



Unsupervised Classification of Acoustic Echoes from Two Krill Species in the Southern Ocean (Ross Sea)

Ignazio Fontana¹, Giovanni Giacalone¹, Riccardo Rizzo², Marco Barra^{3(✉)}, Olga Mangoni⁴, Angelo Bonanno¹, Gualtiero Basilone¹, Simona Genovese¹, Salvatore Mazzola¹, Giosuè Lo Bosco^{5,6}, and Salvatore Aronica¹

¹ IAS-CNR, National Research Council, Campobello di Mazara, TP, Italy

² ICAR-CNR, National Research Council, Palermo, Italy

³ ISMAR-CNR, National Research Council, Napoli, Italy

marco.barra@cnr.it

⁴ Department of Biology, University of Naples Federico II, Napoli, Italy

⁵ DMI, University of Palermo, Via Archirafi, Palermo, Italy

⁶ IEMEST, Via Miraglia 20, Palermo, Italy

Abstract. This work presents a computational methodology able to automatically classify the echoes of two krill species recorded in the Ross sea employing scientific echo-sounder at three different frequencies (38, 120 and 200 kHz). The goal of classifying the gregarious species represents a time-consuming task and is accomplished by using differences and/or thresholds estimated on the energy features of the insonified targets. Conversely, our methodology takes into account energy, morphological and depth features of echo data, acquired at different frequencies. Internal validation indices of clustering were used to verify the ability of the clustering in recognizing the correct number of species. The proposed approach leads to the characterization of the two krill species (*Euphausia superba* and *Euphausia crystallorophias*), providing reliable indications about the species spatial distribution and relative abundance.

Keywords: Krill identification · Ross Sea · Acoustic data · Machine learning for pelagic species classification

1 Introduction

In the last decades, fishery science widely used acoustic-based technique to obtain information about the spatial distribution and abundance of economically and ecologically important pelagic organisms characterized by aggregative behaviour. Such organisms usually live in groups, often referred to as school or shoals, and thus are easily detected by using acoustic methods. The use of scientific echo-sounder allowed to investigate large sea sectors in a relatively

L. B. Giosuè and A. Salvatore—Equal contribution.

© Springer Nature Switzerland AG 2021

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12666, pp. 65–74, 2021.

https://doi.org/10.1007/978-3-030-68780-9_7

small amount of time, leading to a synoptic and spatially detailed view of the status of aquatic resources. Usually, acoustic data are recorded along specific routes following a parallel-transects survey design. Biological sampling is performed to identify the species inhabiting the water column thus partitioning the recorded echoes among the observed species. Even if the acquisition of acoustic data is a non-invasive procedure, the biological sampling is not, and the sampling effort strongly depends on several factors such as the number of species characterizing the considered ecosystem, the spatial overlap among species, and the possibility to discriminate among different species based on specific acoustic characteristics and/or the shape and structure of observed aggregations. In some complex operative scenarios or particularly vulnerable ecosystems, the possibility to discriminate among species utilizing semi-automatic classification procedures, thus avoiding or reducing the biological sampling effort, represents an important aspect. Recently, a number of scientific papers focused the attention on this topic [1–4]. Anyway, in mixed-species ecosystems, due to a number factors affecting the characteristics of observed echoes, it is difficult to develop a fully-automatic procedure matching echoes and species [5], and it is necessary to contextualize and validate the procedure according to a deep knowledge of the biology and behaviour of the target species. In this work, we tested the use of an unsupervised clustering algorithm (k-means), to partition 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. Acoustic data collected during the survey and relative to the upper water column stratum showed mainly the presence of two krill species, namely *Euphausia superba* (Dana, 1850) and *Euphausia crystallophias* (Holt & Tattersall, 1906). In this context, it was evaluated if the performed classification confirmed some general features (related to the spatial distribution, relative biomass and energetic differences) reported in the literature, providing a way to obtain information about population status even in the case the biological sampling was missing or non-representative.

