Contents

Description

The Ellipsoid score (hereafter Escore) algorithm is a multi-frequency backscatter classification tool which involves selecting regions of interest (ROIs) within multifrequency red-green-blue (RGB) echograms and classifying into clusters or echo-classes using S_v differences.

It has been developed under Matlab and R softwares.

Citation: Annasawmy P, Roudaut G, Lebourges Dhaussy A (2024) Impact of an eddy dipole of the Mozambique channel on mesopelagic organisms, highlighted by multifrequency backscatter classification. PLoS ONE 19(9): e0309840. https://doi.org/10.1371/journal.pone.0309840

Glossary of terms

Region of Interest (ROI): Manually chosen rectangular section of the RGB echogram which contains a specific structure of interest with consistent frequency responses.

Echo-type: Each echo-type encompasses a set of individual pixels classified into the same cluster from the ROI definition and the K-means clustering (step 2).

Library: The whole set of echo-types identified from the training dataset.

Echo-class: Each echo-class encompasses several echo-types with similar characteristics and is the result of the echo-type classification (i.e., hierarchical classification) of a library (step 3). The echo-classes form the learning/reference database for the classification of the whole acoustic data.

Implementation

The algorithm is implemented in a series of step-wise procedures summarized below and detailed in Supplementary Material 2 of Annasawmy et al. (2024). The Escore algorithm should be run on processed and echo-integrated multi-frequency acoustic data at high resolutions. We recommend echo-integrating at 1 ping horizontal and 0.5 m vertical resolutions or 3 pings horizontal and 1.5 m vertical resolutions.

Step (1) Selecting ROI

A semi-supervised approach consisting of manually selecting Regions of interest (ROI) from RGB composite echograms.

1.1 Run the script D_echocolors_Define_param

The folder "Echogramme_RGB_fromEI" contains three key MATLAB scripts. The user can modify the settings in the MATLAB script D_echocolors_Define_param to customize various parameters. These include specifying the path to the echointegration file generated by running echo-integration on the acoustic data in the Matecho software, setting the number of elementary sampling units (ESUs), selecting frequency triplets for the RGB echograms, adjusting the low threshold, defining the maximum depth for the echograms, setting the lower limits for each frequency, specifying the range, and deciding whether to apply a convolution to smooth the echogram.

The user can choose to run the script with any combination of three frequencies. In this example, we select 18/38/70 kHz and 38/70/120 kHz as the frequency triplets.

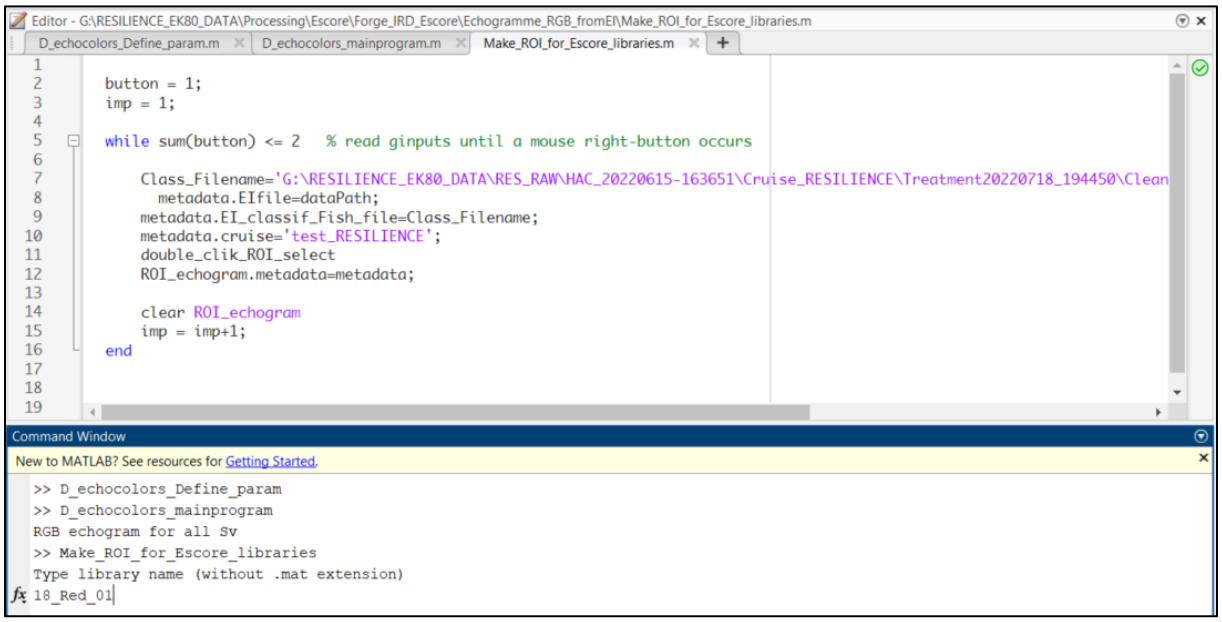
1.2 Run the script D_echocolors_mainprogram

From the same folder "Echogramme_RGB_fromEI", the Matlab script D_echocolors_mainprogram can be run without modifying the code. The script opens an RGB echogram in Matlab. Zoom into a specific portion of the RGB echogram.

1.3 Run the script Make_ROI_for_Escore_libraries

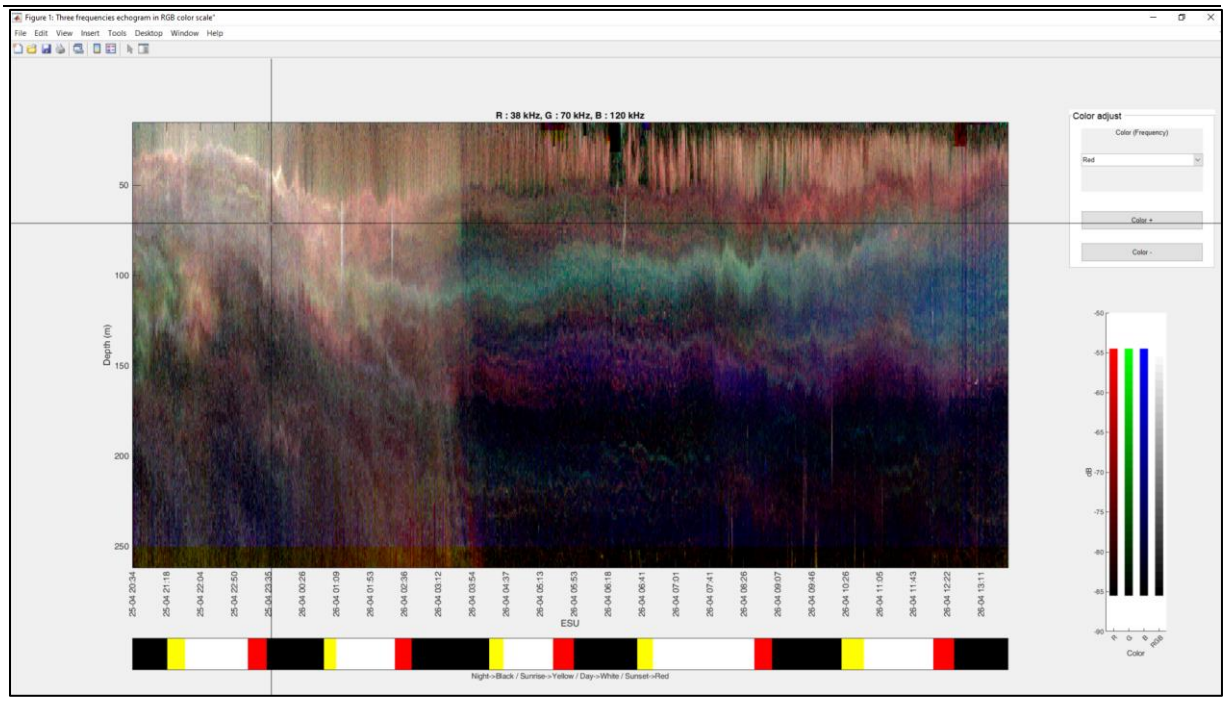
The data path must be provided.

