based on Sv_{38} - Sv_{120} regression allowed to classify only observations related to *E. crystallorophias* and *E. superba*, while all the others were labelled as *unknown* (*unk* for short). Thus, as the "unk" group was potentially made by more than one species, internal validation indices (Table 2) were used to test for the hypothesis of more than 3 groups.

Internal validation indices are based on the concept of "good" cluster structure ([14], [31]). In particular, in the present study, 16 validation indices were used. The majority of them are based on the homogeneity and separation measures, i.e. the sum of distances between elements inside the same clusters and the sum of distances of elements belonging to different clusters. Others such as [29, 28] also on cluster internal variance. Some of them are only usable for the case of Hierarchical clustering [19]. All of them can suggest the correct number of clusters k by the *argmax* or *argmin* of the index, computed in the range $k = 2, ..., 10$.

All the statistical analysis were carried out using NbClust [14] package in the R statistical environment [46].

3. Results

3.1. *Reference classification and exploratory data analysis*

The regression-based (Eq. 3 and Eq. 4) classification clearly identified 3 distinct groups (Fig. 3); each observation was thus classified accordingly as *E. crystallorophias* (375 observations, identified as *C* for short), *E. superba* (703 observations, identified as *S* for short) and "unk" for all the remaining observations (256). Kruskal-Wallis ANOVA carried out for each considered variable evidenced the presence of significant differences among the three identified groups in 21 out of 31 cases (Table 3).

In particular, no significant differences were found in terms of $Sv.min_{120}$, $Length$, $Thickness$, $Area$, $H.rough_{120}$, SD_{120}, $V.rough_{120}$ and $Skew_{200}$. Furthermore, only for 11 variables, significant differences were found among all the three groups, while for the remaining ones significant differences were found only for one or two out of three comparisons. In particular, significant differences among the three groups were found for $Sv.mean_{38}$, $Sv.mean_{120}$, $Sv.min_{38}$, $Sv.max_{38}$, $H.rough_{38}$, $V.rough_{38}$, SD_{38}, $Skew_{38}$, RF_{120} and RF_{200} (Table 3, Fig. 4).

To avoid possible bias in the clustering due to highly skewed data, the skewness index was computed for each variable (Table 3). In many cases, the skewness index evidenced the presence of moderate/high right-skewed variables while only $Sv.min_{120}$ showed a moderate left-skewed distribution.

Table 2
Internal validation indices. The Optimal number of cluster is the argmax or argmin of the index for k ranging from 2 to 10.

Index	Ref. biblio.	Optimal number of clusters
Silhouette	[48]	argmax
Dunn	[20]	argmax
KL	[35]	argmax
Hartigan	[30]	argmin
DB	[16]	argmin
CH	[11]	argmax
Cindex	[34]	argmin
SDindex	[29]	argmin
Sdbw	[28]	argmin
Duda	[19]	argmin
Beale	[4]	argmax
Ratkowsky	[47]	argmax
Ball	[3]	argmax
Ptbiserial	[44, 43]	argmax
Gap	[50]	argmin
Mcclain	[42]	argmin

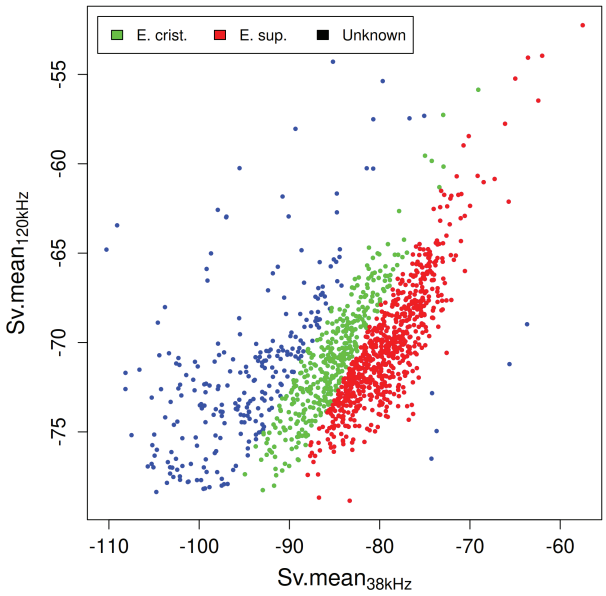

Figure 3: Scatterplot of the observations in the Sv_{38} and Sv_{120} space, presenting the obtained classification according to eq. 3 and 4. All observations outside the classification criteria for *E. crystallorophias* and *E. superba* were classified as "Unknown".

According to Hair et al. [27] and Bryne [10], variables characterized by an absolute value of skewness higher than 2 should not be considered normally distributed; consequently, all the variables characterized by an absolute value of the skewness index higher than 2 were log-transformed, while in the case of $Sv.min_{120}$, that was negatively skewed, the adopted transformation was:

$$Sv.min_{120\,t} = \log(\max(Sv.min_{120}+1) - Sv.min_{120}) \quad (5)$$

Once the variables were transformed, the absolute value of the skewness index was lower than 2 in all the cases (not shown).

