DOI: 10.71568/daga2026.040

# Developing Data Labelling Standards for Ship Recordings

Viktoriia Boichenko[1], Thierry Maquil[2], Carsten Zerbs[2], Andreas Müller[2], Patrick Nowak[3], Delf Sachau[3], Stefan Schael[4], Anton Homm[5], Gerhard Schmidt[1]

[1] *Christian-Albrechts-Universität zu Kiel, Kiel, Germany*

[2] *Müller-BBM Industry Solutions GmbH, Hamburg, Germany*

[3] *Helmut-Schmidt University, Hamburg, Germany*

[4] *Wehrtechnische Dienststelle 71, Eckenförde, Germany*

*Email: {vib, gus}@tf.uni-kiel.de*[1], *{thierry.maquil, carsten.zerbs, andreas.mueller}@mbbm-ind.com*[2], *{patrick.nowak, delf.sachau}@hsu-hh.de*[3], *{StefanSchael}@bundeswehr.org*[4], *{HA_Hydroakustik}@t-online.de*[5]

## Introduction

The use of different labeling schemes for ship datasets remains largely unexplored, although data labeling is a key preprocessing step for specialized machine learning models. This work investigates how user-defined label schemes can be optimally mapped onto ship data. Despite the availability of several public ship noise datasets, there is no consensus on how to map labels efficiently or how to compare datasets that use different schemes.

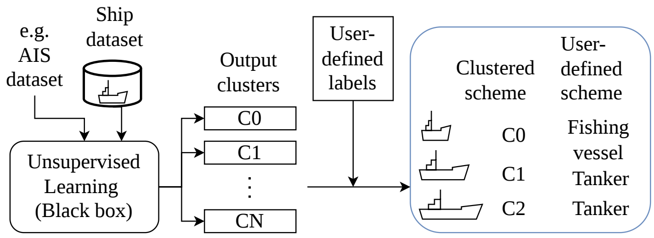



**Figure 1:** A diagram of the proposed data labeling approach with cluster-to-label mapping.

The way to solve this problem can be split into a couple of steps: preprocessing of the data, obtaining an optimal number of clusters, clustering the dataset, and mapping clusters to the chosen labeling scheme. As a result, such a sequence serves as the guide for obtaining an optimal solution during each step of the processing. The steps visualisations is depicted in Figure 1.

## Data Labelling Approaches

Various ship features can be used for labeling ship data. Common labeling strategies for ship classification can be grouped as follows:

- Geometrical characteristics: dimension, propulsion system;
- User-defined: present/absent ship; quiet, noisy, or loud ship; military or civilian ship;
- Acoustic data: low and high frequency content; environmental data;

For example, knowledge of the current weather state can improve acoustic preprocessing, such as denoising, and guide the weighting of input features. Regarding geometrical features, ship length and beam are fundamental characteristics that form a baseline for ship classification. The dataset ShipsEar, published by D. Santos-Domínguez et al. [1], defines four ship classes and characterises them by the size of the vessel (fishing boat, motorboat, passenger ship, and ocean liner). Another dataset, DeepShip, published by M. Irfan et al. [2], classifies into the other four ship types based on AIS data descriptions (oil tanker, tug, passenger ship, and cargo ship).

The current AIS framework for assigning ship types does not always differentiate vessel classes reliably. In particular, AIS types do not clearly separate vessels with similar size and speed. For example, containerships and cruise vessels are not separated in such classification. However, distinguishing between a bulker and a container ship remains challenging when their velocities are similar (e.g., $V < 16$ kn); in such cases, additional characteristics or alternative features are required.

## Clustering Methods

Unsupervised learning is a common tool when useful patterns or characteristics need to be discovered, but no labels are given [4]. The clustering analysis is one of the unsupervised learning methods to discover different prototypes or clusters of features. Among commonly used methods are codebooks that use K-means along with the LGB algorithm and the silhouette score to identify the optimal number of clusters.

## Codebook

Codebooks are a clustering approach to identify patterns in clusters of the input data, to which codewords are calculated and assigned to groups of data points with the help of K-means.