The user is prompted to specify the library name in the MATLAB command window. This includes assigning a name to each Region of Interest (ROI). See example below:

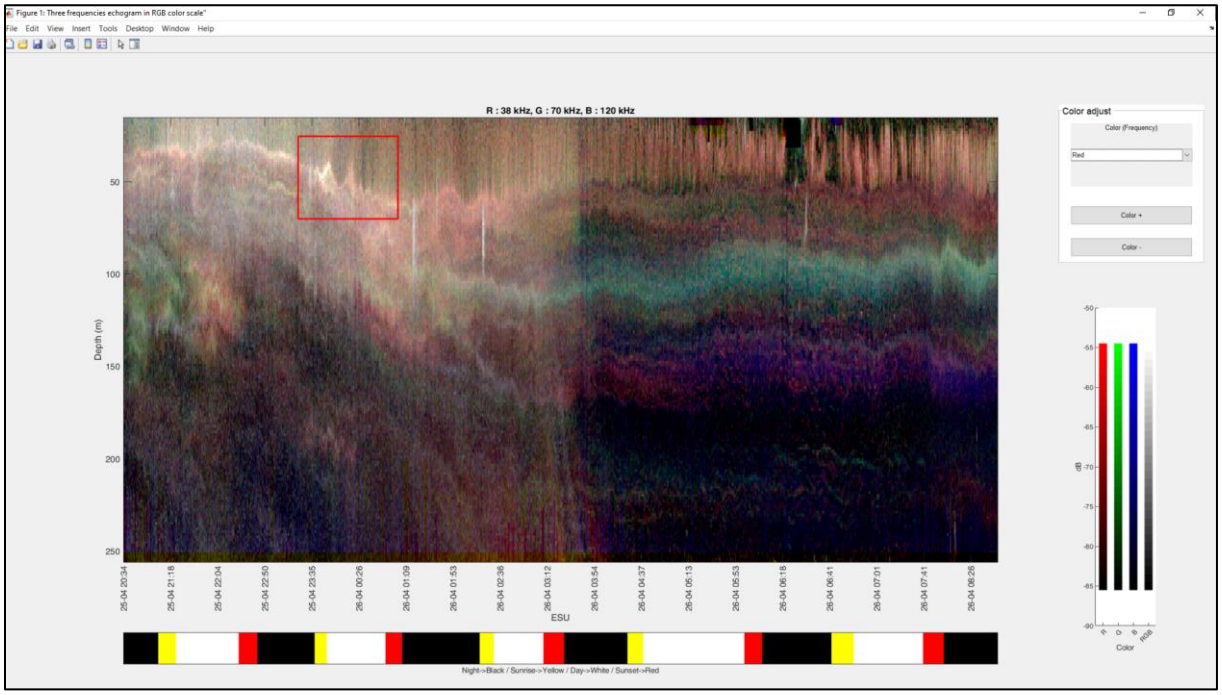


After typing in the Matlab command window, press ENTER on the keyboard.

With the tool that opens in the Matlab figure, click on the RGB plot to select a ROI. Use left-click to continue selecting within that window; complete the selection process by left-clicking and then right-clicking.



ROI should be selected only within the maximum range of the highest frequency. When choosing ROIs, it is recommended to focus on distinct features with Day and Night patterns, differently colored layers, both shallow and deeper structures within the echogram, and it is also good practice to clearly define the lower and upper limits of the feature of interest.



RGB echogram at 18, 38, and 70 kHz showing the selection of one ROI (bounded by the red outlined rectangle). The color bar shows S_v (dB) at each corresponding frequency (Red (R): 18 kHz; Green (G): 38 kHz; Blue (B): 70 kHz).

Run the script Make_ROI_for_Escore_libraries again and specify the name of the next library name before selecting another ROI.

The data for each selected ROI will be stored in a .mat file, and a PNG image of the chosen ROI will also be automatically saved in the "Echogramme_RGB_fromEI" folder.

It is good practice to move all .mat and PNG files to the appropriate subfolder, either "RGB_18_38_70" or "RGB_38_70_120," depending on the frequencies used for ROI selection.

From each set of RGB echograms, aim to select a maximum of more than 50 ROIs.

Step (2) Selecting one echotype

The pixels within each ROI will be subjected to a K-means clustering based on the relative frequency responses at 18, 70, and 120 kHz in dB ($\Delta S_{v,18-38}$, $\Delta S_{v,70-38}$, and $\Delta S_{v,120-38}$) to retain one cluster representing a single coherent visible acoustic structure, i.e., one echo-type. This process ensures that only one acoustic structure with limited variability in its frequency responses is selected. The resulting echo-types will form the training dataset.

2.1 Run the script A_1_Make_library_kmeans_ClusterDiffSv_Calinski

Go to the folder "codes_diff" and run the script A_1_Make_library_kmeans_ClusterDiffSv_Calinski

Several settings can be modified from this script. The user has to specify the path to directory where the ROI.mat files are stored.

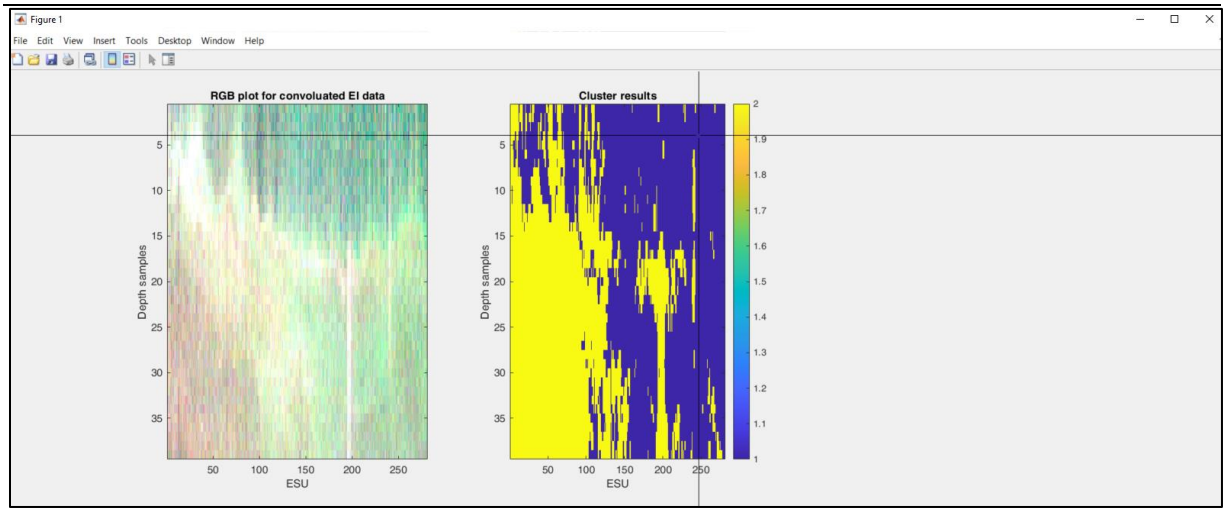
Next, the ROI file number must be specified. Typically, one starts with the 1st file, i.e., by setting file_IDtoProcess=1

The reference frequency can also be chosen. Typically, that is the 38 kHz frequency, so the Freq_ID_for_Ref=[2] since the 38 kHz is the 2nd frequency in our case.

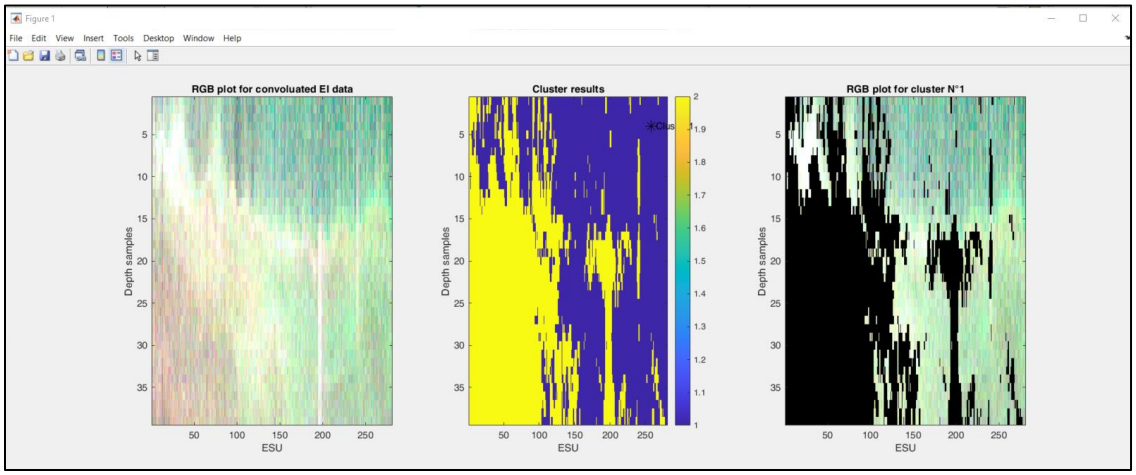
The S_v thresholds can then be set. During the K-means clustering process, the user needs to set the minimum and maximum number of clusters based on the number of distinct features visible within each ROI.

Additionally, the user has the option to plot the frequency response, which is recommended for better visualization. The user can also decide whether to use convoluted data in the clustering process or not.