3.2. *Clustering*

Different hierarchical clustering methods with 3 different distance measures (Euclidean, Manhattan and Minkowski) were tested. Along with hierarchical methods, also k-means was tested but working only on Euclidean and Manhattan distances (as Minkowski distance was not available for k-means). All the above-mentioned combinations of methods and distances were tested on the whole data matrix (considering all the extracted parameters) and on a reduced data matrix accounting only for the variables showing significant differences among all the three groups (reported in bold in Table 3). In a first step the degree of structure of dendrograms obtained by applying hierarchical methods was inspected. When applied to the whole data matrix, hierarchical methods provided a well-structured dendrogram only for Complete-Linkage (Fig. S 1). A similar situation was evidenced when hierarchical methods were applied on the selected variables (reported in bold in Table 3); in this case, only the dendrogram obtained using Average-Linkage and Complete-Linkage were well structured (Figs. S 2 and S 3). In all the other cases obtained dendrograms showed a poor structure (e.g. Fig. S 4). Consequently, the methods showing poorly structured dendrograms were excluded from further analysis. When applied to the whole dataset, validation indices highlighted that the optimal k (i.e. number of clusters) was 4 for the Complete-Linkage with Euclidean and Manhattan distance, while working with Manhattan and Minkowski distances, the optimal k was 3. The presence of only 3 groups was also evidenced in the case of the k-means algorithm with Euclidean and Manhattan distance. The same was true when working with selected variables using the Average-Link, Complete-Link and k-means whatever the distance used. Clustering was then carried out by taking into account, for each combination of method/distance,

Table 3
Skewness values and statistical tests results. Post-hoc tests were carried out only if K-W p-value was found lower than 0.01. Variables in bold present significant differences among the three considered groups. The symbol # indicate that the variable was found highly skewed and log-transformed.

Variable	Skewness	K-W p.value	C vs S	C vs unk	S vs unk
$\boldsymbol{Sv.mean_{38}}$	-0.71	p<0.01	p<0.01	p<0.01	p<0.01
$\boldsymbol{Sv.mean_{120}}$	0.947	p<0.01	p<0.01	p<0.01	p<0.01
$Sv.mean_{200}$	-0.138	p<0.01	p<0.01	0.012	p<0.01
$\boldsymbol{Sv.min_{38}}$	0.802	p<0.01	p<0.01	p<0.01	p<0.01
$Sv.min_{120}$ #	-2.757	p<0.01	0.021	0.079	p<0.01
$Sv.mim_{200}$	0.137	0.081			
$\boldsymbol{Sv.max_{38}}$	-0.488	p<0.01	p<0.01	p<0.01	p<0.01
$Sv.max_{120}$	0.753	p<0.01	0.056	0.006	p<0.01
$Sv.max_{200}$	-0.137	p<0.01	0.001	0.083	p<0.01
$Depth.mean$	0.795	p<0.01	p<0.01	0.013	p<0.01
$Length$ #	5.937	0.001			
$Thickness$ #	2.284	0.02			
$Perimeter$ #	6.304	p<0.01	0.248	0.094	0.004
$Area$ #	7.144	0.516			
$Image.compactness$ #	6.167	p<0.01	0.261	0.001	p<0.01
$Height$	1.864	p<0.01	0.523	p<0.01	p<0.01
$\boldsymbol{H.rough_{38}}$ #	24.317	p<0.01	p<0.01	p<0.01	p<0.01
$H.rough_{120}$ #	22.779	0.112			
$H.rough_{200}$ #	7.363	p<0.01	0.001	0.1	p<0.01
$\boldsymbol{V.rough_{38}}$ #	17.155	p<0.01	p<0.01	p<0.01	p<0.01
$V.rough_{120}$ #	16.103	0.176			
$V.rough_{200}$ #	6.56	p<0.01	0.001	0.09	p<0.01
$\boldsymbol{SD_{38}}$ #	21.968	p<0.01	p<0.01	p<0.01	p<0.01
SD_{120} #	8.788	0.142			
SD_{200} #	6.379	p<0.01	0.001	0.036	p<0.01
$\boldsymbol{Skew_{38}}$ #	3.423	p<0.01	0.001	p<0.01	p<0.01
$Skew_{120}$ #	2.36	p<0.01	0.009	0.885	0.018
$Skew_{200}$ #	2.126	0.064			
$\boldsymbol{RF_{120}}$	1.284	p<0.01	0.001	p<0.01	p<0.01
$\boldsymbol{RF_{200}}$	0.943	p<0.01	0.001	p<0.01	p<0.01

the k value obtained using validation indices. Clustering results were then validated by comparing the regression-based (Eq. 3 and Eq. 4) classification to the clustering one. In this context, it is important to highlight that clustering output was a vector of numeric values identifying the obtained clusters but does not provide the link between cluster number (i.e. 1, 2, 3) and the regression-based labelled groups ("C", "S", "unk"). To link the reference classification and cluster numbers, the pattern of $Sv.mean_{38}$ values (Fig. 4), showing good contrast among the predefined groups, was used. In particular, based on the reference classification, *E. superba* (S) was characterized by highest $Sv.mean_{38}$ median values, the unknown group ("unk") by the lowest, while *E. crystallorophias* (C) showed intermediate $Sv.mean_{38}$ median value. Thus, the obtained clusters were labelled according to such pattern, assigning the label “Sn” to the cluster showing the highest $Sv.mean_{38}$ median value, “unkn” to the cluster characterized by the lowest while the remaining cluster (*E. crystallorophias*) was identified as (Cn). Confusion matrix (Tabs. 4 and 5) were used to evaluate the performances of each selected method according to the number of clusters suggested by validation indices. The boxplots of $Sv.mean_{38}$ categorized by cluster (labelled according to the pattern evidenced in $Sv.mean_{38}$) were also generated to visually evaluate the obtained results (Figs. 5 and 6). When working on the whole dataset, the Complete Linkage method assigned most of the observations in a single cluster (Table 4). In particular, using the Euclidean distance, most of the observations belonging to C and S were allocated in a single cluster showing intermediate $Sv.mean_{38}$ median values. Even if the minimum and maximum $Sv.mean_{38}$ values of Cn and Sn resembled quite well the reference classification, the respective median values were lower than the reference ones (Fig. 5). Working with Manhattan distance, the obtained results showed similar problems and using the Minkowski distance, most of the observations were allocated to 2 clusters only (Table 4). On the contrary, when using the k-means method on the whole dataset, the final classification almost equally distributed the observations among the different groups (Table 4). Working on the selected variables only, hierarchical methods showed the same problems observed in the previous cases (Table 5 and Fig. 6). In the case