## K-means

Unsupervised learning methods for grouping features include clustering algorithms such as K-means. It estimates the locations of a predefined number of centroids in the dataset. Its goal is to partition the data into $K$ clusters by minimizing the distance between data points and their assigned centroids. Initialization is performed by selecting $K$ distinct training vectors as initial codebook vectors. The cost function can be defined using Equation 1:

$$\sum_{n=0}^{N-1} \min_{i=0,K-1} d(\boldsymbol{x}(n), \boldsymbol{c}_i) \rightarrow \min \tag{1}$$

where $N$ is number of elements, $K$ is number of clusters, $d$ is the distance between a feature vector $\boldsymbol{x}$ and a codebook

**Table 1:** A view of the collected physical characteristics of the ship from AIS data in the Baltic Sea.

| AIS Code | AIS Type Decoded | Length (m) | Beam (m) | DWT (t) | Engine Model | MCR (RPM) | Power (kW) |
|---|---|---|---|---|---|---|---|
| 72 | Bulk carrier | 90 | 15 | 18,900 | mtu 16V 2000 M86 | 2,450 | 1,630 |
| 32 | Tug | 24 | 9 | 3,024 | mtu 8V 4000 M54R | 1,600 | 746 |
| 71 | Container ship | 180 | 32 | 122,400 | mtu 20V 8000 M91L | 1,150 | 10,000 |
| 32 | Tug | 33 | 14 | 6,468 | mtu 16V 2000 M96 | 2,450 | 1,790 |
| 71 | Container ship | 183 | 31 | 120,551 | mtu 16V 4000 M73 | 1,970 | 2,560 |

**Table 2:** A view of the collected physical characteristics of the ship from the ShipsEar dataset.

| Vigo ID | AIS Type Decoded | Length (m) | Beam (m) | DWT (t) | Engine Model | MCR (RPM) | Power (kW) |
|---|---|---|---|---|---|---|---|
| 6 | Passenger ship | 27 | 10 | 45 | MTU 12V 2000 M72 | 2250 | 1800 |
| 7 | Passenger ship | 27 | 10 | 47 | MTU 12V 2000 M72 | 2250 | 1800 |
| 8 | Passenger ship | 19 | 6 | 30 | Caterpillar C18 | 2100 | 746 |
| 9 | Passenger ship | 19 | 6 | 30 | Caterpillar C18 | 2100 | 746 |
| 10 | Passenger ship | 19 | 6 | 30 | Caterpillar C18 | 2100 | 746 |

vector $\boldsymbol{c}$.

## Silhouette score

The silhouette score measures how well each data point fits within its assigned cluster and how well-separated it is from other clusters. It is used in finding an optimal number of clusters. The value is calculated with Equation 2:

$$\frac{b_i - a_i}{\max(a_i, b_i)} \tag{2}$$

where $i$ is a data point, $a_i$ is the intra-cluster distance, $b_i$ is the nearest cluster distance.

## LBG algorithm

Another way to find the optimal number of clusters in the dataset is using the LBG algorithm, which stands for the names of the inventors of the algorithm: Linde, Buzo, and Gray. At initialization, the start codebook consists of only one entry, which is chosen as the mean over all training vectors. After that, follow the iteration steps: increasing the codebook size, classification, codebook correction, and termination condition.

## Mapping function

The mapping between clusters and labels can be classified as the assignment problem. The solution to it lies in the discrete optimisation algorithms, and one of the most famous ones is the Hungarian algorithm or Kuhn-Munkres algorithm [3]. The main principle is to minimise the cost function of the sum of the product of the matrix $X$, sequences of the mapping, and cost matrix $C$, value of each element being assigned to a new element (Equation 3):

$$\min \sum_i \sum_j C_{i,j} X_{i,j} \tag{3}$$

where $C_{i,j}$ is a cost matrix, $i$ is an element of the first set (e.g. clusters), $j$ is an element of the second set (e.g. labels), $X$ is a boolean matrix with $X[i,j] = 1$ if row $i$ is assigned to column $j$. In order to find an optimal zeros assignment to find the desired mapping sequence, a minimum value from each row and column needs to be substracted iteratively. The found zeros correspond to the cluster-to-label mapping.