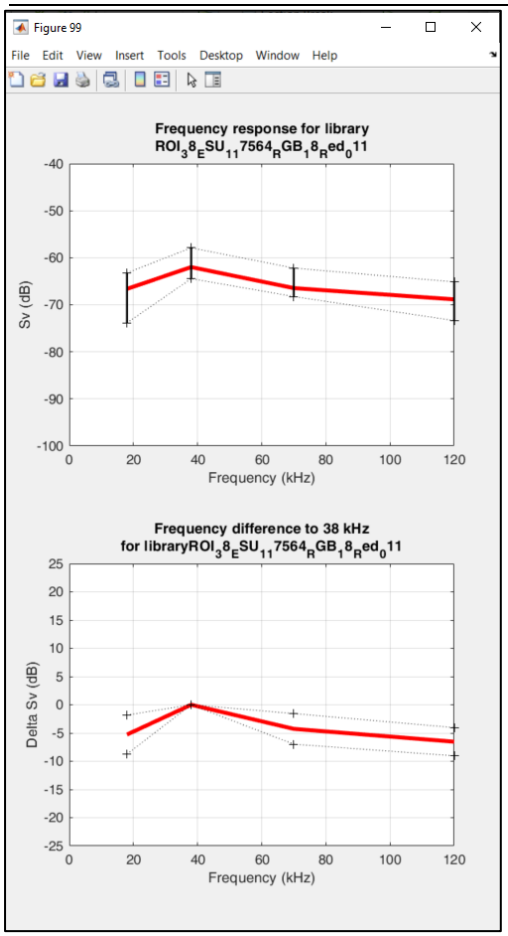
As illustrated below, the K-means algorithm identifies two clusters. The user is then prompted to select the cluster that represents the feature of interest. To do this, left-click on the desired cluster in the "Cluster results" plot.



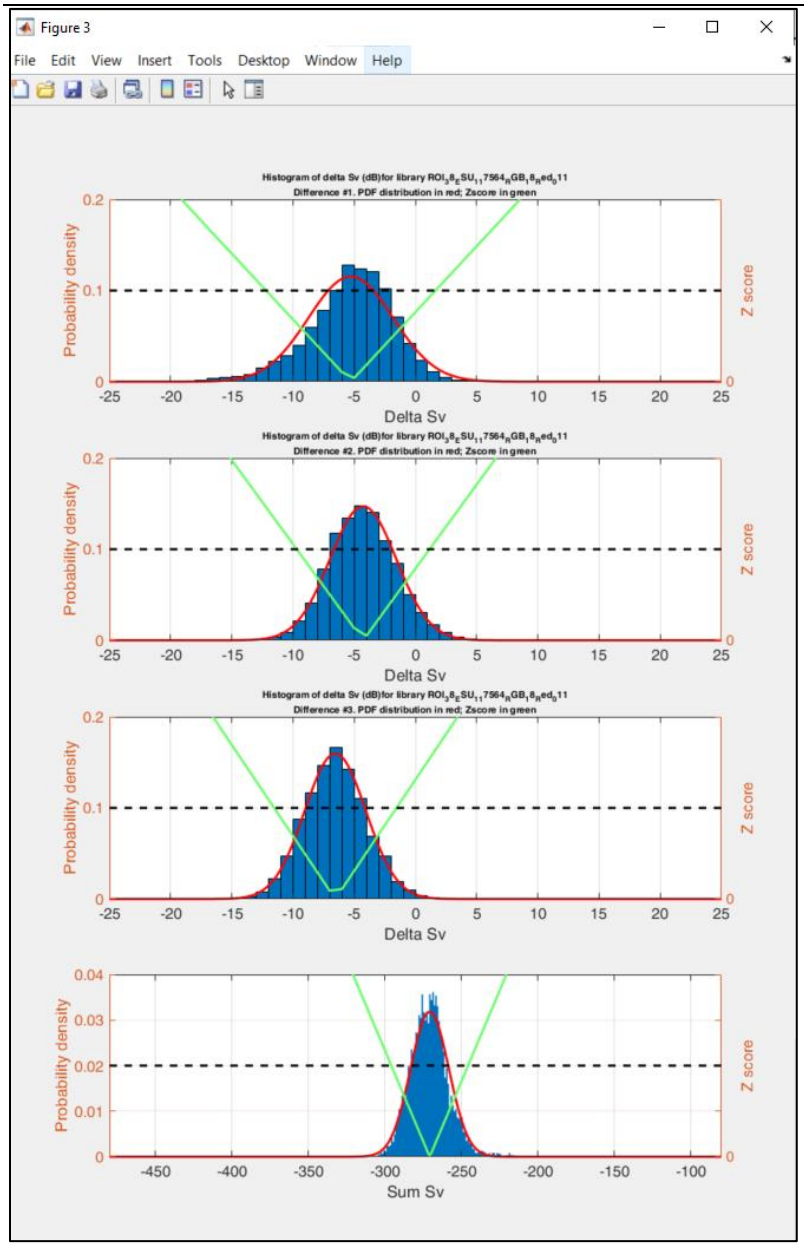
Three figure windows will then open. The right-hand side plot shows the RGB representation of the selected feature of interest.



In the selected ROI, one cluster showing higher frequency response at 38 kHz was chosen following K-means clustering.



Mean frequency response curve (red) with standard deviations (black) of selected echo-type and the curve of frequency difference relative to 38 kHz for the chosen echo-type.



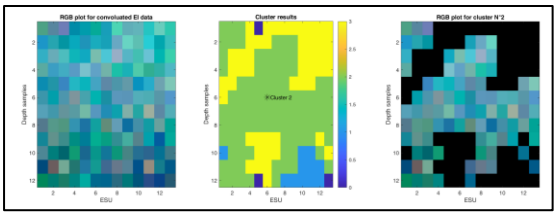
Histograms of $\Delta S_{v,18-38}$, $\Delta S_{v,70-38}$, $\Delta S_{v,120-38}$ for the echo-type is given with the corresponding probability density for a normal distribution.

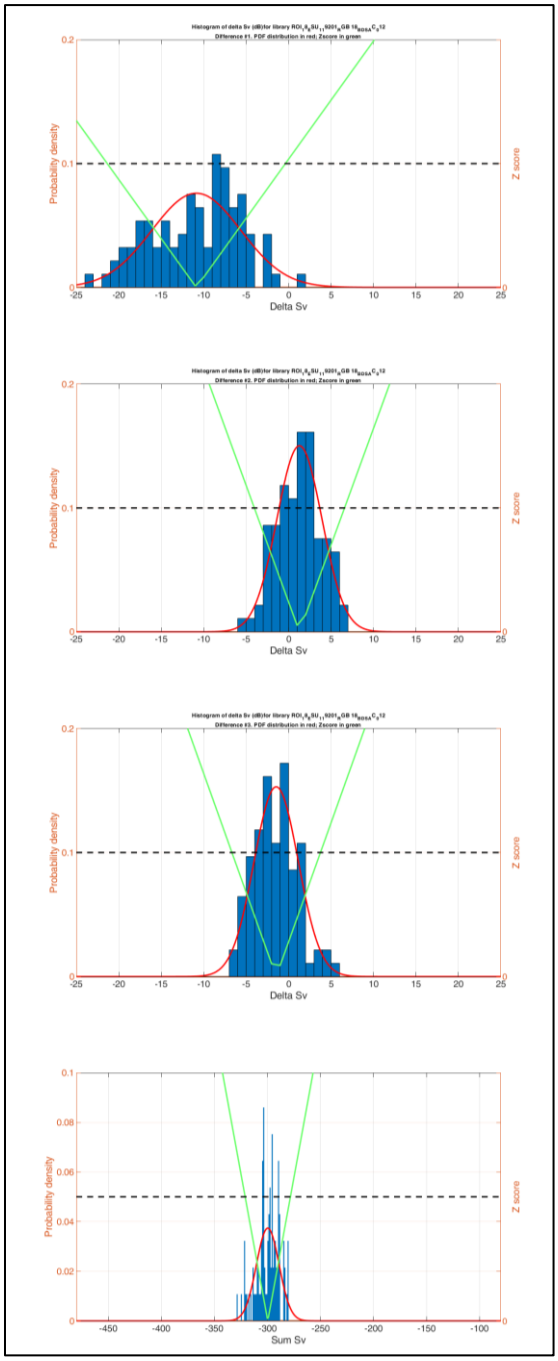
During this step, the user can visualize both the frequency response curves and the probability density functions for each echo-type. Typically, the user should select echo-types that exhibit:

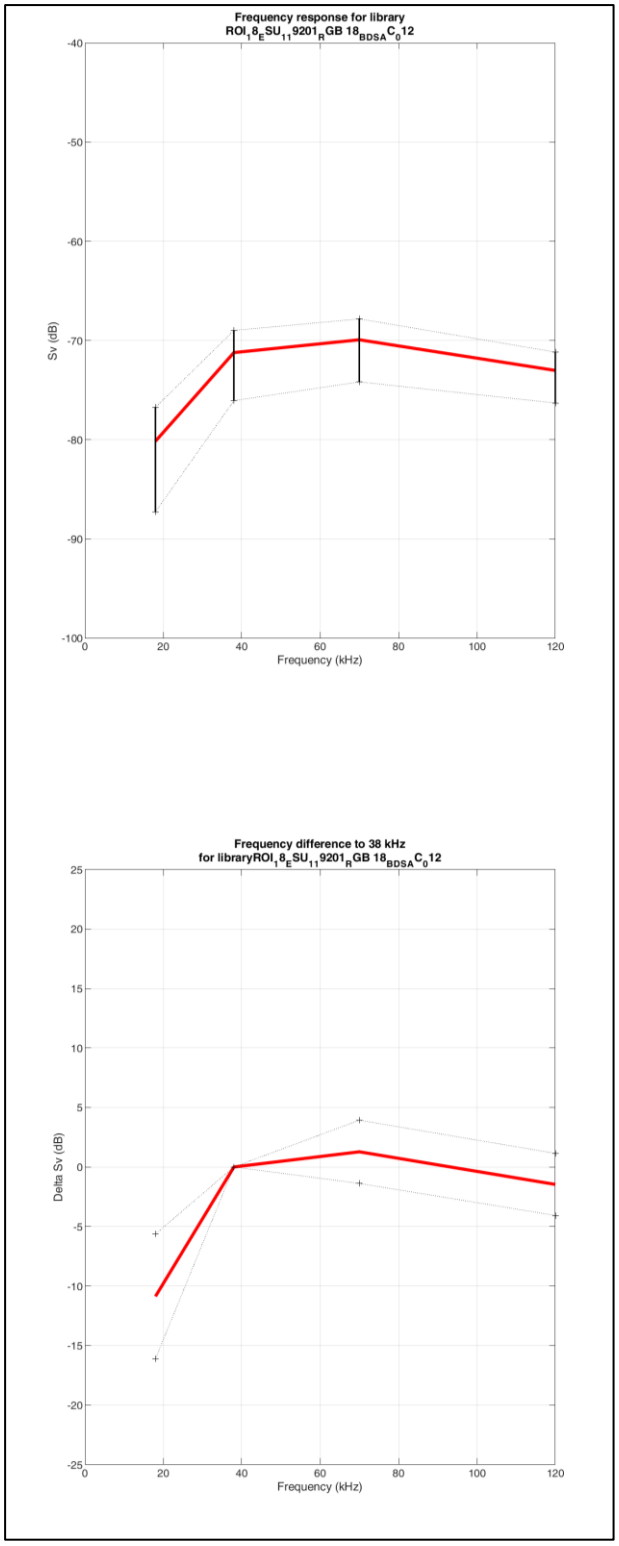
Frequency Response Curves with minimal variations in standard deviations.

Probability Density Functions with normal, Gaussian, or unimodal distributions with "slender" histograms.

Echo-types showing non-normal distributions or significant variations in their frequency response curves should be discarded from further analysis, as illustrated below:







Echo-types showing large variations in the frequency response curves and non-normal or non-Gaussian probability density functions, as shown above, should be discarded from further analysis.

Run the script again for the second ROI.mat file by setting file_IDtoProcess = 2 to process the 2nd ROI.mat file.

Additional Notes

The script may encounter issues if the selected ROI extends deeper than the maximum range of the highest frequency.

It is possible to choose two clusters from the same figure/ROI. To do this:

1) Copy and rename the ROI.mat file (also, if the current file ID is 7, you would create a new file with an ID of 8).

2) Re-run the script, setting file_IDtoProcess = 8 to process the newly created ROI file.

2.2 Run the script A_2_Make_meanSv_diff_forClassif

From the same folder "codes_diff", run the script A_2_Make_meanSv_diff_forClassif

This script calculates S_v frequency differences from the selected echotypes.

The user needs to specify the directory where the final selected echo-types are stored. Additionally, the user must indicate the number of Sv frequency differences to compute. For example:

- If the user is working with echo-types at 4 frequencies (18, 38, 70, and 120 kHz), the number of Sv frequency differences (Ndiff) to compute would be 3.
- If the user is working with 3 frequencies, Ndiff would be 2.

A file named "RES_calinskiClusters_Jan_2023.mat" file will be saved. This file contains the mean and standard deviations of the S_v frequency differences.

Step (3) Hierarchical clustering of mean S_v differences

Mean S_v differences ($\Delta S_{v,18-38}$, $\Delta S_{v,70-38}$, and $\Delta S_{v,120-38}$) are calculated from the echo-types and classified using hierarchical cluster analysis.

3.1 Open the R script region_of_interest_classification

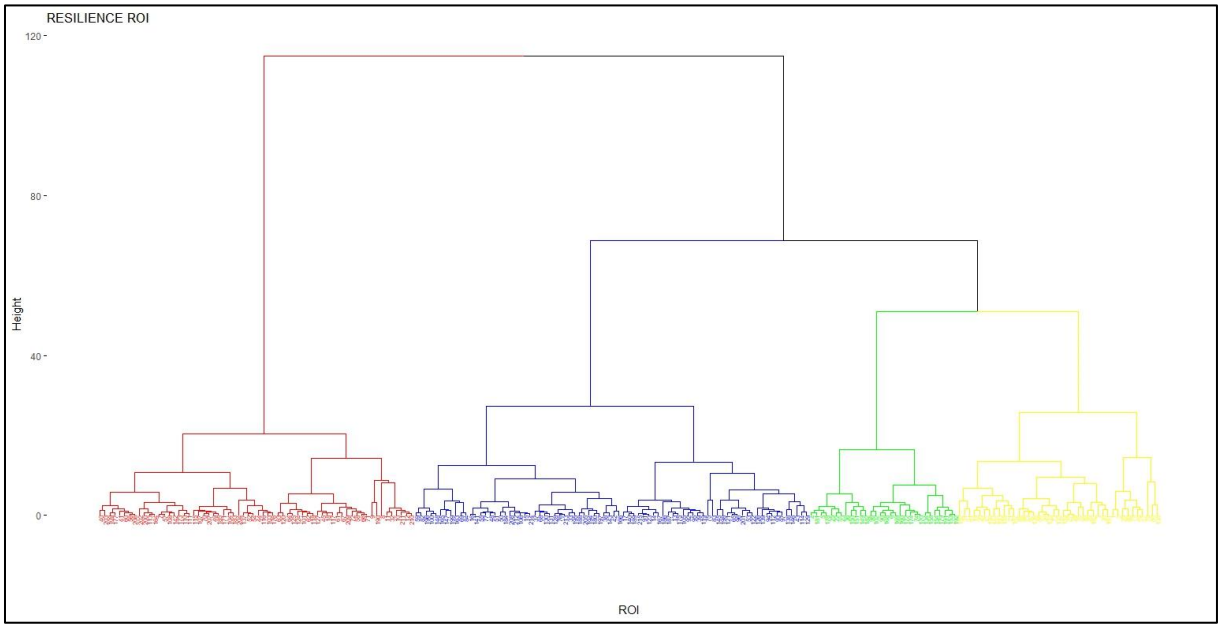
Install and run the required packages. The user has to specify the path to the "RES_calinskiClusters_Jan_2023.mat" file.

A hierarchical clustering is performed on a distance matrix (d.roi) of the mean S_v differences (ROI_meanSvDiff). Prior to running the hierarchical clustering, the optimum number of echo-classes is determined using the NbClust package.

Typically, the optimal number of echo-classes is selected based on the maximum consensus among indices indicating the ideal number. For example, if out of 20 indices implemented in NbClust, 11 indices suggest that 4 is the optimal number of echo-classes, then 4 echo-classes should be chosen.

Additionally, the user can visualize bar plots of inertia and ladder plots to further aid in determining the optimal number of echo-classes.

A dendrogram is plotted with the fviz_dend function by specifying the optimum number of echo-classes (i.e., by setting k=4 for 4 echo-classes).



A dendrogram using Ward's clustering method showing the relative level of dissimilarity between the selected echo-types. The echo-types were classified into 4 echo-classes in this example.

The R cutree function is used to cut the dendrogram into the specified number of clusters. A .csv file "ROI_meanSvDiff.csv" and .txt file "RESILIENCE_classification_groups_4.txt" are saved.

In the .csv file, add a new column named "group_classification." Then, copy the values from the second column of the .txt file and paste them into the newly created "group_classification" column in the .csv file.

Cluster performance can be evaluated using a random forest algorithm using the .csv file. The classification error is estimated from a random subset of the bootstrap data, and the error rate is returned for each echo-class.

Step (4) Visualize the echo-classes in 3-D space, and their frequency responses

A three-dimensional ordination plot is created to visualize the echo-classes. This plot represents the echo-classes along the S_v difference axes: $S_{v18\text{-}38}$ (x-axis), $S_{v70\text{-}38}$ (y-axis) and $S_{v120\text{-}38}$ (z-axis).

