WMU Journal of Maritime Affairs (2021) 20:377–400
https://doi.org/10.1007/s13437-021-00241-3

IAMU SECTION ARTICLE

K-means clustering for SAT-AIS data analysis

Marta Mieczyńska[1] · Ireneusz Czarnowski[2]

Received: 19 December 2019 / Accepted: 16 May 2021 / Published online: 5 July 2021

Abstract

The paper deals with a problem of automatic identification system (AIS) data analysis, especially eliminating the impact of AIS packet collision and detecting existing outliers in AIS data. To solve this problem, a clustering-based approach is proposed. AIS is a system that supports the exchange of information between vessels about their trajectories, e.g. position, speed or course. However, SAT-AIS, which enables the system to work on a global scale, struggles against packet collisions due to the fact that the satellite, which receives AIS data from ships, has a field of view that covers multiple areas that are not synchronized among themselves. As a result, the received data is difficult to process by AIS receivers, because most of the messages have a character of noise. In this paper, results of a computational experiment using *k*-means algorithm for packet recovery and for dealing with noise have been presented. The outcome proves that a clustering-based approach could be used as an initial step in AIS packet reconstruction, when the original data is incorrect .

Keywords *K*-means · Clustering · SAT-AIS · Data analysis · Maritime data analytics

1 Introduction

An automatic identification system (AIS) is an automatic tracking system that has been developed according to the International Maritime Organisation (IMO) regulations. The aim of creating such system was to develop a technology that would

✉ Ireneusz Czarnowski
i.czarnowski@umg.edu.pl

Marta Mieczyńska
m.mieczynska@we.umg.edu.pl

[1] Department of Marine Telecommunications, Gdynia Maritime University, Morska 81-87, 81-225 Gdynia, Poland

[2] Department of Information Systems, Gdynia Maritime University, Morska 81-87, 81-225 Gdynia, Poland

provide information about ships, including their unique identifier, type, position, speed, course and current state, to other vessels and shore stations automatically (International Maritime Organisation (IMO) 2019). The dynamic information is obtained from the ship's navigational sensors such as its global navigation satellite system (GNSS) receiver and gyrocompass. On the other hand, static information (e.g. ship's identifier MMSI) is permanently programmed on the ship's equipment. Both of them are formed into binary format to create AIS messages and transmitted regularly using dedicated transponders. The reception of AIS messages is performed by either ships or land-based systems (e.g. vessel traffic systems) (exactEarth 2015).

Most of AIS messages are transmitted on a regular basis. For instance, messages containing dynamic information are exchanged every 2 to 180 s (European Space Agency 2019). Hence, during a specific recording time period a significant amount of data can be received. To process this huge dataset and actually derive some meaningful information from it, the use of modern, advanced technology is required (Czarnowski 2019). Machine learning methods might be one of the possible approaches here, since it provides algorithms that cope with, among others, finding a pattern in a huge dataset (Mieczyńska and Czarnowski 2019).

Nowadays, a need for carrying out the analysis of AIS data appears more and more often. The reason is that such functionality is utilized by various applications. The importance of AIS data analysis is crucial especially for maritime industry since the usage of data analysis may lead to improved performance of monitoring and optimization of maritime processes. Examples of those applications might be related to the maritime safety. For instance, the usage of a system that would predict the vessels' movement may result in an early collision avoidance between ships (Zhang et al. 2015). The same system may be indispensable when it comes to predict a vessel's location (Liang et al. 2019) in emergency situations, when the connection with that ship is lost. Another example of analysis of both real-time and historical data is an identification of abnormal vessels' activity that may lead to the detection of an act of piracy (Lane et al. 2010). On the other hand, AIS data might also be useful in a research of industrial usage in the form of maritime traffic analysis — prediction of the load in seaports and its optimization (Millefiori et al. 2016) or route planning (He et al. 2019).

The original, terrestrial AIS utilizes two VHF (very high frequency) frequencies (161.975 MHz and 162.025 MHz) with the bandwidth of 25 kHz. To manage the access to the wireless medium by multiple AIS transponders, the TDMA (time division multiple access) method is used. A single device is allowed to transmit only during a pre-determined period of time (called slot). More specifically, each AIS transponder must preannounce the time slots it wants to use (this technique is called self-organizing TDMA (SOTDMA)). Time slots filled with information from various devices form a time frame. Nine 1-min-long time frames (consisting of 2250 26.6-ms time slots per radio frequency channel) are then grouped into a communication cell. Within such a communication cell, slot selection is organized randomly. Devices choose their time slots so they can transmit in a pre-assumed rate (which depends on such factors as the speed of the vessel or its heading). If the AIS transponder changes its slot assignment, it must transmit its new assignment and timeout for that assignment.