## Dataset of Ship Characteristics

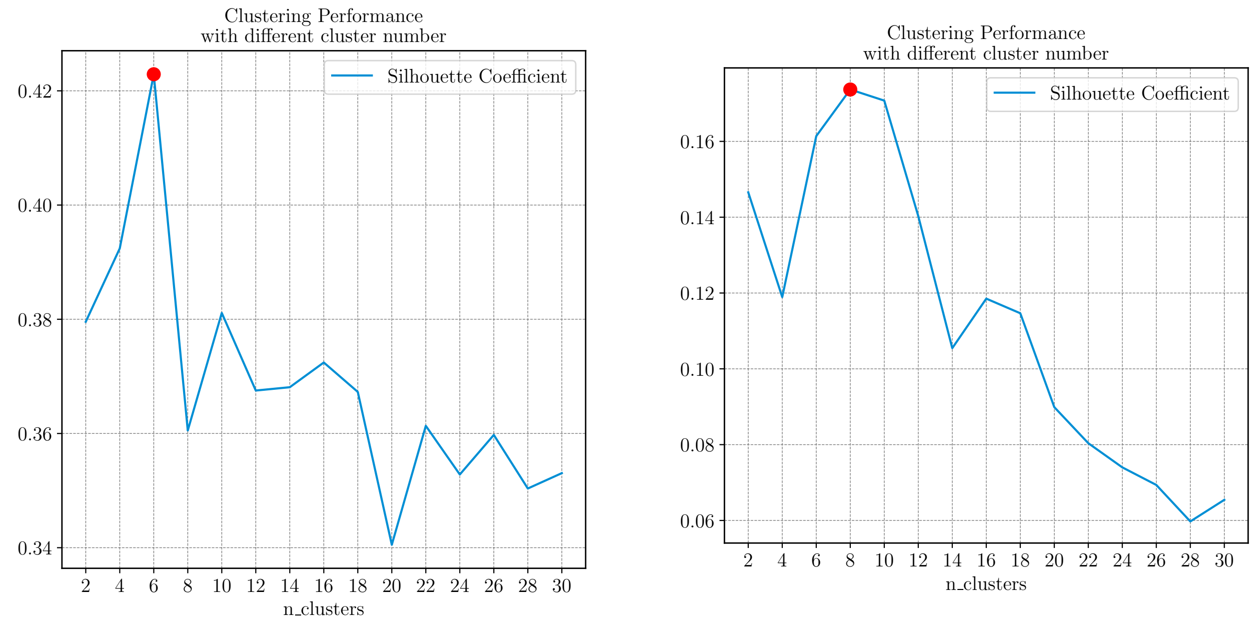
The data were collected from publicly available ship tracking systems and technical datasheets. The resulting database contains AIS data from 3846 ships observed in the area of the Baltic Sea between Germany and Denmark (Table 1). It includes static ship characteristics such as its propulsion system features, geometrical properties, and assigned ship type. The collected data entries were in the following columns: total activity count, total AIS signals, engine RPM, engine number of cylinders, auxiliary engine number of strokes, number of propeller blades, propeller count, and firing frequency in Hz.

The same characteristics are gathered for the publicly available dataset ShipsEar. The view of the first entries of the first entries of the extended ShipsEar metadata is presented in Table 2. Additionally, acoustic features extracted from hydrophone recordings are combined with the physical ship characteristics. A segment of 5 seconds is analysed to obtain values of the spectral centroid, rolloff, bandwidth, zero crossing rate, RMS energy, and 20 mel-frequency cepstral coefficients per second. Therefore, the extended ShipsEar dataset comprises 42 ship entries with 202 features. Multiple recordings are available for most vessels. For the clustering analysis, columns such as MMSI and IMO were excluded to avoid automatically combining ships into the same clusters solely based on their IDs.

## Results & Discussion

The clustering analysis was focused on geometrical properties of the ship; thus, propulsion system-related characteristics were excluded from the analysis. The chosen features and column names were length, beam, deadweight tonnage (DWT), maximum continuous rating (MCR RPM), and engine power in kW.

The silhouette coefficient was evaluated for different codebook sizes to estimate the optimal number of clusters. Figure 2a shows the K-means performance; the maximum value of 0.42 occurs at $K = 6$, whereas all other values remain below 0.4. Regarding the ShisEar




**(a)** Baltic Sea dataset, $K = 6$ is optimal number of clusters.

**(b)** ShipsEar extended dataset, $K = 8$ is optimal number of clusters.

**Figure 2:** A clustering performance of Baltic Sea and ShipsEar extended datasets.

extended dataset, the optimal number was chosen to be 8 clusters, according to the results with the silhouette coefficient calculation in Figure 2b.

The mapping results in Figure 3 indicate that the datasets do not yield separate clusters for small vessels, because the data are strongly imbalanced toward large ship types. Nevertheless, Figure 3b shows that the ShipsEar dataset, when augmented with spectral features, still allows meaningful cluster-to-label mappings despite the limited number of ship entries.