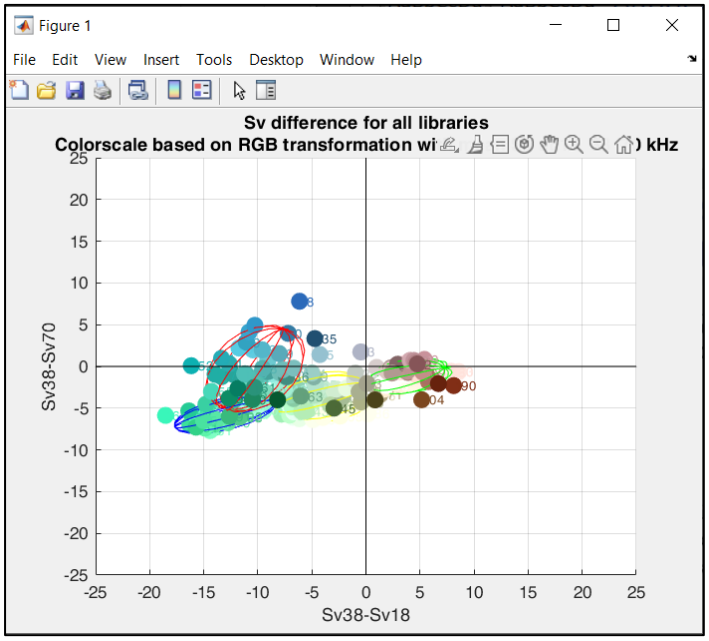
4.1 Run the script B_1_Plot_classif_new3D

From the folder "codes_diff", run the script B_1_Plot_classif_new3D

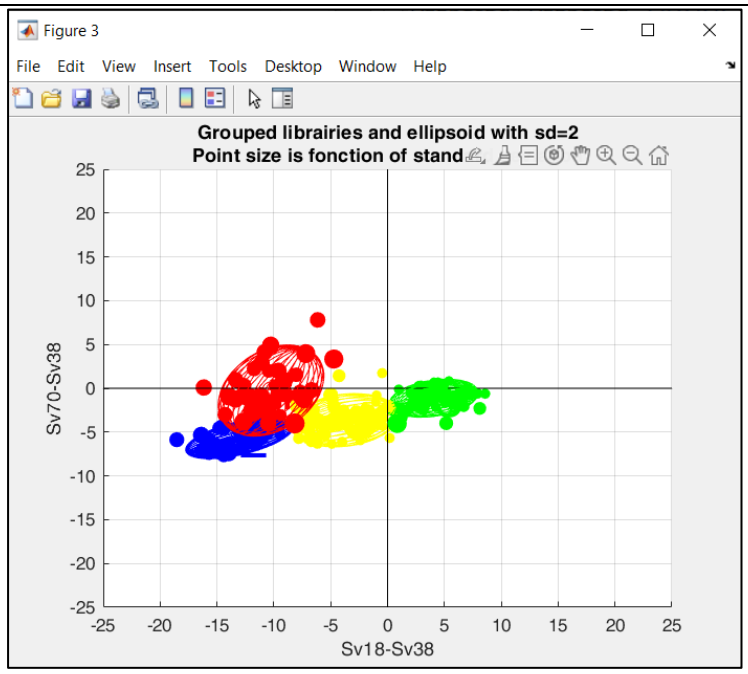
The user has to specify the path to the .mat file resulting from the Calinski clustering (Step 2.2), and the .txt file resulting from the hierarchical clustering (Step 3.1), and the directory containing all ROI files.

The user also has to specify the number of clusters/echo-classes (number_group = 4 in this case), and the colors of each echo-class.

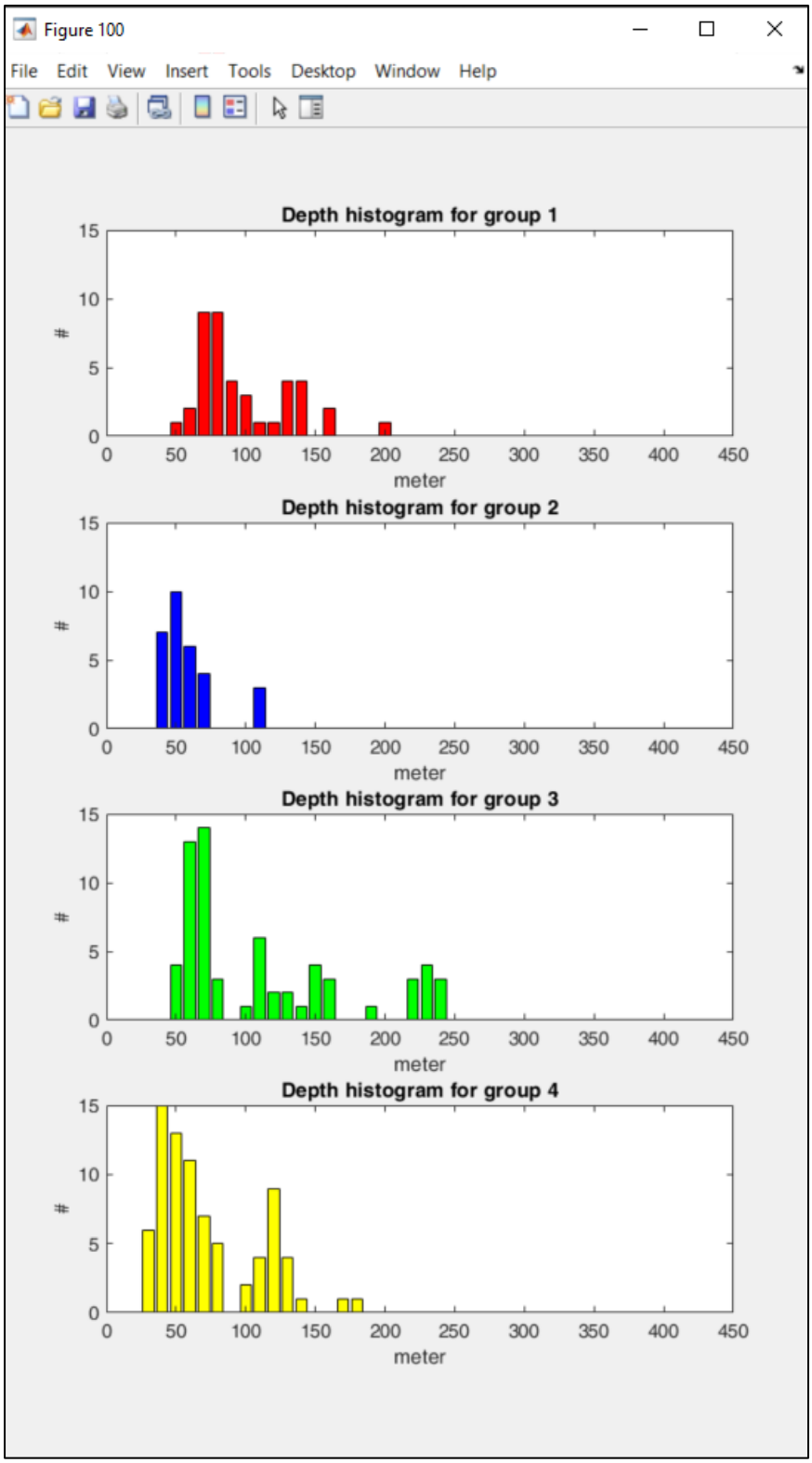
The following figures are plotted after running the script:



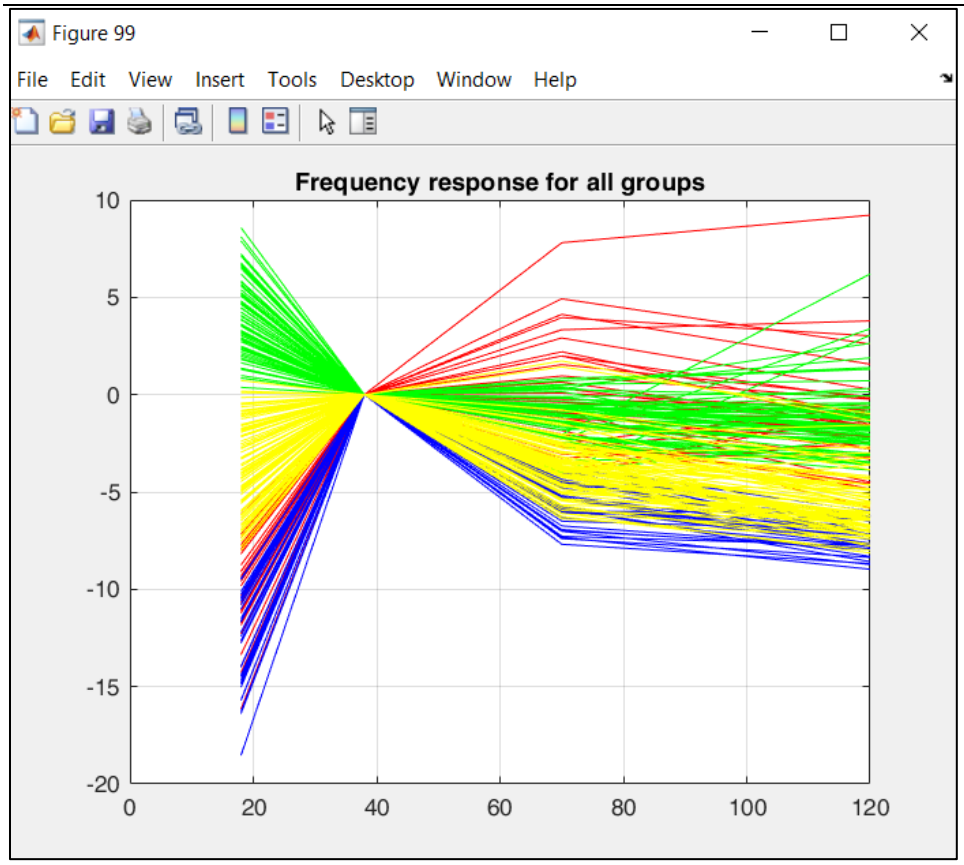
Dot color represents the frequency response, mapped according to the chosen frequencies for the RGB triplet. The numbers represent the file IDs.