Although original (terrestrial) AIS itself has many advantages and potential applications, there are some drawbacks of this system as well. As mentioned before, it has been originally developed to provide information about nearby vessels that could be used to prevent collisions of vessels. The information about ships' movement (course, position, speed) is exchanged between them and shore stations regularly, so they are able to recognize other vessels that may appear on their paths. However, the main limitation of this communication is its range. Due to the Earth's curvature, the horizontal range of terrestrial AIS' visibility is about 74 km (40 nautical miles) from shore (European Space Agency 2019). Consequently, this indicates that the original AIS is a system working on a local scale, i.e. on a ship-to-ship basis or around coastal zones only.

To overcome such a problem and enable AIS to work on a global scale, a SAT-AIS system has been proposed (European Space Agency 2019). In general, SAT-AIS utilizes satellites (e.g. AAUSAT3) on low-earth-orbit to increase the range of transmission. Messages sent by ships are recorded by a satellite (which has a broader range of view due to its altitude) and then transmitted to ground stations for further processing and distribution (Wawrzaszek et al. 2019). Although it seems to solve many of terrestrial AIS' restrictions, SAT-AIS also struggles against its own limitations. The main problem of this satellite system is packet collision. The problem of package collision is an example of a normal wireless system behaviour due to its technological restrictions regarding synchronization between AIS' transponders (Swetha et al. 2018). When multiple vessels start or stop transmitting (thus assigning slots) in a communication cell, other devices may receive information from various cells (e.g. from both terrestrial AIS base station and SAT-AIS satellite), which are not organized within themselves — that is why slot (and packet) collisions appear (exactEarth 2015).

The packet collision means that the AIS message cannot be correctly decoded from signals recorded by a satellite. Thus, the AIS message is incomplete or consists of incorrect values (for example, includes incorrect information about ship's position or speed). In other words, it means that in a chain of received information, a noise exists and the AIS message can be lost or refused. From the perspective of ship monitoring, it means that the ship's trajectory and ship's behaviour cannot be monitored in some period of time and the ship's trajectory have gaps.

To make AIS-based systems working efficiently, not only the high-efficient algorithms should be implemented, but also the dataset needs to meet some quality requirements. Several ways of handling with problem of packet collision can be encountered in the literature. Some of them are based on data analysis and signal reconstruction (Prevost et al. 2012; Seta et al. 2016). The approach proposed in this article is focused on the latter issue. Therefore, machine learning algorithms can be applied not only for deriving information from data but also for recovering the incomplete dataset by learning the patterns between delivered features and trying to predict the missing values from AIS messages or those assumed to be incorrect.

In this paper, a cluster-based approach is presented as a one of machine learning techniques for AIS data analysis. The proposed approach is going to be the first stage of further detection of the abnormal AIS messages and prediction of their correct form.

The initial results of an implementation of cluster-based approach in general way for AIS data analysis and with respect to the AIS data reconstruction have been presented at the International Association of Maritime Universities Conference (IAMUC) in 2019 (Czarnowski 2019). The k-means algorithm implemented for vessel trajectory reconstructions has also been discussed in Mieczyńska and Czarnowski (2019). This article consists of an extension of the research results presented before in the mentioned papers. In this paper, the AIS data recovery using k-means algorithm is considered again, but the discussion is wider and more results are presented. The main contribution of the paper is the evaluation (through computational experiment) of the clustering approach based on k-means for AIS data recovering and for deciding on a character of detected noise (outliers) within AIS data stream. The performance of the proposed clustering approach has been evaluated using a benchmark dataset obtained from the original AIS system.

The paper is organized as follows: the SAT-AIS packet collision is shortly discussed in Section 2. Section 3 contains a description of the proposed approach. Section 4 provides details on SAT-AIS dataset. Details on the computational experiment setup and discussion on experiment results are included in Section 5. Finally, the last section contains conclusions and suggestions for future research.

2 SAT-AIS packet collision and problem formulation

The main problem of the satellite-based AIS system is the aforementioned packet collision. Packet collisions occur when a satellite receives two or more messages at the same time. The synchronized allocation of AIS message time slots is guaranteed only within a limited area (Wawrzaszek et al. 2019), called a cell or a terrestrial AIS service area. Satellite's field of view (FOV) covers several such areas (see Fig. 1), all of them not being synchronized between themselves and sending messages to the satellite simultaneously. As a result, a package collision is observed. The satellite receives signals from different ships that are characterized by various amplitudes, time delays and Doppler frequency shifts, which make it difficult to distinguish the corresponding messages.