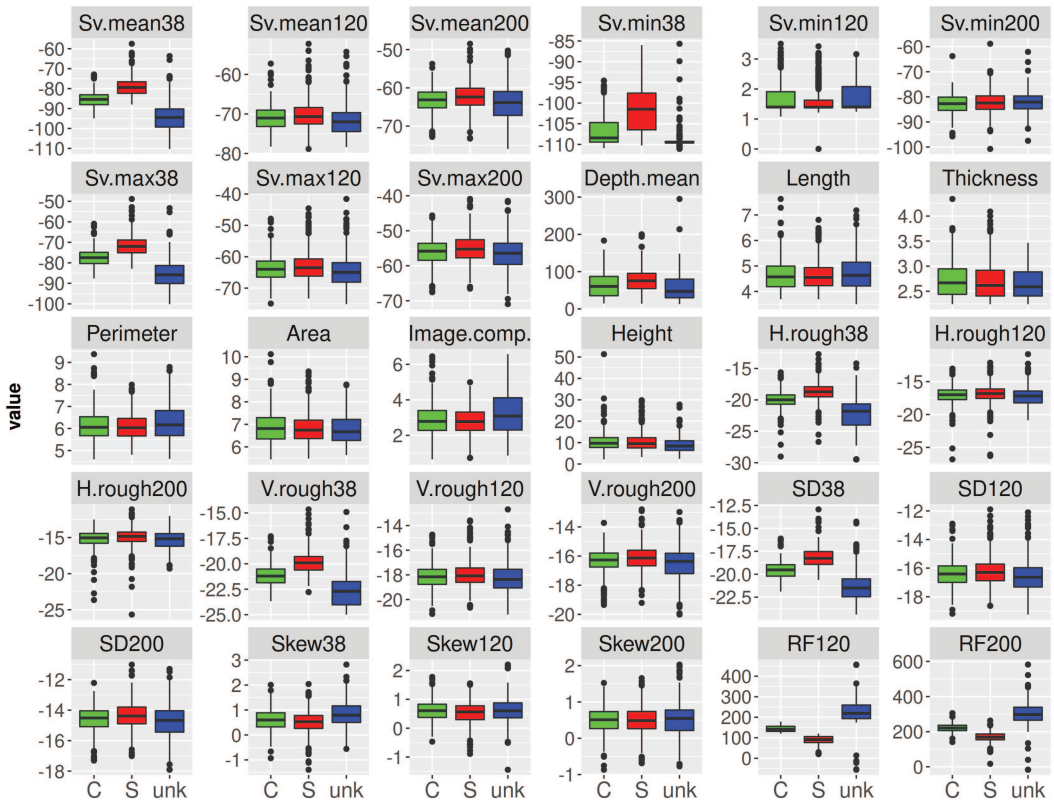

Figure 4: Boxplot by group of the considered variables according to the reference classification.

of Complete-Linkage, a high number of "unk" observations were correctly classified, while most observations related to C and S were assigned in the same cluster (Table 5). On the contrary, working with Average Linkage, the "unk" observations were largely misclassified. Finally, when working on the selected variables with the k-means algorithm, the best results were obtained. In particular, using the euclidean distance 85.1% of C, 67.6% of S and 65.6% of "unk" were correctly classified. Similar results were obtained by using the Manhattan distance. In such cases, the $Sv.mean_{38}$ pattern (Fig. 6) was well resembled for C and S, while in the case of "unk", even if the median value was quite similar to the original one, the maximum value was strongly lower than the correct one, evidencing the misclassification in both cases of about 30% of the unknown as C. It must also be noted that in both cases only a few of C and S were misclassified as "unkn".

4. Discussion

Identifying at species level the echoes recorded during acoustic surveys is one of the most important aspects in the estimation of the spatial distribution and abundance of pelagic organisms. Such a task represents a critical step for the interpretation of acoustic data [33] and is usually accomplished by using expert echogram scrutinizing (knowledge-based approach) and/or looking at the species composition in the nearest sampling station [49]. It is important to highlight that the species identification process could represent an important source of error [45], leading to biased estimates of abundance and spatial distribution of targeted organisms. In this context, the possibility to improve the species identification step through an automatic procedure is of great interest, especially if biological sampling was not representative due to technical problems. To the aim of finding specific rules to differentiate among different krill species, some authors analyzed the acoustic characteristics of krill working with different frequencies and focusing mainly on energetic values and frequency responce ([39], [9], [37], [38], [2]). In particular, Madureira et al. (1993) [39] and Brierly et al. (1998) [9], provided regression parameters, based on Sv_{38} and Sv_{120} values for different krill species. It is important to highlight that even if Madureira et al. (1993) [39] were able to discriminate three species based on Sv_{38} and Sv_{120} values, the results obtained by Brierley, working on additional species, showed that none of them was uniquely identified, even if most of the *E. superba* echoes could be distinguished from all the others adopting specific threshold during data analysis. As a rule of thumb the differences between Sv_{38} and Sv_{120} values should be in the range 2 - 12 dB for *E. superba* and 12 - 18 dB for *E. crystallorophias*. In this context, Fallon et al. (2016) [22], working on trawl-verified acoustic aggregations, observed that only 61% of the *E. superba* were characterized by a 2 - 12 dB difference between Sv_{120} and Sv_{38} values. In this work, we tested the use of several well-

Table 4
Confusion matrix related to clustering results obtained using the whole set of variables. C: *Euphausia crystallorophias*; S: *Euphausia superba*; unk: *unknown*. The suffix "n" in the column names indicate the new classification obtained by each combination of method/distance. In this context the clusters were labeled according to the $Sv.mean_{38}$ pattern observed in the original classification.