Data points representing the S_v differences of each echo-class are plotted around a centroid of an ellipsoid at mean ± 2 standard deviations.



The histograms show the vertical depth distributions (in meters) below the sea level of each echo-class.



Relative frequency response curves of each echo-class are shown above so as to assign the frequency with dominant backscatter to each echo-class in the next steps. Echo-class 1 (red), echo-class 2 (blue), and echo-class 4 (yellow) are dominant at 38 kHz; echo-class 3 (green) is dominant at the 18 kHz.

4.2 Run the script B2_Frequency_response_per_group_plot

This script generates the ellipsoids, i.e., a .mat file containing the centroids, eigenvalues, and eigenvectors.

Step (5) Perform sensitivity tests on the training dataset

5.1 Run the script C_Sensibility_Analysis_Escore_3D

From the folder "codes_diff", run the script C_Sensibility_Analysis_Escore_3D

Sensitivity tests are conducted to determine the classification accuracy of echo-integrated cells. These tests assess:

well-classified Cells: The percentage of cells correctly classified into each echo-class.

misclassified Cells: The percentage of cells incorrectly assigned to an echo-class.

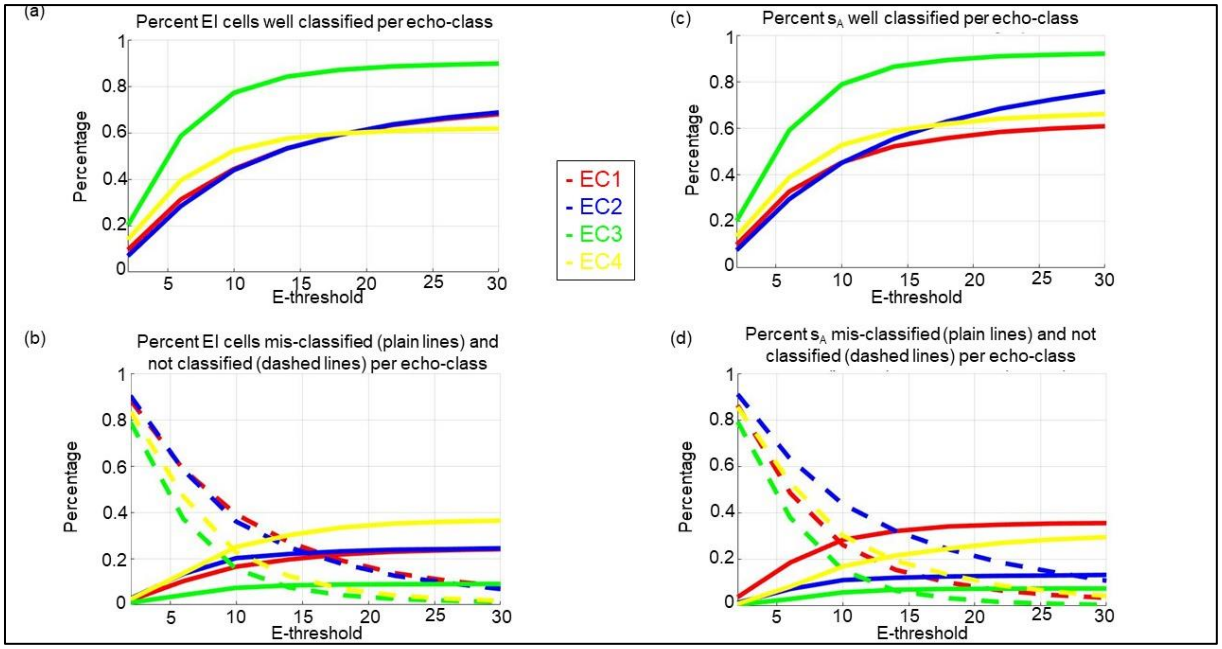
unclassified Cells: The percentage of cells that were not classified into any echo-class.

The user has to specify the path to the .mat file resulting from the Calinski clustering (Step 2.2), the path to the echotypes, the .txt file resulting from the hierarchical clustering (Step 3.1), and the .mat file created from script B2 (Step 4.2).

The user also has to specify an Ellipsoid-threshold (Escore_tresh). In the example script, this is defined as a vector ranging from 2 to 30 in increments of 4. The user can also modify the colors of the echo-classes.

Additionally, in lines 106 to 116, the user needs to assign a dominant frequency to each echo-class based on the frequency response curves shown in the previous figure. For instance, in the example provided:

- Echo-classes 1, 2, and 4 are assigned to the dominant 38 kHz frequency (i.e., 2).
- Echo-class 3 is assigned to the 18 kHz frequency (i.e., 1).



Graphs of the percentage echo-integrated (EI) cells (a) well classified, (b) mis-classified (plain lines) and not classified (dashed lines) per echo-class. Graphs of the percentage s_A (c) well classified, (d) mis-classified (plain lines) and not classified (dashed lines) per echo-class.

In the figures above, an Ellipsoid-threshold value of 25 appears to optimally classify each data point into one of the four echo-classes while limiting the percentage echo-integration cells and s_A mis-classified and decreasing those which were not classified.

Each point in this three-dimensional space can be characterised by its distance from the centroid of each echo-class. This is what we call the Escore which is a sum of squared independent normal random variables following a chi-square distribution with three degrees of freedom. The Escore can be represented by the following equation:

$Escore_{(i,k)} = \frac{x_i^2}{\lambda_{1(k)}} + \frac{y_i^2}{\lambda_{2(k)}} + \frac{z_i^2}{\lambda_{3(k)}}$, where x_i, y_i, and z_i are the coordinates of the echo-integrated cell i in the three-dimensional space defined by the ellipsoid k, and the origin is the arithmetic mean. For a given cell i, $Escore_{(i,k)}$ is calculated for each echo-class k. For that given cell, the higher the Escore, the lower is the similarity of the frequency response to the echo-class. In principle, the cell is assigned to the echo-class for which $Escore_{(i,k)}$ is the minimum.

In order to avoid classifying cells having frequency responses which are different from the defined echo-classes, a maximum threshold (Ellipsoid-threshold) is defined.

The Escore algorithm is:

If $Escore_{(i)}$ for all k echo-classes >= Ellipsoid-threshold, cell i is unclassified

If not, cell i is assigned to echo-class k according to smallest $Escore_{(i,k)}$

Step (6) Run the Escore algorithm

The Escore algorithm is then applied to the entire acoustic dataset, which includes all echo-integration cells, including the cells from the training dataset. This step classifies each data point into one of the predefined echo-classes.

6.1 Run the script D_define_param_Ellipse3D_score

From the folder "codes_diff", run the script D_define_param_Ellipse3D_score

The user has to specify the directory containing all echotypes, and the name of the .mat file created from script B2 (Step 4.2).

The user also has to specify the reference frequency (Freq_ref_for_diff) which is 38 kHz (i.e., 38000) in our case, and the frequencies for the S_v differences (i.e., 18000, 70000. 120000) in our case. The user can also modify the colors for each echo-class, choose to plot echograms of the classified data, add a trawl path or not and set minimum and maximum depths.

6.2 Run the script D_Main_program_Ellipse3D_score_loop

From the folder "codes_diff", run the script D_Main_program_Ellipse3D_score_loop

ESU_number_in_loop=30000 means that the script will divide the entire echointegration file into portions of 30 000 ESUs.

A .mat file is generated containing details of the classification of each ESU.

The following plots are generated:

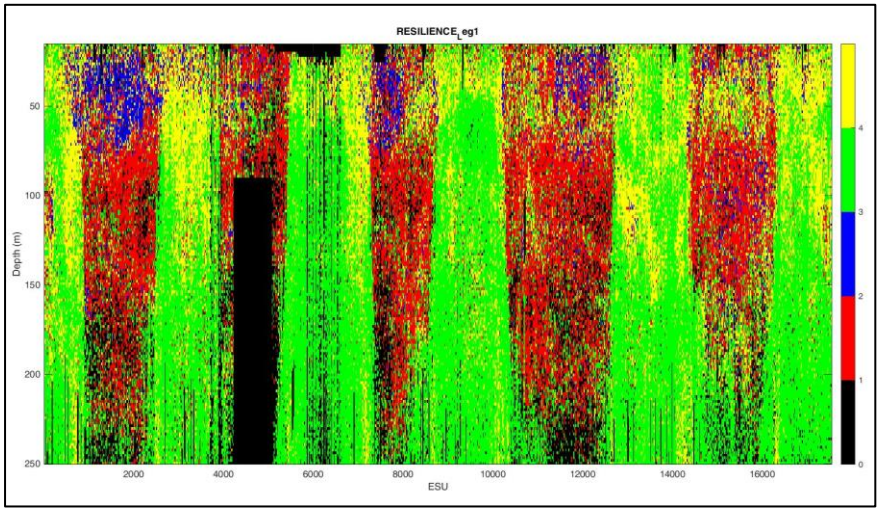
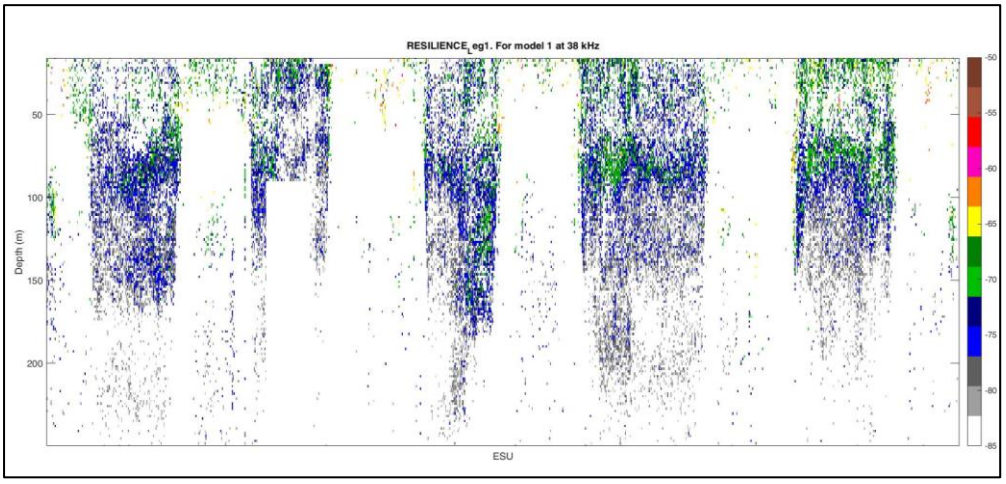
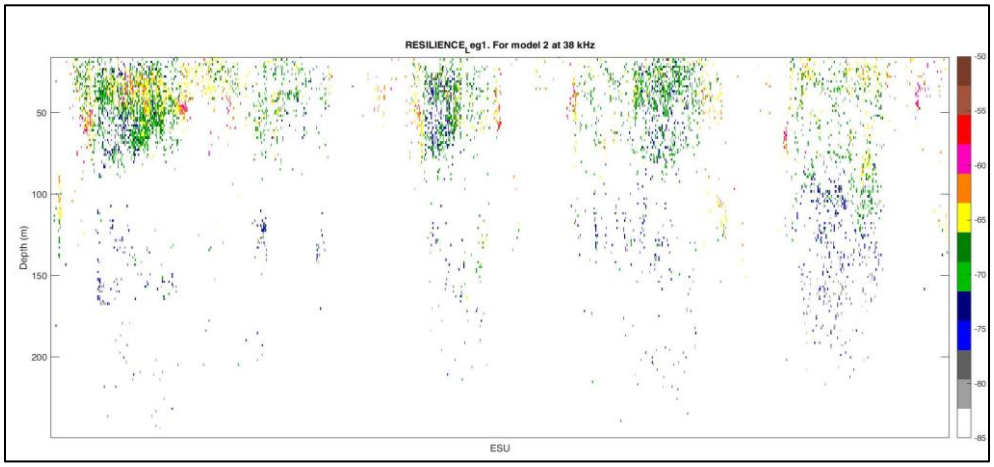


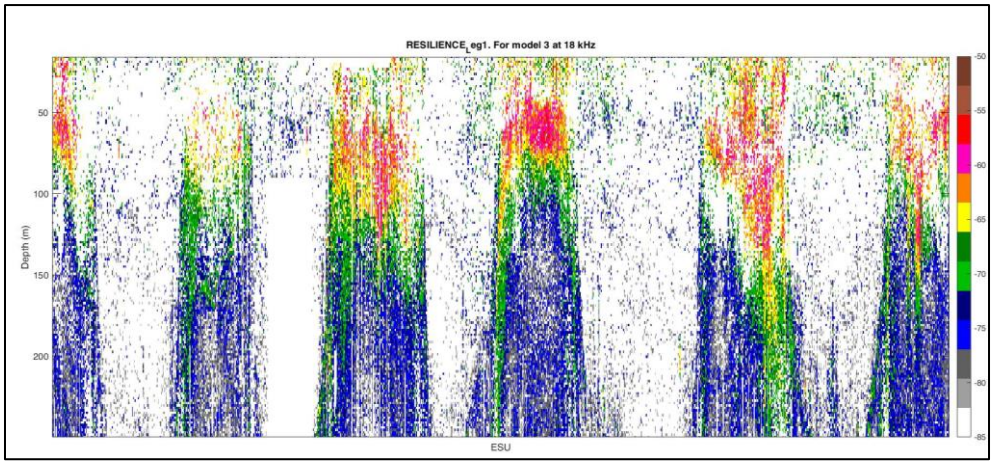
Figure showing the vertical distributions (Depth, m) of each elementary sampling unit classified into one of the four echo-classes. ESUs classified in black represent those which were not classified.



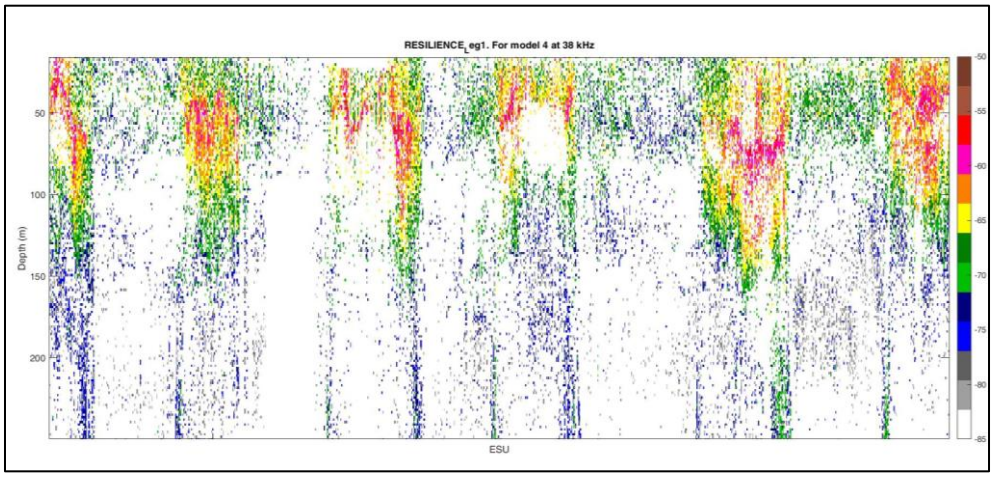
Echogram showing the vertical and temporal distributions of echo-class 1. The color bar represents S_v in dB.



Echogram showing the vertical and temporal distributions of echo-class 2. The color bar represents S_v in dB.



Echogram showing the vertical and temporal distributions of echo-class 3. The color bar represents S_v in dB.



Echogram showing the vertical and temporal distributions of echo-class 4. The color bar represents S_v in dB.

The following additional steps can also be conducted:

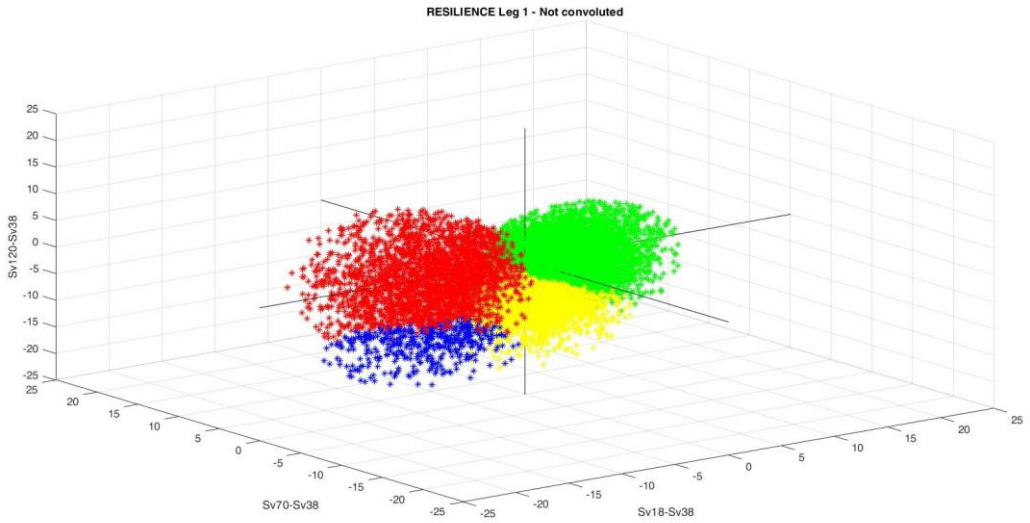
Step (7) Plots to visualise the classified echo-integrated cells of the whole dataset

7.1 Run the script E_1_Check_EchogramSv_classified

From the folder "codes_diff", run the script E_1_Check_EchogramSv_classified

The user has to specify the path to the .mat file generated in Step 6.2.

The following figure is generated:



Three-dimensional ordination plot representing the four echo-classes of the whole acoustic dataset along S_{v18-38} (x-axis), S_{v70-38} (y-axis) and $S_{v120-38}$ (z-axis).