2 Materials and Methods

2.1 Acoustic Data: Acquisition and Processing

Acoustic data were collected in the period 05/01/2017–11/02/2017 during the XXXII Antarctic expedition on board of the R/V *Italica* under the Italian National Antarctic Research Program and in the framework of P-ROSE project (Plankton biodiversity and functioning of the Ross Sea ecosystems in a changing southern ocean). In particular, acoustic data were collected through EK60 scientific echo-sounder at three different frequencies (38 kHz, 120 kHz and 200 kHz) and calibrated following standard techniques [6]. Acoustic sampling followed an opportunistic strategy (Fig. 1), recording data among the sampling stations. A total of 2200 nmi were recorded. Acoustic raw data were then processed through Echoview[®] software [7] to extract all the echoes related to aggregations of pelagic

organisms. In the first step, the depth range for the analysis was defined. In particular, the region between 0 and 8.5 m depth was excluded, avoiding artefacts due to beam formation distance and noise due to cavitation and waves. Similarly, the echogram region related to depths higher than 350 m was removed due to the strong attenuation of signals at 120 kHz and 200 kHz. In a second step, background noise was removed by applying the algorithm proposed by De Robertis and Higginbottom [8]; all the echogram regions affected by another noise type (i.e. instrumental, waves, ice etc.) were identified and removed manually. Finally, working on the 120 kHz frequency, all the aggregations (schools) were identified using school detection module in Echoview[©]. The school detection was applied on the 120 kHz as it was the reference frequency for krill species [9].

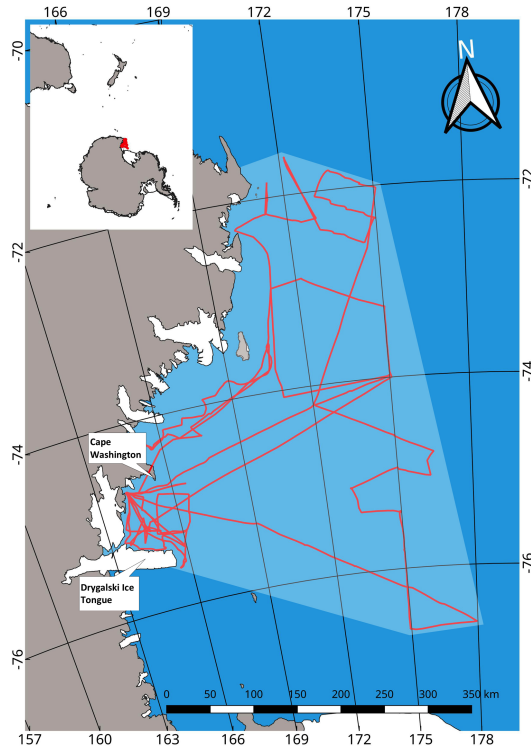


Fig. 1. Study area and acoustic tracks.

Once the school were identified, for each aggregation several parameters related to the energetic, geometric and positioning characteristics were extracted (Table 1).

In addition to the parameters computed by means of Echoview[©] software, four more parameters were computed, namely: the frequency response at 120 and 200 kHz (respectively computed as $FR_{120-38} = Nasc_{120}/Nasc_{38}$ and $FR_{200-38} =$

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. The variables that were log-transformed (see Sect. 2.2) are indicated by using the # symbol (in the case of Sv min, the log transformation was applied only on the 120 kHz)

Parameter	Units	Description
Lat		Average latitude
Lon		Average longitude
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
Beam volume sum#	m^3	Sum of the beam volumes
Sv mean*	$\text{dB re } 1 \text{ m}^{-1}$	Average recorded Sv value
NASC*#	m^2/nmi^2	Nautical Area Scattering Coefficient
Sv max*	$\text{dB re } 1 \text{ m}^{-1}$	Maximum recorded Sv value
Sv min*	$\text{dB re } 1 \text{ m}^{-1}$	Minimum recorded Sv value
Standard deviation*#		Standard deviation of Sv values
Skewness*#	$\text{dB re } 1 \text{ m}^{-1}$	Skewness of Sv values
Horizontal roughness*#	$\text{dB re } 1 \text{ m}^2/\text{m}^3$	Horizontal dispersion of acoustic energy within the school
Vertical roughness*#	$\text{dB re } 1 \text{ m}^2/\text{m}^3$	Vertical dispersion of acoustic energy within the school
FR_120_38#		Frequency ratio
FR_200_38#		Frequency ratio
$\Delta MVBS$	$\text{dB re } 1 \text{ m}^{-1}$	Difference of average Sv 120 kHz minus 38 kHz of the school
$AP.ratio$ #		Ratio between perimeter and area

$Nasc_{200}/Nasc_{38}$ [2]), the difference of MVBS at 120 and 38 kHz ($\Delta MVBS = MVBS_{120} - MVBS_{38}$) [10], and the ratio between the school area and perimeter as an index of shape compactness ($AP.ratio = Perimeter^2/Area$). The resulting data matrix was characterized by 4482 rows and 35 columns.