	Complete Linkage					k-means			
Euclidean		**Cn**	**Sn**	**unkn1**	**unkn2**		**Cn**	**Sn**	**unkn**
	C	83.2	10.1	5.6	1.1	C	53.9	30.4	15.7
	S	82.4	16.9	0.7	0	S	56.3	42.2	1.3
	unk	41	5.1	30.9	23	unk	18.4	19.1	62.5
Manhattan		**Cn**	**Sn**	**unkn1**	**unkn2**		**Cn**	**Sn**	**unkn**
	C	67.7	2.9	3.5	25.9	C	61.3	23.2	15.5
	S	82.1	4.8	0.1	12.9	S	54.1	43.2	2.7
	unk	21.5	3.5	40.2	34.8	unk	28.9	17.6	53.5
Minkowski		**Cn**	**Sn**	**unkn**					
	C	62.4	37.6	0					
	S	48.2	51.8	0					
	unk	74.2	19.1	6.6					

Table 5
Confusion matrix related to clustering results obtained working on selected variables only. The suffix "n" in column names indicate the new classification obtained by each combination of method/distance. In this context the clusters were labeled according to the $Sv.mean_{38}$ pattern observed in the original classification.

	Complete Linkage				Average Linkage				k-means			
Euclidean		**Cn**	**Sn**	**unkn**		**Cn**	**Sn**	**unkn**		**Cn**	**Sn**	**unkn**
	C	90.9	1.3	7.7	C	98.4	1.6	0	C	85.1	13.1	1.9
	S	97.2	2.8	0	S	97.3	2.7	0	S	32.4	67.6	0
	unk	34.8	2.7	62.5	unk	70.3	2.7	27	unk	30.1	4.3	65.6
Manhattan		**Cn**	**Sn**	**unkn**		**Cn**	**Sn**	**unkn**		**Cn**	**Sn**	**unkn**
	C	63.7	0.5	35.7	C	98.9	0	1.1	C	83.7	14.4	1.9
	S	95.3	3.4	1.3	S	99.1	0.9	0	S	31.4	68.6	0
	unk	16.8	0.8	82.4	unk	34	0.8	65.2	unk	30.9	4.3	64.8
Minkowski		**Cn**	**Sn**	**unkn**		**Cn**	**Sn**	**unkn**				
	C	93.9	1.3	4.8	C	100	0	0				
	S	86.8	13.2	0	S	100	0	0				
	unk	35.9	2.3	61.7	unk	77.3	1.2	21.5				

known and easy to implement clustering methods, comparing the obtained classification to the reference one accomplished by considering Sv_{120} and Sv_{38} values only.

Our results showed that among the considered methods and distances, the k-means with Manhattan distance was the one performing better. The use of internal validation indices, to test for the hypothesis of more than three groups, showed that was not possible to evidence the presence of sub-groups within the "unknown" group. In this context, it must be considered that k-means is sensitive to very unbalanced groups, that is the case of the "unknown" that accounted for about 19% of the total number of observations.

The variable selection was an important step to obtain an acceptable classification; all the considered methods were not able to perform a correct classification when variables characterized by a lack of contrast among the identified groups were used. The hierarchical methods, even in presence of a well-structured dendrogram, lead to misleading results probably due to the presence of outliers (even if proper transformations were adopted to reduce the effect of outliers and skewed distributions). On the contrary, the k-means algorithm was more robust to the presence of outliers. Obtained classification rate for *E. crystallorophias* was 85.1% while was lower for *E. superba* (68%) and the "unknown" group (66%). Considering that Fallon et al. (2016) [22] observed that *E. superba* backscatter could be underestimated, our reference classification is probably biased as it was developed looking at Sv_{38} and Sv_{120} variables only. Thus, taking into account that the k-means classification was based on a higher number of variables, it is possible, that some of the observations that we considered as misclassified effectively belong to the correct species.

According to D'Elia et al. (2014) [21], highlighting very poor performances using morphometric variables only, also in our case such variables were found not useful in the clustering procedure, showing no significant differences among