7.2 Run the script E_2_Plot_RGB_one_group

From the folder "codes_diff", run the script E_2_Plot_RGB_one_group

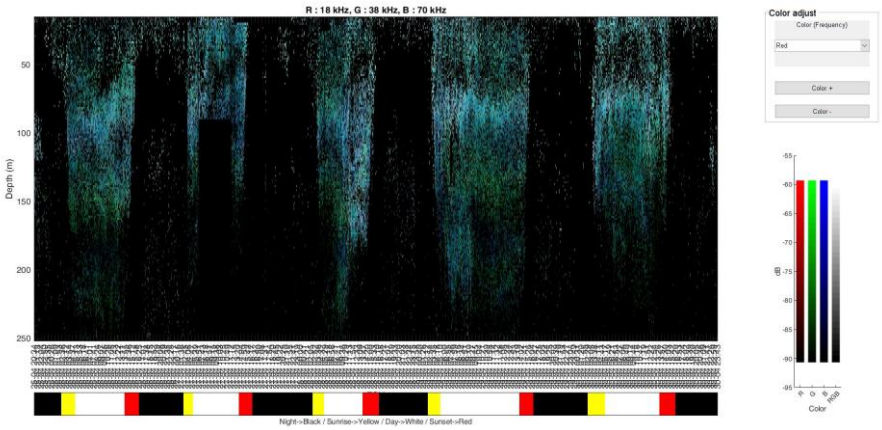
This script generates RGB echograms of each echo-class.

Prior to running the script, the user has to comment lines 21 and 22 (i.e., clear all and close all) of the script D_echocolors_mainprogram from the folder Echogramme_RGB_fromEI. Uncomment once the script E_2 has been run.

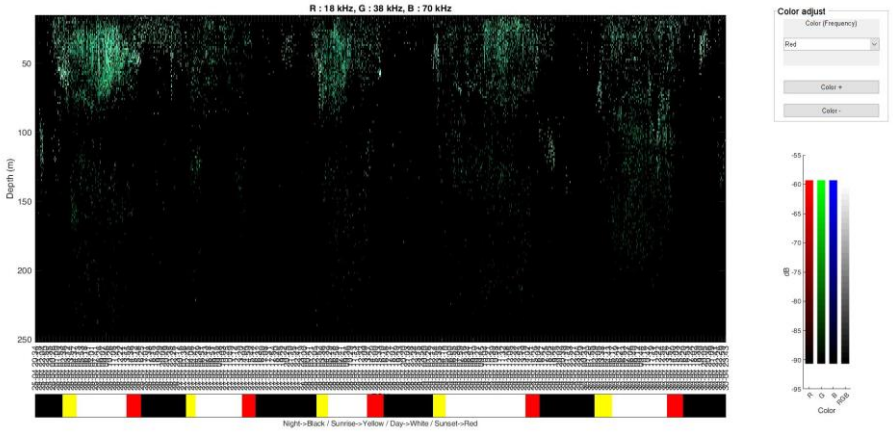
The user will have to specify the path to the echo-integrated file and the .mat file generated in Step 6.2.

The user also has to specify the number of ESUs to be plotted, the number of echo-classes to plot, the frequencies to be plotted with the RGB color scale, the low threshold, the maximum depth on the RGB echogram, the ranges and whether to do a convolution or not.

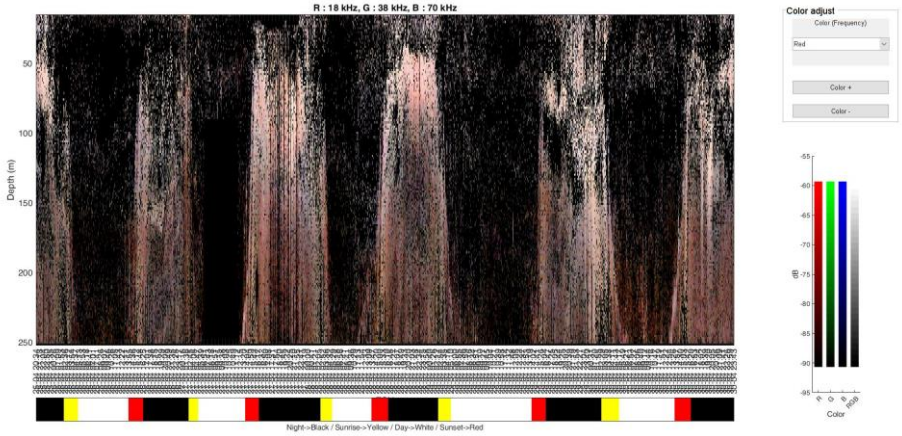
The following plots are generated:



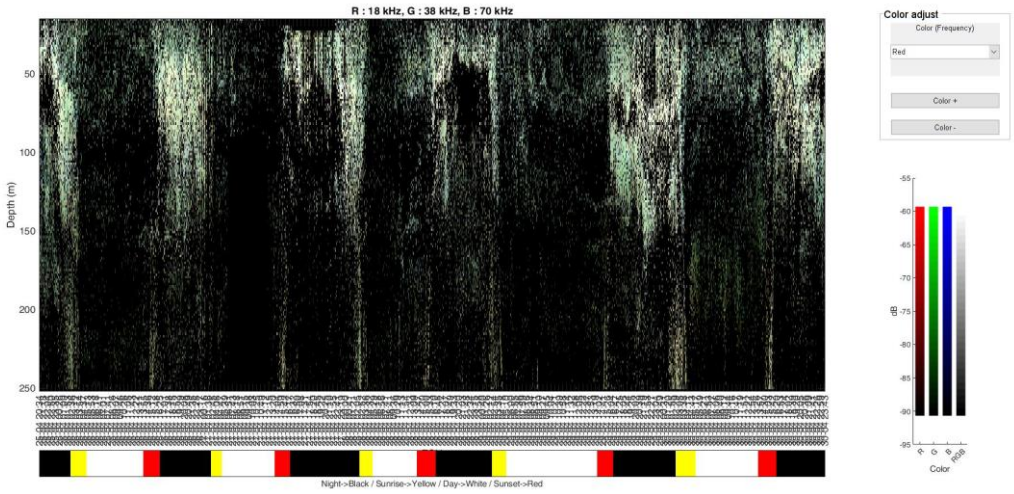
RGB echogram of echo-class 1, with depth (m) on the y axis, and time on the x-axis.



RGB echogram of echo-class 2, with depth (m) on the y axis, and time on the x-axis.



RGB echogram of echo-class 3, with depth (m) on the y axis, and time on the x-axis.



RGB echogram of echo-class 4, with depth (m) on the y axis, and time on the x-axis.

7.3 Run the script F_create_EI_classified_for_Matecho

From the folder "codes_diff", run the script F_create_EI_classified_for_Matecho

This script creates a .mat file which can be opened in Matecho. It merges the echo-integrated file of the acoustic dataset with results of the classification.

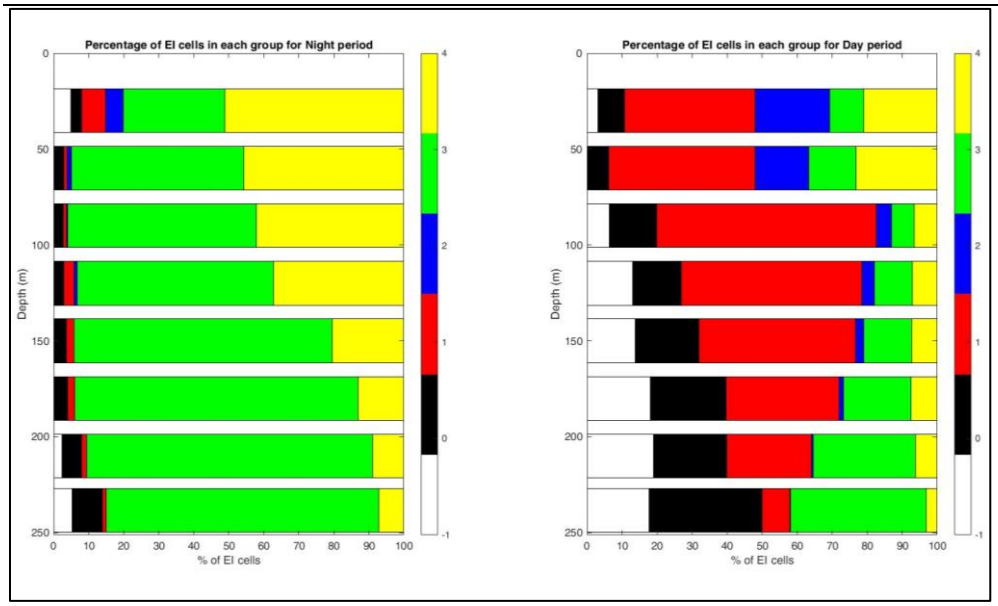
The user has to specify the path to the echo-integrated file and the .mat file generated from Step 6.2

7.4 Run the script G_Plot_vertical_classif

From the folder "codes_diff", run the script G_Plot_vertical_classif

The user has to specify the colors of each echo-class, the number of samples in each vertical bin, the maximum depth, the path to the .mat file generated from Step 6.2 and the echo-integrated file.

The following plot is generated:



Day and night distributions of each-echoclass with depth (m). Unclassified cells are colored black and low S_v cells are colored white.