Table 3 present the accuracy of mapping clusters to labels for AIS labeling scheme. As it was mentioned earlier, the number of clusters within the two datasets is smaller than the number of labels from available AIS ship types. Thus, some labels have not been included in generated clusters, and corresponding labels with percentages are filled with zeros, such as „Bulk carrier“, „Towing vessel“, and „Other cargo“. Besides, it can be seen at Table 3 that „Cargo ship“, „Tanker“, „Container ship“, and „Tug“ have higher accuracy of mapping than other labels because these ship types are dominating within the dataset.

**Table 3:** AIS Ship Type Labeling Scheme: Cluster-to-label Accuracy.

| AIS Ship Type Labels | Baltic Sea Data Accuracy | ShipsEar Data Accuracy |
|---|---|---|
| Bulk carrier | 0.000 | 0.000 |
| Cargo ship | 0.523 | 0.750 |
| Container ship | 0.489 | 0.000 |
| Fishing vessel | 0.091 | 0.000 |
| Other cargo | 0.000 | 0.000 |
| Passenger ship | 0.160 | 0.500 |
| Tanker | 0.375 | 0.000 |
| Towing vessel | 0.000 | 0.000 |
| Tug | 0.355 | 1.000 |
| **Mean** | 0.221 | 0.25 |

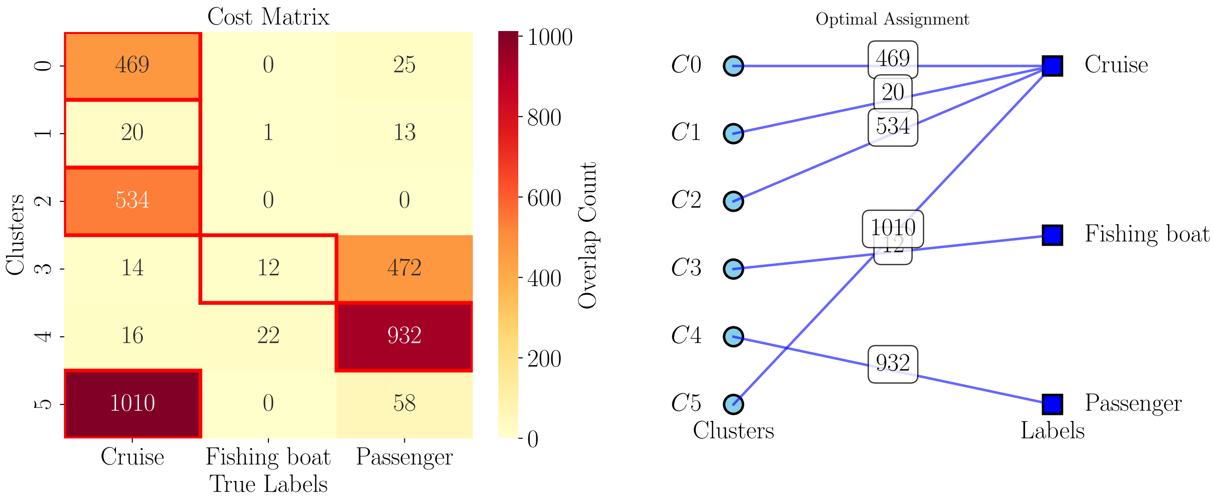
**Table 4:** ShipsEar Labeling Scheme: Cluster-to-label Accuracy.

| ShipsEar Labels | Baltic Sea Data Accuracy | ShipsEar Data Accuracy |
|---|---|---|
| Cruise | 0.985 | 0.571 |
| Fishing boat | 0.343 | 0.783 |
| Passenger | 0.621 | 0.917 |
| **Mean** | 0.650 | 0.757 |