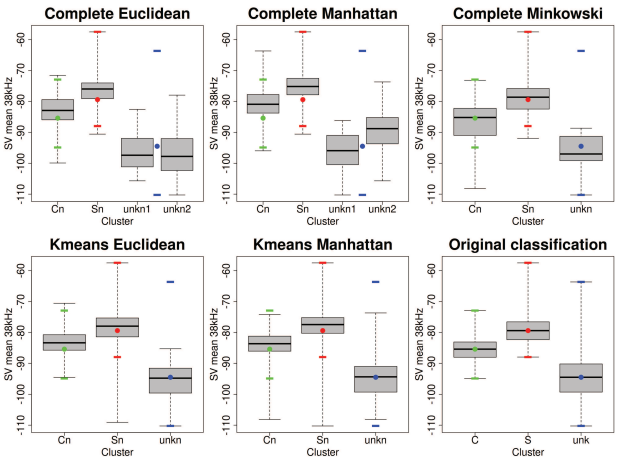

Figure 5: Boxplot of $Sv.mean_{38}$ classified according to the cluster analysis results obtained working on the whole set of variables. The $Sv.mean_{38}$ boxplot of the reference classification is also reported in the bottom-right panel. For complete Euclidean and Complete Manhattan, based on $Sv.mean_{38}$ boxplot of the reference classification (bottom-right panel), the third and fourth clusters were identified as two sub-group of "unk" and were labelled as "unkn1" and "unkn2". Coloured dots represent the median value observed for the reference classification, while coloured lines the minimum and maximum observed values.

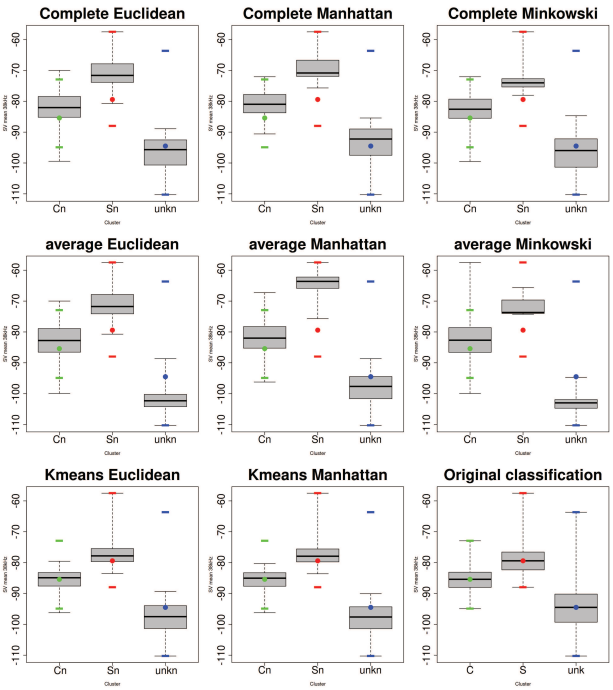

Figure 6: Boxplot of $Sv.mean_{38}$ classified according to the cluster analysis results obtained working on selected variables only. The boxplot of the reference classification is also reported in the lower-right panel. Coloured dots represent the median value observed for the reference classification, while coloured lines the minimum and maximum observed values.

the groups identified by the reference classification. It must be considered that a single-beam echo-sounder can record only a specific slice of the insonified school, and thus is not able to properly characterize the difference in the shape of the different organism's schools. Differently from our results, the above-mentioned authors reported a general decrease of classification rates when the average school depth was removed from the dataset (total successful rate lowered to 59%). Such a decrease in performances highlighted the importance of a variable accounting for specific behavioural and ecological aspects of small pelagic populations inhabiting the Mediterranean Sea.

In our case, the average depth of the aggregations was found not useful in classifying the three groups. In this context, it is important to single out that the acoustic data related to small pelagics in the Mediterranean Sea are recorded over the continental shelf only. The behaviour of small pelagics concerning depth is also modulated by species-specific ecological characteristics ([26], [51], [6]).

Similarly, krill species are characterized by species-specific depth preferences related to complex biological processes influenced by environmental factors. In our case during the survey very different environments, such as coastal ice-free areas, sectors closed to ice tongues as well as very offshore areas, were explored along with a great latitudinal range. The average depth of schools for the same species changes with changes in the environmental conditions; as an example, strong winds could lead to changes in the stratification of the upper part of the column water, influencing biological processes and consequently the depth of aggregations. Such variability, if not explicitly taken into account by considering environmental factors, could make the average depth of aggregations not informative in distinguishing the different species, since each species may react differently according to the considered environment. Aronica et al. (2019) [1], working on small pelagics in a different environment, along with energetic, morphometric and bathymetric variables, accounted also for a set of environmental descriptors. Temperature, salinity, fluorescence and oxygen were explicitly taken into account in the classification procedure. The authors obtained for the considered groups a classification rate of about 95% highlighting the good performance of the method and the importance of environmental variables for the classification. Thus, environmental variables, carrying important information about the ecological preference of considered species, are of great importance to improve classification results and probably, also in our case, considering environmental factors could lead to better classification performance.

5. Conclusions

Improving the procedures used to identify echoes represents a fundamental requirement to obtain reliable estimates of the spatial distribution and abundance of pelagic organisms. Several authors tried to tackle this problem by adopting different methods and working on different species and ecosystems. Our results, adopting the k-means algorithm, showed acceptable classification rates, that could be proba-

bly improved by explicitly considering, along with energetic parameters, the environmental factors. It is important to evidence that compared to other methods, the clustering algorithm considered in this work was found to be robust to outliers, is easier to implement and does not require specific assumptions. Thus k-means seems promising and could represent a valid tool to improve species identification procedures by reducing the post-processing time and obtaining more reliable estimates. Nonetheless, to fully validate this method, further studies are needed; in particular specific tests should be carried out including environmental variables and testing the k-means performance on trawl-verified aggregations.

Acknowledgements

The authors thanks the Italian National Antarctic Research Program (PNRA) and the P-ROSE project (Plankton biodiversity and functioning of the Ross Sea ecosystems in a changing southern ocean) that allowed collecting the acoustic dataset. The authors are also grateful to the captain and the crew of "Italica" Research Vessel.