2.2 Exploratory Analysis and Data Preparation

A preliminary data analysis was carried out to evaluate the presence of outliers and multicollinearity, as they can negatively impact on the clustering performance. The presence of outliers could lead to a bad clustering, while strongly correlated variables could over-weight a specific aspect in building the clusters. To reduce the effect of multicollinearity and to highlight the presence of multivariate outlier Principal Component Analysis (PCA) was carried out. In applying PCA it is important to scale the variable if they are expressed in different units, to avoid that a specific set of variables gain much importance only due to a scale problem. Furthermore, the presence of highly skewed distribution and/or non-linear relationships among variables could introduce distortion in the axes rotation thus leading to incorrect ordination. Before applying PCA all the variables characterized by highly skewed probability distribution were natural log-transformed (Table 2, variables marked by using the # symbol). The performance

of transformation was checked both using a statistical test (Shapiro-Wilks) and by inspecting the qq-plot. Small deviations from normality were considered not impacting PCA results and were ignored. Subsequently, all the variables were scaled and centred. Based on the PCA results, only the principal components (PC) accounting for more than 80% of the total variance was retained and used for clustering.

2.3 Clustering

Let $X = \{x_1, \dots, x_n\}$ a dataset, d a distance measure between element of X , and C_1, \dots, C_k the k clusters found by a generic clustering algorithm. K-means clustering algorithm is one of the most used among the unsupervised clustering methods. The clustering procedure identify the clusters by minimizing the following function:

$$J = \sum_{i=1}^k \sum_{x \in C_i} d^2(x, c_i) \quad (1)$$

where c_i is the mean (centroid) of elements belonging to cluster C_i .

In this context, standardization is an important preprocessing step to avoid scale problems. Besides, the number of clusters must be defined a priori. In the present study, the correct k value is 2, as only two species were found in the echogram. Anyway, to test if the number of groups was an intrinsic property of the data matrix, or if sub-group could be found in terms of specific acoustic and morphological features, the number of clusters was validated employing validation indices. In particular, due to the lack of the *true* schools classification, internal indices were used. Internal validation indices are based on the concept of “good” cluster structure [11, 12]. In particular, in the present study, four validation indices were used, to verify if the number of clusters was correctly identified. All of them are based on the *compactness* and *separation* measures, i.e. the average distance between elements inside the same clusters and the average distance of elements belonging to different clusters. In the following, the used indices are formally defined.

The Silhouette index [13] combines the compactness and separation according to the following formula:

$$S(k) = \frac{1}{k} \sum_{i=1}^k \frac{1}{n_i} \sum_{x \in C_i} \frac{b(x) - a(x)}{\max\{b(x), a(x)\}} \quad (2)$$

Assuming C_i as the cluster of n_i elements where x belongs, $a(x)$ is the average distance between x and all other elements in C_i while $b(x)$ is the average distance between x and all the elements belonging to all the clusters C_j with $j \neq i$.

The Calinski-Harabasz index [14] evaluate the number of cluster according to the following:

$$CH(k) = \frac{\sum_{i=1}^k n_i \sum_{x \in C_i} \frac{d^2(c_i, g)}{(k-1)}}{\sum_{i=1}^k \sum_{x \in C_i} \frac{d^2(x, c_i)}{(n-k)}} \quad (3)$$

In the case of the Dunn index [15] the clustering is evaluated based on the ratio between the separation and compactness:

$$D(k) = \frac{\min_{i \neq j, i, j \leq k} d(c_i, c_j)}{\max_{i \leq k} \max_{x, y \in C_i, x \neq y} d(x, y)} \quad (4)$$

Finally, the Hartigan index [16] takes into consideration the ratio between the compactness of two clustering solutions relative to $k - 1$ and k , in the following way:

$$H(k) = \left(\frac{\sum_{i=1}^{k-1} \sum_{x \in C_i} d^2(x, c_i)}{\sum_{i=1}^k \sum_{x \in C_i} d^2(x, c_i)} - 1 \right) (n - k - 1) \quad (5)$$

For all the above-mentioned indices, the optimal number of clusters is the one maximizing the index value.