Ultimately, not much of the received signal can be useful. Only partially received data packets can be correctly decoded into AIS message which is a result of the lack of synchronization and lack of the possibility to identify the beginning and ending of a packet. The lack of synchronization and the overlapping of signals from several ships can also result in mistakes (errors) within decoded AIS message. Those mistaken fields with respect to all received messages can be recognized as outliers.

To eliminate the aforementioned issue, several techniques have been already proposed in the literature. Most of them concentrate on packet recovery, either through recognition of the received sequence's shape, decoding methods using the Viterbi algorithm (Prevost et al. 2012), blind source separation algorithm (Swetha et al. 2018) or statistical estimation (Seta et al. 2016).

However, the problem of packet collision and elimination of its effects is still open. Today, the advanced machine learning methods create a possibility to solve this problem using slightly different approach than technical or statistical analysis.

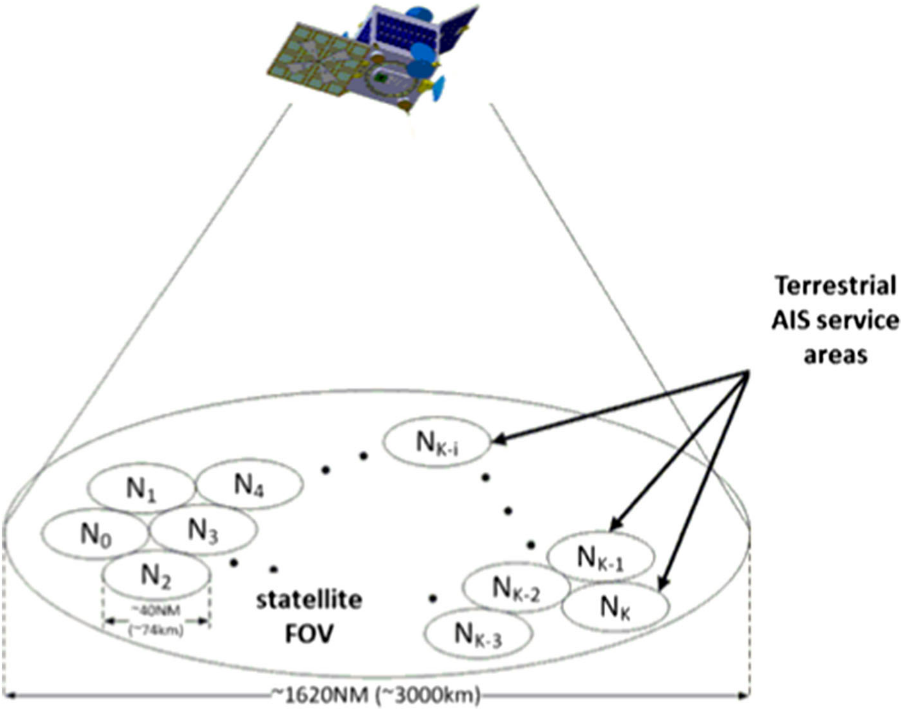

Fig. 1 AIS service areas in the satellite field of view (Swetha et al. 2018)

In the next section of this paper, the cluster-based approach for reconstruction of AIS messages and elimination of incorrect AIS data is presented.

3 Cluster analysis for SAT-AIS data

3.1 Unsupervised learning and clustering

In general, cluster analysis is a method of distinguishing groups called clusters in a set of available data. It is assumed that these groups are disjoint, which means that data belonging to different clusters differ between themselves much more than the data belonging to the same cluster. From a practical point of view, when data is defined as a set of objects or instances described by a set of attributes (features), a result of cluster analysis are groups of instances where each instance belongs only to one cluster.

Thus, it can be concluded that the role of cluster analysis is to uncover a certain kind of natural structure in the dataset. For performing this task, a certain measure of similarity or dissimilarity is usually defined (Wierzchon and Klopotek 2015).

Cluster analysis or clustering belongs to a class of unsupervised learning. Unsupervised learning is a machine learning technique where learning from data is carried out without supervision. In such case, the machine learning process is based on using information that is neither classified nor labelled and allowing the algorithm to act without guidance.

The crucial question formulated in a domain of unsupervised learning concerns identification of similarities among the data and determining the number of clusters in a dataset.

Among the set of different algorithms dedicated for solving the task of cluster analysis, there is a k-means algorithm. K-means is an iterative and so-called partitional clustering algorithm. The k-means algorithm divides the given data into k clusters. Each cluster has a cluster centre called centroid and the clustering process is focused on these centroids. k defines the number of clusters and must be specified by the user. Therefore, k-means is not an algorithm for determining the potential similarity between the data, thus for determining numbers of these similarities, but k-means is a relatively effective algorithm for partitioning the data into clusters, when the number of clusters is defined beforehand. In general, k-means allow to find, if not optimal, the near optimal partition of data in reasonable amount of computation time. The pseudocode of the k-means algorithm is presented as Algorithm 1.