References

[1] Aronica, S., Fontana, I., Giacalone, G., Lo Bosco, G., Rizzo, R., Mazzola, S., Basilone, G., Ferreri, R., Genovese, S., Barra, M., Bonanno, A., 2019. Identifying small pelagic mediterranean fish schools from acoustic and environmental data using optimized artificial neural networks. Ecological Informatics 50, 149–161.

[2] Azzali, M., Leonori, I., De Felice, A., A., R., 2006. Spatial-temporal relationships between two euphausiid species in the ross sea. Chem. Ecol. 22, 219–233.

[3] Ball, G., Hall, D., 1965. ISODATA: A novel method of data analysis and pattern classification. Technical Report. Stanford Research Institute.

[4] Beale, E., 1969. Euclidean Cluster Analysis. Scientific Control Systems Limited.

[5] Bolinesi, F., Saggiomo, M., Ardini, F., Castagno, P., Cordone, A., Fusco, G., Rivaro, P., Saggiomo, V., Mangoni, O., 2020. Spatial-related community structure and dynamics in phytoplankton of the ross sea, antarctica. Frontiers in Marine Science 7, 1092.

[6] Bonanno, A., Barra, M., Basilone, G., Genovese, S., Rumolo, P., Goncharov, S., Popov, S., Buongiorno Nardelli, B., Iudicone, D., Procaccini, G., Aronica, S., Patti, B., Giacalone, G., Ferreri, R., Fontana, I., Tranchida, G., Mangano, S., Pulizzi, M., Gargano, A., Di Maria, A., Mazzola, S., 2016. Environmental processes driving anchovy and sardine distribution in a highly variable environment: the role of the coastal structure and riverine input. Fisheries Oceanography 25, 471–490.

[7] Bonanno, A., Goncharov, S., Mazzola, S., Popov, S., Cuttitta, A., Patti, B., Basilone, G., Di Nieri, A., Patti, C., Aronica, S., Buscaino, S., 2006. Acoustic evaluation of anchovy larvae distribution in relation to oceanography in the cape passero area (strait of sicily). Chemistry and Ecology 22, S265–S273.

[8] Bonanno, A., Zgozi, S., Basilone, G., Hamza, M., Barra, M., Genovese, S., Rumolo, P., Nfate, A., Elsger, M., Goncharov, S., Popov, S., Mifsud, R., Bahri, T., Giacalone, G., Fontana, I., Buongiorno Nardelli, B., Aronica, S., Ceriola, L., Patti, B., Ferreri, R., Colella, S., G., V., Mazzola, S., 2015. Acoustically detected pelagic fish community in relation to environmental conditions observed in the central mediterranean sea: a comparison of libyan and sicilian–maltese coastal areas. Hydrobiologia 755, 209–224.

[9] Brierley, A., Ward, P., Watkins, J., Goss, C., 1998. Acoustic discrimination of southern ocean zooplankton. Deep-Sea Res. II 45, 1155–1173.

[10] Byrne, B.M., 2010. Structural equation modeling with amos: Basic concepts, applications, and programming. Routledge Taylor and Francis Group 2nd ed., 396 p.

[11] Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics 3, 1–27.

[12] Campanella, F., Christopher, T.J., 2016. Investigating acoustic diversity of fish aggregations in coral reef ecosystems from multifrequency fishery sonar surveys. Fisheries Research 181, 63–76.

[13] Cavan, E.L., Belcher, A., Atkinson, A., Hill, S.L., Kawaguchi, S., McCormack, S., Meyer, B., Nicol, S., Ratnarajah, L., Schmidt, K., Steinberg, D.K., Tarling, G.A., Boyd, P.W., 2019. The importance of antarctic krill in biogeochemical cycles. Nature Communications 10, 4742.

[14] Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., 2014. Nbclust: An r package for determining the relevant number of clusters in a data set. Journal of Statistical Software 61, 1–36.

[15] Choi, S., Yoon, E.A., An, D., Chung, S., Lee, J., Lee, K., 2018. Characterization of frequency and aggregation of the antarctic krill (euphausia superba) using acoustics. Ocean Sci. J. 53, 667–677.

[16] Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 1, 224–227.

[17] Davis, L.B., Hofmann, E.E., Klinck, J.M., Pinones, A., Dinniman, M.S., 2017. Distributions of krill and antarctic silverfish and correlations with environmental variables in the western ross sea, antarctica. Marine Ecology Progress Series 584, 45–65.

[18] De Robertis, A., Higginbottom, I., 2007. A post-processing technique to estimate the signal-to-noise ratio and remove echosounder background noise. ICES Journal of Marine Science 64, 1282–1291.

[19] Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. John Willey and Sons.

[20] Dunn, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics 4, 95–104.

[21] D'Elia, M., Patti, B., Bonanno, A., Fontana, I., Giacalone, G., Basilone, G.and Fernandes, F., 2014. Analysis of backscatter properties and application of classification procedures for the identification of small pelagic fish species in the central mediterranean. Fisheries research 149, 33–42.

[22] Fallon, N.G., Fielding, S., Fernandes, P.G., 2016. Classification of southern ocean krill and icefish echoes using random forests. ICES Journal of Marine Science 73, 1998–2008.

[23] Fernandes, P.G., 2009. Classification trees for species identification of fish–school echotraces. ICES Journal of Marine Science 66, 1073–1080.

[24] Foote, K.G., Knudsen, H.P., Vestnes, G., MacLennan, D.N., Simmonds, E.J., 1987. Calibration of acoustic instruments for fish density estimation: a practical guide. ICES Coop. Res. Rep. 144, 69.