3 Results

The skewness index computed for each considered variables showed the presence of highly positive skewed variables. In order to reduce the degree of skewness, all the variables characterized by a skewness index higher than 2 were log-transformed. PCA highlighted the presence of strong patterns (Table 2); the first 5 PC's accounted for about 83% of the total variance and were selected to be subjected to k-means clustering. In particular, the first and second PCs accounted for more than 50% of the total variance and the first PC was strongly related to energetic-related variables (Table 2), while the second one to geometric ones (Perimeter, Area, Length and AP.ratio). The remaining PCs were correlated to a lower number of variables all related to energetic aspects except the 5th PC that was found significantly correlated to the Height_mean only. In terms of outliers, plotting the observation in the PC spaces does not evidence the presence of erratic data points. Internal validation indices (Table 2) were computed on the first 5 PCs (accounting for most of the variance), by testing a vector of cluster numbers from 2 to 10 using the k-means algorithm with d corresponding to Euclidean distance. All considered validation indices highlighted $k = 2$ as the best solution (Table 3).

K-means partitioning was then applied considering the first 5 PCs and $k = 2$. Partitioning results identified 2367 observations as belonging to the cluster 1 and 2217 belonging to cluster 2. Plotting observations (categorized by cluster id) in the PCs space, highlighted that the clustering was mainly driven by the 1th PC (Fig. 2).

In particular, looking at variables correlation values of the 1th PC, the first cluster was characterized by lower energetic values and a more homogeneous internal structure than the second one. Finally, to evidence possible differences in the spatial distribution, the two clusters were plotted in the geographical space (Fig. 3).

Table 2. PC variables correlation. Only variables characterized by a correlation value higher than 0.6 (absolute value) are reported.

	PC1		PC2		PC3		PC4		PC5	
	cor	cos ²	cor	cos ²	cor	cos ²	cor	cos ²	cor	cos ²
Sv_mean_38	0.82	0.68								
Sv_mean_120	0.74	0.55								
Sv_mean_200	0.71	0.5								
Sv_max_38	0.88	0.78								
Sv_max_120	0.81	0.65								
Sv_max_200	0.78	0.61								
NASC_38	0.88	0.77								
Sv_max_120	0.79	0.62								
Sv_max_200	0.77	0.6								
Horizontal_roughness_38	0.84	0.7								
Horizontal_roughness_120	0.71	0.5								
Horizontal_roughness_200	0.67	0.45								
Vertical_roughness_38	0.81	0.65								
Vertical_roughness_120	0.72	0.51								
Vertical_roughness_200	0.73	0.53								
Standard_deviation_38	0.86	0.75								
Standard_deviation_120	0.77	0.6								
Standard_deviation_200	0.75	0.57								
Perimeter			0.83	0.69						
Length			0.77	0.59						
Area			0.73	0.53						
AP_ratio			0.72	0.52						
RF_200_38					0.82	0.68				
Sv_min_200					0.67	0.45				
$\Delta MVBS$							0.82	0.66		
RF_120_38							0.82	0.66		
Height_mean									-0.72	0.52