Algorithm 1 The k-means algorithm.

```
Input: $D$ — set of data; $k$ — predefined number of clusters
Choose randomly $k$ data points (seeds) to be the initial centroids
while convergence criterion is not met do
    Assign each data point to the closest centroid
    Re-compute the centroids using the current cluster memberships
end while
```

An alternative way for machine learning is based on the concept of semi-supervised learning, originally introduced to eliminate a basic disadvantage of any supervised learning algorithm, which generally needs labelled data. Semi-supervised learning uses both tagged and untagged data to find a model.

3.2 The proposed approach

In this paper, k-means algorithm is proposed to analyse AIS data when there are outliers within the data. We assume that those outliers are a result of packet collision in a satellite-based AIS system. The proposed approach is also based on semi-supervised learning in such a meaning that the decision on qualification of outliers is based on clustering results.

We assume that it is well known how many vessels are within the considered sea area — it can be established based on previously registered information. The stream of data from AIS system for a given period time can be partitioned into number of groups equal to a number of vessels in a considered sea area. Such clustering can be carried out regardless of whether the data contains errors or not. Assuming that the data contains errors or a noise, such clustering requires an approach which can detect similarities or dissimilarities within the data, for example the k-means algorithm.

Thus, when an individual data consists of errors or a noise and available dataset will be divided into a predefined number of clusters, then there is a big probability that the data with errors and a noise will be allocated to clusters including data similar to them. Subsequently, the data with errors can be reconstructed to their correct form, which means that the AIS message can be reconstructed to its useful form. In the considered approach, the correct data is considered as labelled, when the damaged data as unlabelled, so it is a reason for including the proposed approach to a semi-supervised learning paradigm.

The pseudocode of the proposed k-means clustering for SAT-AIS data analysis is shown as Algorithm 2.

Algorithm 2 The k-means clustering for SAT-AIS data analysis.

```
Input: $D$ — set of data; $k$ — predefined number of clusters (equal to the number of
       vessels)
1: Run $k$-means on $D$ and map instances from $D$ into $k$ clusters
2: Let $D_1, \ldots, D_k$ denote the obtained clusters such that $D = \bigcup_{i=1}^{k} D_i$
3: for all $i = 1, \ldots, k$ do
4:     Identify outliers or damaged feature values in $D_i$ and run their correction
5: end for
```

In the next sections, computational experiment results of a research where k-means algorithm has been used for the considered problem will be presented. Moreover, answers to the questions raised about the legitimacy of the proposed approach have been formulated.

4 SAT-AIS dataset

4.1 AIS message encoding

To predict the trajectory of the ship and, subsequently, to reconstruct lost or incorrect AIS messages, it was necessary to obtain a specific dataset. In the AIS system, the information regarding vessels' trajectory is carried by messages of 3 different types: 1, 2 and 3 (International Telecommunications Union 2014). Messages of each of those types are called position reports and share the same data format which is presented in Table 1. The difference between them is that messages type 1 carry scheduled position reports, messages type 2 — assigned position reports, and type 3 — special position reports (International Telecommunications Union 2014).

The AIS data that has been used to conduct this experiment was obtained from messages of types 1–3, each consisted of 168 bits in total, gathered in 3 different datasets:

- The first one was obtained from 22 vessels from the area of Gulf of Gdansk, recorded during a 35-min time slot (during that time 850 messages have been received),

Table 1 Fields of AIS messages types 1–3 (International Telecommunications Union 2014)

Field	Bits	Format
Message ID	1–6	Unsigned integer
Repeat indicator	7–8	Unsigned integer
User ID (MMSI)	9–38	Unsigned integer
Navigational status	39–42	Enumerated (unsigned integer)
Rate of turns	43–50	Signed integer with scale
Speed over ground	51–60	Unsigned integer with scale
Position accuracy	61–62	Boolean
Longitude	62–89	Signed integer with scale
Latitude	90–116	Signed integer with scale
Course over ground	117–128	Unsigned integer with scale
True heading	129–137	Unsigned integer
Time stamp	138–143	Unsigned integer
Special manoeuvre indicator	144–145	Enumerated (unsigned integer)
Other (spare, radio status)	148–168	–

- The second covers the broader area of Baltic Sea (19,999 messages, 387 ships), and
- The third consists of data collected around Gibraltar (also 19,999 messages, 524 ships).

The original data, right after the recording, was in a binary form. Each of the messages had to be then decoded, i.e. some fields (mentioned in the next section) have been transformed into a decimal form (multiplied by some scale in certain cases). The decoding process of a sample AIS message type 1 is shown in Table 2.