[25] Giannoulaki, M., Iglesias, M., Tugores, M.P., Bonanno, A., Patti, B., De Felice, A., Leonori, I., Bigot, J.L., Tičina, V., Pyrounaki, M.M., Tsagarakis, K., Machias, A., Somarakis, S., Schismenou, E., Quinci, E., Basilone, G., Cuttitta, A., Campanella, F., Miquel, J., Oñate, D., Roos, D., Valavanis, V., 2013. Characterizing the potential habitat of european anchovy engraulis encrasicolus in the mediterranean sea, at different life stages. Fisheries Oceanography 22, 69–89.

[26] Giannoulaki, M., Valavanis, V.D., Palialexis, A., Tsagarakis, K., Machias, A., Somarakis, S., Papaconstantinou, C., 2008. Modelling the presence of anchovy engraulis encrasicolus in the aegean sea during early summer, based on satellite environmental data. Hydrobiologia 612, 225–240.

[27] Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., 2010. Multivariate Data Analysis: A Global Perspective. Pearson Education International, New Jersey.

[28] Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001. On clustering validation techniques. Journal of Intelligent Information Systems 17.

[29] Halkidi, M., Vazirgiannis, M., Batistakis, Y., 2000. Quality scheme assessment in the clustering process, in: Zighed, D.A., Komorowski, J., Żytkow, J. (Eds.), Principles of Data Mining and Knowledge Discovery, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 265–276.

[30] Hartigan, J.A., 1975. Clustering Algorithms. 99th ed., John Wiley

and Sons, Inc., USA.
[31] Hassani, M., Seidl, T., 2017. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. Vietnam J. Comput. Sci. 4, 171–183.
[32] Higginbottom, I., Pauly, T.J., Heatley, D.C., 2000. Virtual echograms for visualization and post-processing of multiple-frequency echosounder data. In Proceedings of the Fifth European Conference on Underwater Acoustics , 1497–1502.
[33] Horne, J., 2000. Acoustic approaches to remote species identification: a review. Fish. Oceanogr. 4, 356–371.
[34] Hubert, L., Levin, J., 1976. A general statistical framework for assessing categorical clustering in free recall. Psychological Bulletin 83, 1072–1080.
[35] Krzanowski, W., Lai, Y.T., 1988. A criterion for determining the number of groups in a data set using sum-of-squares clustering. Biometrics 44, 23.
[36] La, H., Lee, H., Kang, D., Lee, S., Shin, H., 2015a. Ex situ target strength of ice krill (euphausia crystallorophias). Chin. J. Oceanol. Limnol. 33, 802–808.
[37] La, H.S., Lee, H., Fielding, S., Kang, D., Ha, H.K., Atkinson, A., Park, J., Siegel, V., Lee, S., Shin, H.C., 2015b. High density of ice krill (euphausia crystallorophias) in the amundsen sea coastal polynya, antarctica. Deep Sea Research 95, 75–84.
[38] Leonori, I., De Felice, A., Canduci, G., Costantini, I., Biagiotti, I., Giuliani, G., Budillon, G., 2017. Krill distribution in relation to environmental parameters in mesoscale structures in the ross sea. Journal of Marine Systems 166, 159–171.
[39] Madureira, L., Everson, I., Murphy, E., 1993a. Interpretation of acoustic data at two frequencies to discriminate between antarctic krill (euphausia superba dana) and other scatterers. Journal of Plank-ton Research 15, 787–802.
[40] Madureira, L., Ward, P., Atkinson, A., 1993b. Differences in backscattering strength determined at 120 and 38 khz for three species of antarctic macroplankton. Marine Ecology Progress Series 93, 17–24.
[41] Mangoni, O., Saggiomo, V., Bolinesi, F., Margiotta, F., Budillon, G.and Cotroneo, Y., Misic, C., Rivaro, P., Saggiomo, M., 2017. Phytoplankton blooms during austral summer in the ross sea, antarctica: driving factors and trophic implications. PLoS ONE 4, 12.
[42] McClain, J.O., Rao, V.R., 1975. Clustisz: A program to test for the quality of clustering of a set of objects. Journal of Marketing Research 12, 456–460.
[43] Miligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179.
[44] Milligan, G., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45, 325–342.
[45] Petitgas, P., 2003. Biomass-dependent dynamics of fish spatial distributions characterized by geostatistical aggregation curves. ICES Journal of Marine Science 3, 443–453.
[46] R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.
[47] Ratkowsky, D., Lance, G., 1978. A criterion for determining the number of groups in a classification. Australian Computer Journal 10, 115–117.
[48] Rousseeuw, P., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65.
[49] Simmonds, J., N., M.D., 2008. Fisheries acoustics: Theory and practice. John Wiley and Sons .
[50] Tibshirani, R., Guenther, W., Hastie., T., 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society Series B .
[51] Tugores, M.P., Giannoulaki, M., Iglesias, M., Bonanno, A., Tičina, V., Leonori, I., Machias, A., Tsagarakis, K., Díaz, N., Giráldez, A., Patti, B., De Felice, A., Basilone, G., Valavanis, V., 2011. Habitat suitability modelling for sardine sardina pilchardus in a highly diverse ecosystem: the mediterranean sea. Mar. Ecol. Prog. Ser. 443, 181–205.
[52] Ventero, A., Iglesias, M., Córdoba, P., 2019. Krill spatial distribution in the spanish mediterranean sea in summer time. Journal of Plankton Research 4, 491–505.