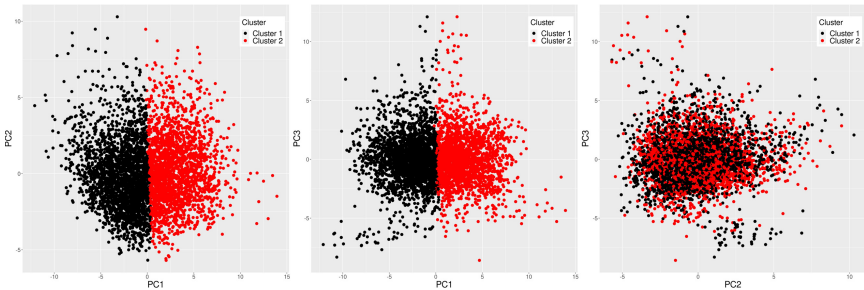
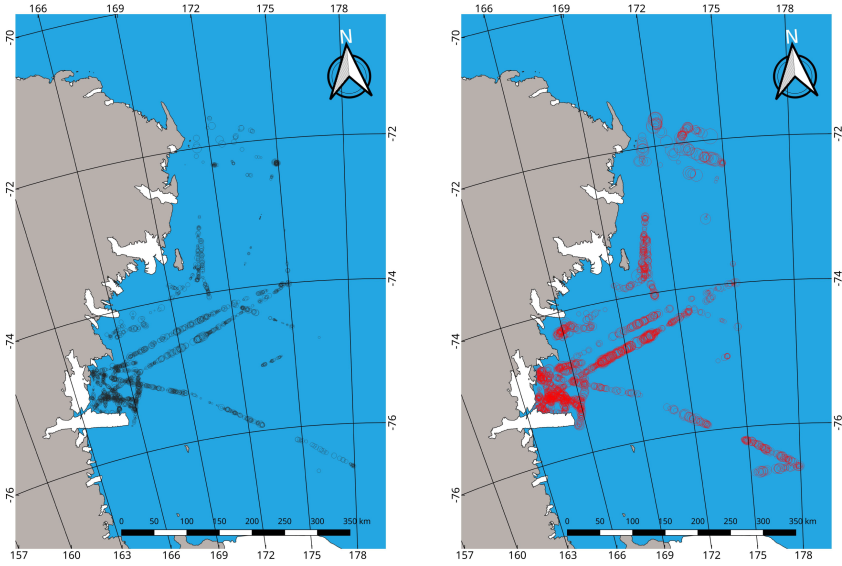


Fig. 2. PCA biplot for the first three components.

Table 3. Validation values of internal indices as the parameter k varies for the “k-means” clustering algorithm and Euclidean distance.

Index	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
S	0.2644	0.1809	0.1857	0.1919	0.1764	0.1679	0.1622	0.1685	0.1677
CH	1910.46	1391.28	1192.81	1095.63	1024.38	957.06	894.72	849.85	812.33
D	1.4078	1.0655	1.1846	1.1246	1.1636	1.0462	0.8965	0.8449	0.8605
H	1910.46	611.68	491.29	447.37	374.14	289.89	228.59	223.85	203.83

**Fig. 3.** Spatial distribution of the two identified cluster (*Euphausia crystallophias* cluster 1, left panel; *Euphausia superba*, cluster 2, right panel). Circle size are proportional to the natural logarithm of $NASC_{120}$ values.

4 Discussion

Euphausia superba and *Euphausia crystallophias* are two key species in the Ross Sea trophic web. Due to their importance, several studies focused on their spatial distribution and abundance through acoustic methods [17–19]. Validation clustering indices successfully identified the correct number of clusters, providing the first indication of k-means performance. Also, according to literature, *Euphausia superba* is most abundant than *Euphausia crystallophias*; Azzali et al. (2006) [17] evidenced a value of 8.6 for the ratio between the biomass of the former and one of the latter species. By considering the $NASC_{120}$ as an abundance index, the above-mentioned ratio according to the obtained classification is 7.9 thus comparable to the one reported in the literature. Looking at spatial distribution, according to literature [18], obtained classification also evidenced the dominance of *Euphausia superba* in the northern sector of the study area (Fig. 3). Finally, it must be considered that in terms of acoustic properties

the two species evidenced a clear separation when comparing, using a scatter-plot, the Sv_{38} vs Sv_{120} [19]. This separation was also confirmed by the k-means classification results (Fig. 4).

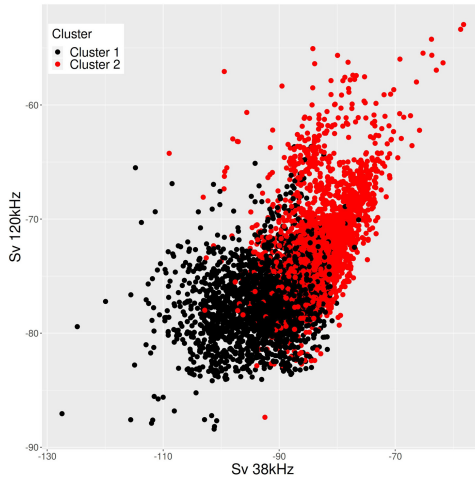


Fig. 4. Scatter plot of Sv_{mean_38} vs Sv_{mean_120} values, categorized according to clustering results (cluster 1 is black).