Figures 2, 3 and 4 show the visualization of the used data. As mentioned before, the AIS message carries numerous amounts of information; however, it was necessary to select no more than 2 or 3 fields to make the visualization comprehensible. The most intuitive ones seemed to be those related to the location of the vessel since that could present the trajectory of the ship, i.e. how the vessel moved during the recording interval. That is why the plot in Figs. 2 and 3 consists of the longitude on X axis and latitude on Y axis. Furthermore, each trajectory has been marked with a different colour to distinguish datapoints originating from each of the vessels.

4.2 The proposed data model

In order to implement the proposed approach to analyse AIS data, the gathered data which represents vessels' trajectories had to be arranged into an input matrix of features.

One method of expressing the trajectory of the ship in a mathematical way is to create a set of the following vectors (Mieczyńska and Czarnowski 2019):

$$T_i^{t_m} = [x_1, x_2, x_3, \ldots, x_N]_{t_m}^{i}, \tag{1}$$

Table 2 A sample of AIS type 1 message in a binary form and its equivalent in a decimal form

Field	Binary form	Decimal form
Message ID	000001	1
Repeat indicator	00	0
User ID (MMSI)	001111100100011010010000111000	261203000
Navigational status	0000	0
Rate of turns	00000000	0
Speed over ground	0000000000	0
Position accuracy	1	1 (high)
Longitude	0000101010011110001111110101	18,55 (deg)
Latitude	001111100110101100001100000	54,54 (deg)
Course over ground	100101111000	2424
True heading	000011000	24 (deg)
Time stamp	000100	4 (s)
Special manoeuvre indicator	00	0

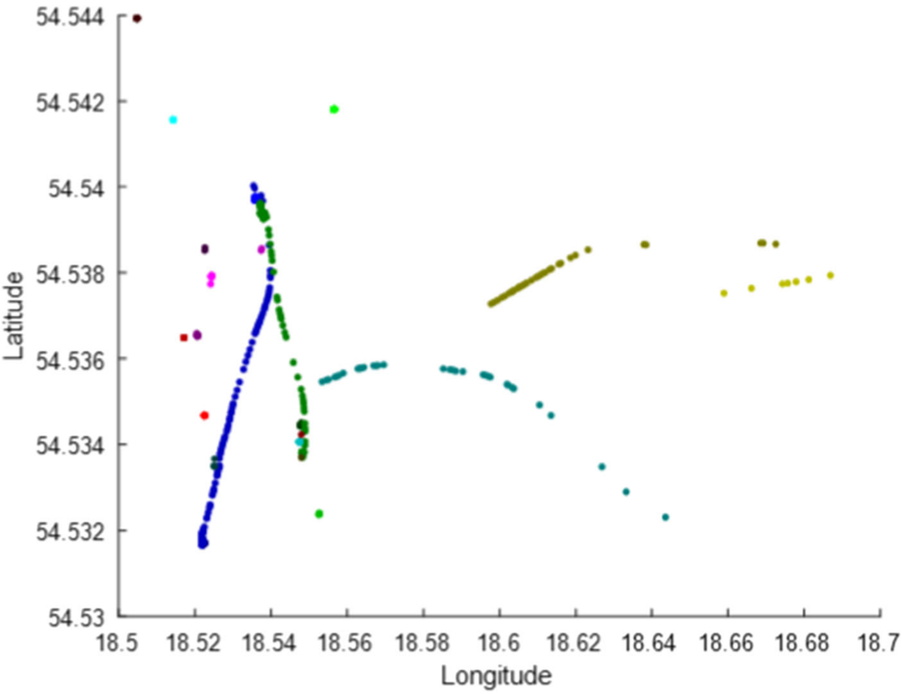

Fig. 2 Recorded AIS data from the first dataset — 22 ships and its representation in the form of trajectories of every ship (each marked with a different colour)