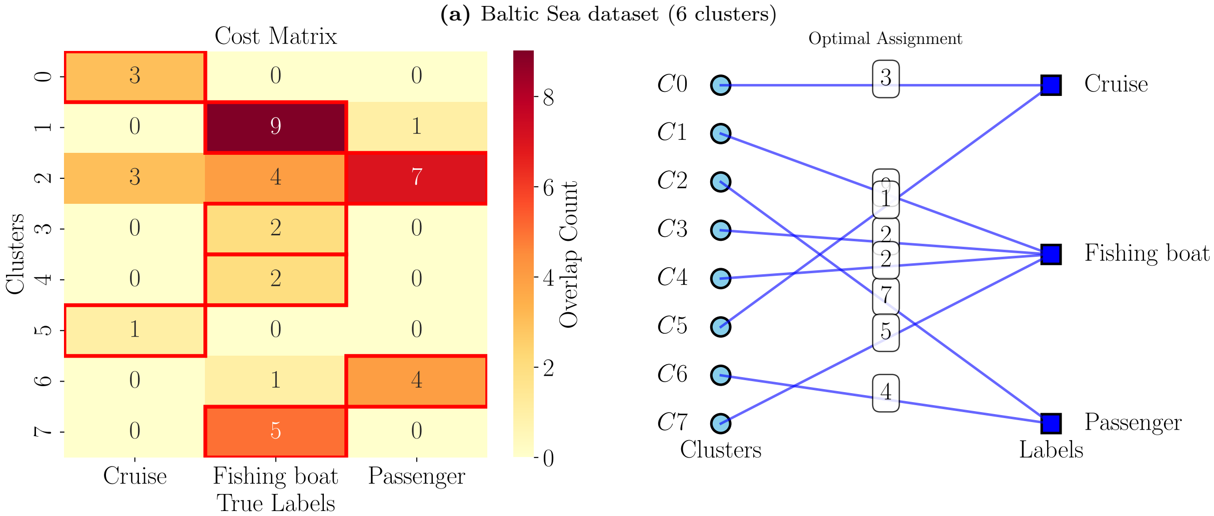
Table 4 compares mapping accuracy based on the ShipsEar labeling scheme that is based on the length of common ship groups. The available ship data contains three different labels that were initially mapped onto generated clusters in this dataset. It can be observed that the average accuracy of both datasets is higher than that of the AIS labeling scheme, Table 3. Consequently, the ShipsEar labeling scheme, which is based on ship length, achieves higher mapping accuracy. For these datasets, a more generalized human-defined labeling overlaps better with machine-generated clusters, whereas AIS ship types, which are mainly function-based, align less well with the mathematically-defined cluster structure.

## Conclusion and Outlook

Applying the Hungarian algorithm demonstrates that human labeling of ships differs markedly from mathematically based labeling, and the overlap between computed and human-defined schemes is limited. Moreover, the pseudo-optimal scheme does not reflect market requirements well. Nevertheless, geometric and spectral features provide a useful baseline for mapping between datasets with different labeling schemes, and spectral features in particular can serve as a foundation for cluster-to-label mapping when no other labeling information is available.




(a) Baltic Sea dataset (6 clusters)




(b) ShipsEar dataset (8 clusters)

**Figure 3:** Cost matrix and assignment of clusters to ShipsEar-type labels. (a) Baltic Sea dataset with 6 clusters. (b) ShipsEar dataset with 8 clusters.

Future work may explore additional methods at each step of the mapping approach to further improve cluster-to-label accuracy. One direction is to incorporate more ship characteristics, including detailed engine and propeller data and acoustic features derived from LOFAR and DEMON analysis. Another is to apply more advanced clustering methods, such as hierarchical clustering, Gaussian mixture models, or DBSCAN, and to refine subclusters for alternative label mappings. Finally, scenarios with varying amounts of labeling information, from partially labeled data to the ultimate goal of individual ship classification, should be investigated.

## Acknowledgement

This work was supported by the WTD 71 Wehrtechnische Dienststelle für Schiffe und Marinewaffen, Maritime Technologie und Forschung.

## References

[1] D. Santos-Domínguez, S. Torres-Guijarro, A. Cardenal-López, and A. Pena-Gimenez, "ShipsEar: An underwater vessel noise database," Applied Acoustics, vol. 113, pp. 64–69, Dec. 2016, doi: 10.1016/j.apacoust.2016.06.008.

[2] M. Irfan, Z. Jiangbin, S. Ali, M. Iqbal, Z. Masood, and U. Hamid, "DeepShip: An underwater acoustic benchmark dataset and a separable convolution-based autoencoder for classification," Expert Systems with Applications, vol. 183, p. 115270, Nov. 2021, doi: 10.1016/j.eswa.2021.115270.

[3] Munkres, J. (1957). Algorithms for the Assignment and Transportation Problems. Journal of the Society for Industrial and Applied Mathematics, 5(1), 32 - 38

[4] M. J. Bianco et al., "Machine learning in acoustics: Theory and applications," The Journal of the Acoustical Society of America, vol. 146, no. 5, pp. 3590–3628, Nov. 2019, doi: 10.1121/1.5133944.