Due to the agreement between the information reported in the literature and the ones obtained from the classification results, the application of the k-means algorithm could be considered as a stable, fast and reliable solution to extract the main features of the two population from acoustic data, allowing the extraction of summary indices about the population status despite the lack of biological sampling.

5 Conclusions

Unsupervised classification of aggregations detected during acoustic surveys represents a useful tool in the post-processing of large acoustic dataset. In the present work, k-means clustering was able to distinguish between the two considered krill species, recognizing the correct number of clusters and providing indications about the species spatial distribution and relative abundance coherent to the ones reported in the literature. We plan to use the same methodology by employing other clustering algorithms, such as the hierarchical ones, and other cluster validation indices, to improve the consensus about the clustering solution.

References

1. Fernandes, P.G.: Classification trees for species identification of fish-school echotraces. *ICES J. Mar. Sci.* **66**, 1073–1080 (2009)

2. D'Elia, M., et al.: Analysis of backscatter properties and application of classification procedures for the identification of small pelagic fish species in the central Mediterranean. *Fish. Res.* **149**, 33–42 (2014)
3. Fallon, N.G., Fielding, S., Fernandes, P.G.: Classification of Southern Ocean krill and icefish echoes using random forests. *ICES J. Mar. Sci.* **73**, 1998–2008 (2016)
4. Aronica, S., et al.: Identifying small pelagic Mediterranean fish schools from acoustic and environmental data using optimized artificial neural networks. *Ecol. Inform.* **50**, 149–161 (2019)
5. Campanella, F., Christopher, T.J.: Investigating acoustic diversity of fish aggregations in coral reef ecosystems from multifrequency fishery sonar surveys. *Fish. Res.* **181**, 63–76 (2016)
6. Foote, K.G., Knudsen, H.P., Vestnes, G., MacLennan, D.N., Simmonds, E.J.: Calibration of acoustic instruments for fish density estimation: a practical guide. *ICES Coop. Res. Rep.* **144**, 69 (1987)
7. Higginbottom, I., Pauly, T.J., Heatley, D.C.: Virtual echograms for visualization and post-processing of multiple-frequency echosounder data. In: *Proceedings of the Fifth European Conference on Underwater Acoustics*, pp. 1497–1502. Ecua (2000)
8. De Robertis, A., Higginbottom, I.: A post-processing technique to estimate the signal-to-noise ratio and remove echosounder background noise. *ICES J. Mar. Sci.* **64**, 1282–1291 (2007)
9. Leonori, I., et al.: Krill distribution in relation to environmental parameters in mesoscale structures in the Ross Sea. *J. Mar. Syst.* **166**, 159–171 (2017)
10. Watkins, J.L., Brierley, A.S.: Verification of the acoustic techniques used to identify Antarctic krill. *ICES J. Mar. Sci.* **59**, 1326–1336 (2002)
11. Hassani, M., Seidl, T.: Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. *Vietnam J. Comput. Sci.* **4**(3), 171–183 (2016). <https://doi.org/10.1007/s40595-016-0086-9>
12. Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A.: NbClust: an R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* **61**, 1–36 (2014)
13. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
14. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. *Comm. Stat.* **3**, 1–27 (1974)
15. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **4**, 95–104 (1974)
16. Hartigan, J.A.: *Clustering Algorithms*. Wiley, New York (1975). ISBN 047135645X
17. Azzali, M., Leonori, I., De Felice, A., Russo, A.: Spatial-temporal relationships between two euphausiid species in the Ross Sea. *Chem. Ecol.* **22**, 219–233 (2006)
18. Davis, L.B., Hofmann, E.E., Klinck, J.M., Pinones, A., Dinniman, M.S.: Distributions of krill and Antarctic silverfish and correlations with environmental variables in the western Ross Sea, Antarctica. *Mar. Ecol. Progress Ser.* **584**, 45–65 (2017)
19. La, H.S., et al.: High density of ice krill (*Euphausia crystallorophias*) in the Amundsen sea coastal polynya, Antarctica. *Deep Sea Res.* **95**, 75–84 (2015)