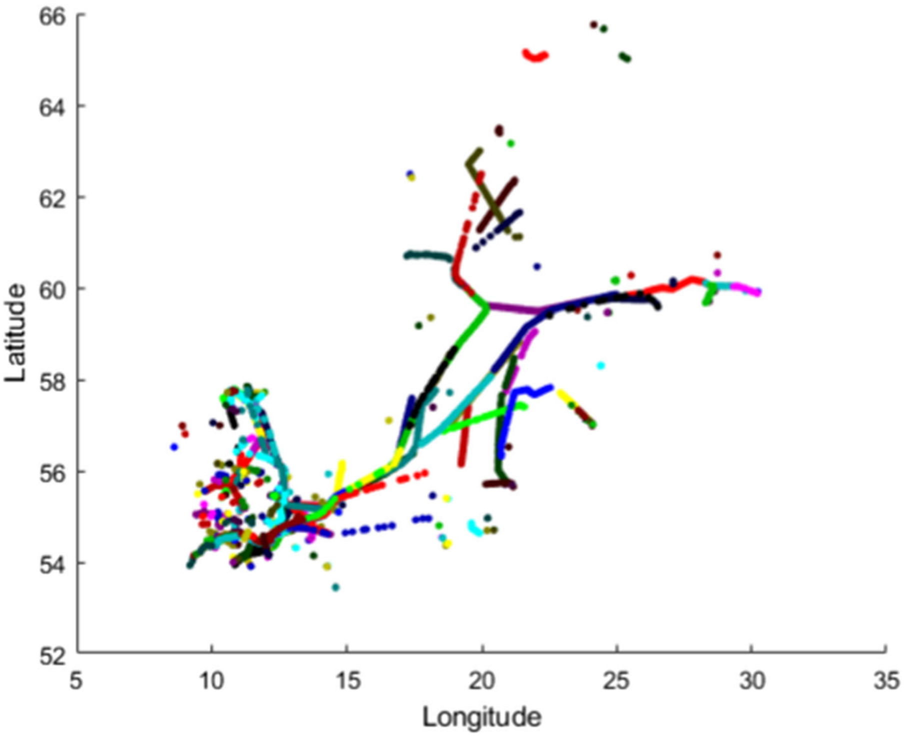

Fig. 3 Recorded AIS data from the second dataset — 387 ships and its representation in the form of trajectories of every ship (each marked with a different colour)

where

- $T_i^{t_m}$ is a trajectory datapoint from time t_m,
- i represents an individual vessel (with a unique MMSI identifier),
- m is the number of points (time steps) that build a trajectory sequence,
- N defines dimension vector of reported AIS dynamic information,
- $x_1, x_2, \ldots, x_N$ are the features derived from AIS messages types 1–3 (see Table 1.).

At this point, an important question arises: which fields of AIS messages types 1–3 should be included in the input matrix as features. After conducting multiple auxiliary tests, the decision has been made to use the following fields: longitude, latitude, navigational status, speed over ground, course over ground, true heading, special manoeuvre indicator, ship identifier and country identifier. The last two features have been obtained from the same message field, since the 3 first digits from MMSI indicate the number unique to each country and the following 6 digits form individual code for each vessel.

Moreover, it is worth mentioning that in machine learning, a special attention is required while using codes/identifiers as features (Brownlee 2017). The algorithm might mistakenly try to interpret the increasing/decreasing code numbers as values

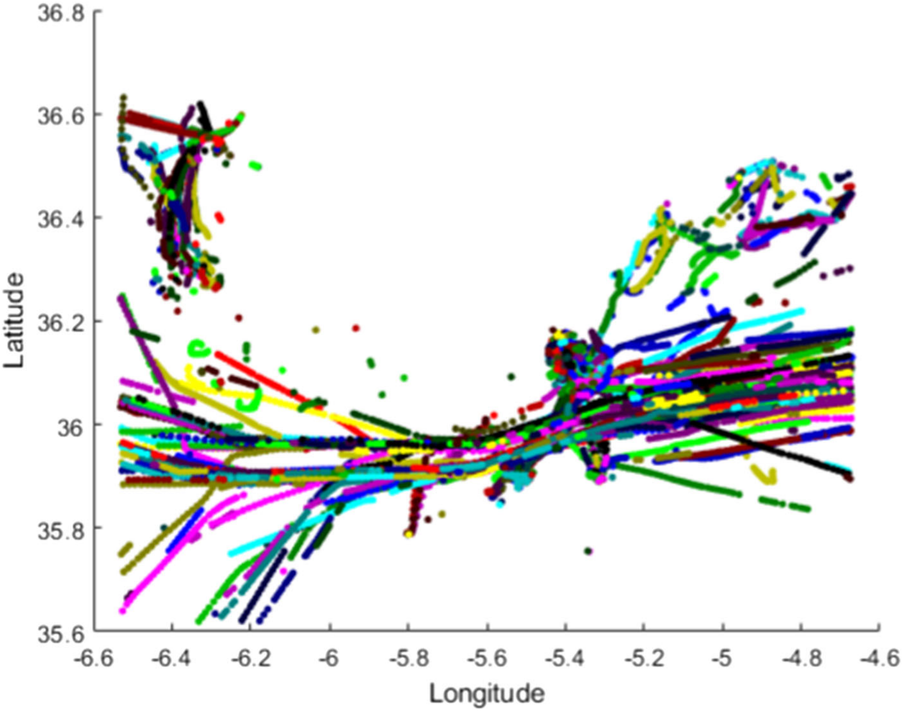

Fig. 4 Recorded AIS data from the third dataset — 524 ships and its representation in the form of trajectories of every ship (each marked with a different colour)

that indicate the increase/decrease of that feature's level (e.g. temperature, speed, brightness). To avoid this behaviour, it is advisable to further encode those identifier features using one hot encoding. With this technique, it is possible to convert each digit of identifier into a set of additional features, whereof each consists of only two binary values: 0 and 1, while 1 can be placed on only one position and other positions are filled with 0s.