A. Supplementary materials

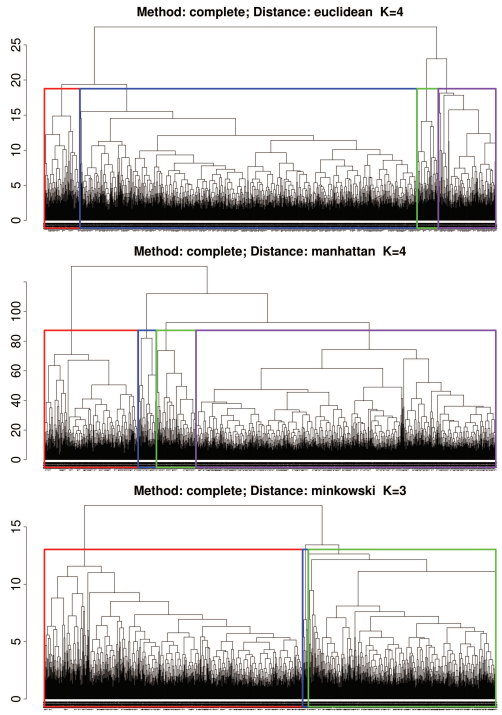

Figure S1: Dendrogram obtained working on the whole set of variables applying Complete Linkage and using Euclidean, Manhattan and Minkowski distance. The coloured rectangles indicate the clusters identified according to the number of clusters suggested by validation indices.

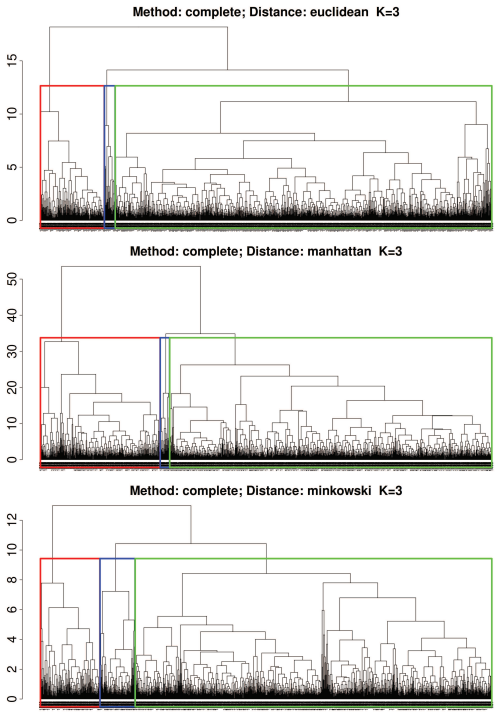

Figure S2: Dendrogram obtained working on the selected variables only and applying Complete Linkage and using Euclidean, Manhattan and Minkowski distance. The coloured rectangles indicate the clusters identified according to the number of clusters suggested by validation indices.

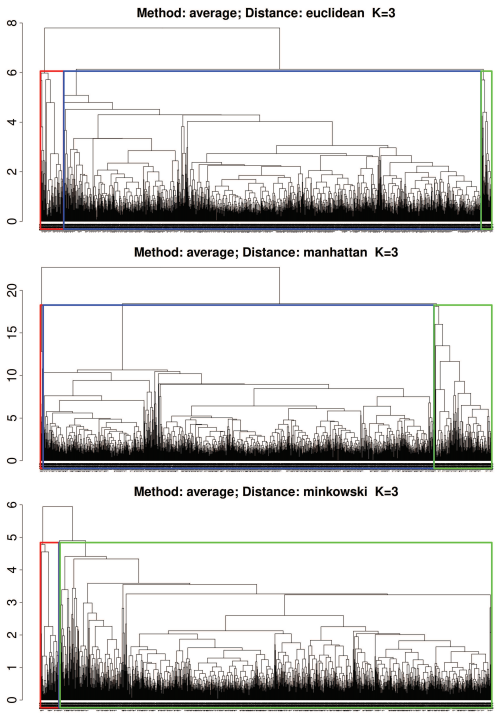

Figure S3: Dendrogram obtained working on the selected variables only and applying Average-Linkage and using Euclidean, Manhattan and Minkowski distance. The coloured rectangles indicate the clusters identified according to the number of clusters suggested by validation indices.

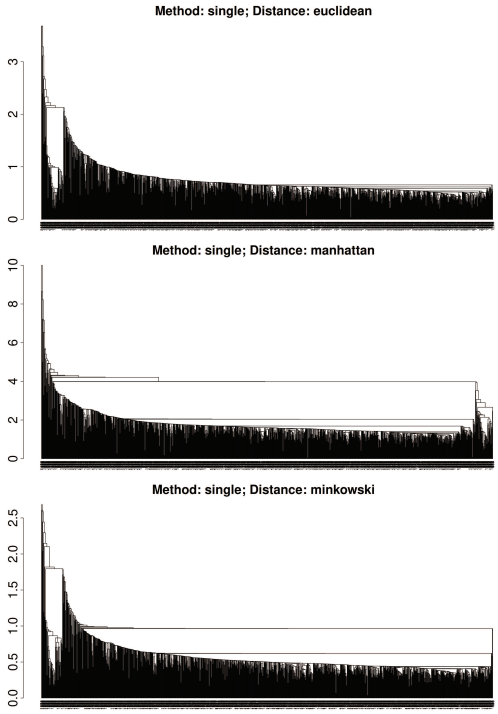

Figure S4: Dendrogram obtained working on the whole set of variables and showing a poor structure.