With the use of one hot encoding technique, the following features have been converted: navigational status, special manoeuvre indicator, ship identifier and country identifier. Each digit of MMSI has been converted into 10-dimensional vectors filled with the value of 0 and one value of 1 on the kth position, where k is simply that digit. If the identifier is allowed to take only restricted l values (like navigational status, which is carried in 2 bits and can only take values in the range from 0 to 3), then the output vector of one hot encoding is l-dimensional.

According to the aforementioned rules, the established features are set as follows:

– x_1 — longitude,
– x_2 — latitude,
– $x_3 - x_{18}$ — navigational status,
– x_{19} — speed over ground,
– x_{20} — course over ground,

- x_{21} — true heading,
- $x_{22} - x_{25}$ — special manoeuvre indicator,
- $x_{26} - x_{85}$ — ship identifier,
- $x_{86} - x_{115}$ — country identifier,

that gives 115 features in total.

Finally, the entire dataset has been standarized, i.e. all features have been turned into values in the range from -1 to 1: the distribution of each of the features has a mean of 0 and a variance equal to 1.

5 Computational experiment results

To validate the proposed approach, it has been decided to carry out the computational experiment. Those experiments aimed at answering the main question whether the proposed approach can help in decoding AIS messages when they contain errors, resulting, for example, from packet collisions.

5.1 Assumption on the number of clusters

The *k*-means algorithm does not provide build-in methods to assess the optimal quantity of clusters which the entire dataset should be divided into. Wherefore, the selection of that value should have been done in a different manner, preceded by some preliminary analysis of the dataset.

The goal of clustering was to separate AIS messages originating from one vessel from the messages sent by any other vessel in the given area. That would further allow to carry a deeper analysis of those selected messages to find abnormal (either disrupted or incomplete) ones among them. Following this line of reasoning, we could expect to observe that ideally our dataset would be divided into clusters that only consist of messages from one vessel, with no messages from any other ship attached. Further reasoning leads to the conclusion that the optimal quantity of clusters should be equal to the number of individual vessels from which the messages forming the dataset originated.

To examine the aforementioned conclusion, the following approach has been proposed: iteratively, the *k*-means algorithm has been run to cluster the entire dataset into different numbers of groups. For example, for the dataset from the area of Gulf of Gdansk with 22 vessels, the experiment started with 2 clusters. To find the final number of iterations, firstly it had been checked that there were messages from 22 individual ships in our dataset, then we multiplied it by 1.5, that gives 33 in this scenario. Then, we computed the value of the silhouette coefficient for each of the obtained clustering results and plotted it to visually see how the value of the silhouette coefficient changes with the increasing number of clusters. Figure 5 presents the graph with the computed silhouette values for each number of clusters.

Analysing Fig. 5, the following conclusion can be drawn: the number of vessels in the dataset is indeed an optimal quantity of clusters, as one may notice that it is close to the number (which in this scenario is around 24) that corresponds to the

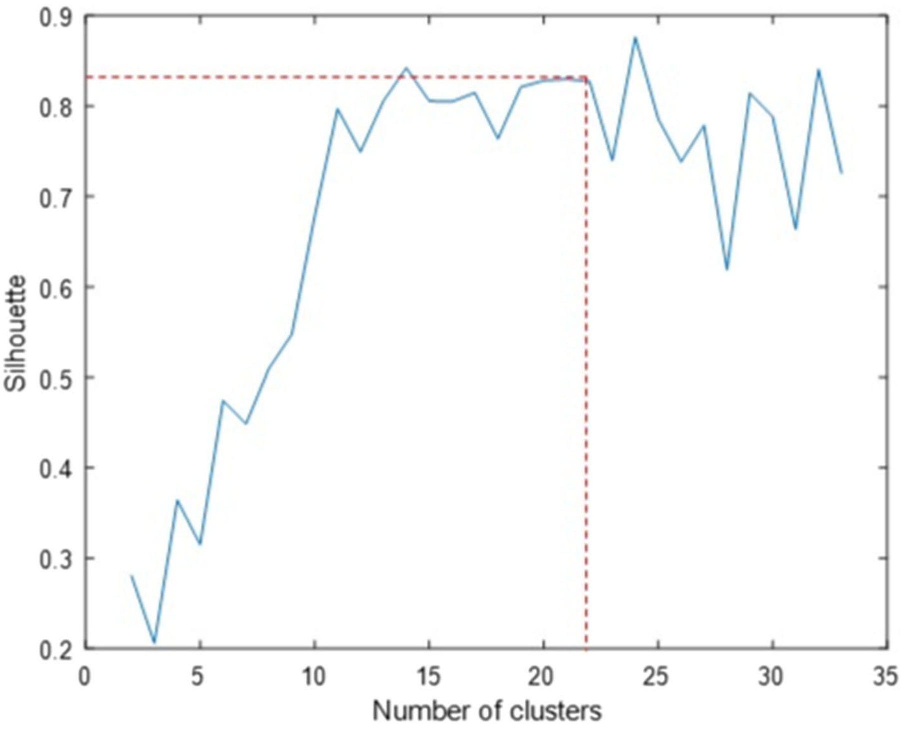

Fig. 5 Examples of silhouette coefficient values obtained while dividing the first dataset into various quantities of clusters (ranging from 2 to 33 clusters). Silhouette value corresponding to 22 clusters (i.e. the number of individual ships in this dataset) has been additionally marked

maximum value of silhouette; thus, it gives sufficient results in terms of silhouette and also proves that there is no need to conduct additional, computationally costly calculations to find the actual maximum of the silhouette value.

5.2 Clustering results

After the number of clusters was established, the clustering itself was conducted. The k-means algorithm has been run to divide the first dataset into 22 groups, the second dataset into 387 groups and the third dataset into 524.

For example, the results of the clustering for the first dataset are presented in Figs. 6 and 7. In case of the first dataset, the average silhouette value of the clustering, as shown in Fig. 7, is equal to 0.927456.

Using the second dataset — clustered into 387 groups — the value of the silhouette coefficient equals 0.9325, while for the third dataset — clustered into 524 groups — the coefficient equals 0.8696 (Figs. 8, 9, 10 and 11).

Moreover, a visual comparison of Figs. 2 and 6 (as well as Fig. 3 and 8, Fig. 4 and 10) may lead to the conclusion that datapoints clustered into the same group indeed mostly follow the trajectory of one ship, i.e. consists of points originating from one

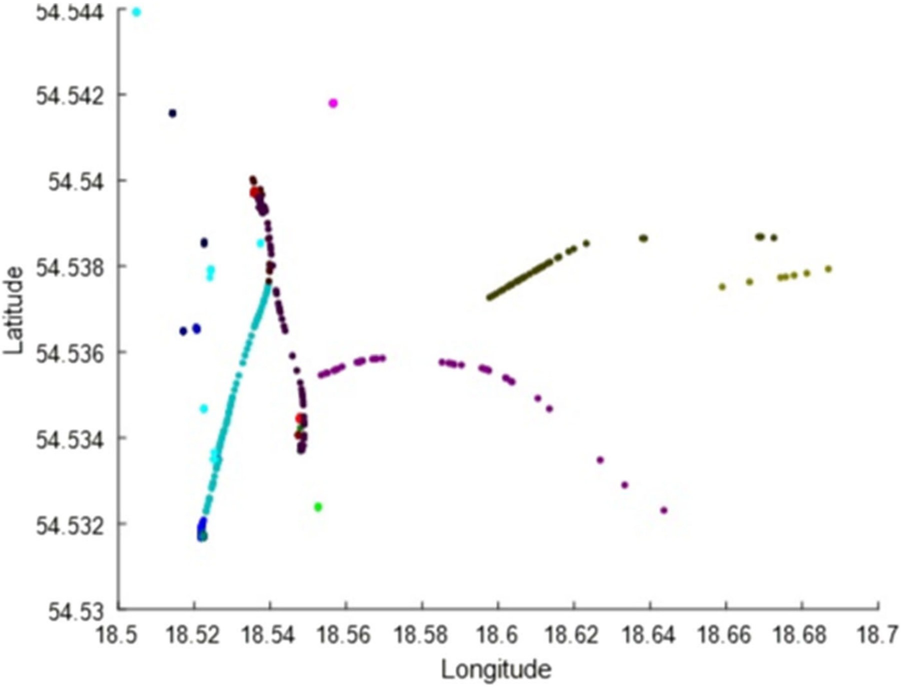

Fig. 6 Recorded AIS data from 22 ships clustered into 22 groups (each marked with a different colour)

particular vessel. However, relying on a visual rating may sometimes be misleading. For this reason, in the next section, for each considered dataset, a mathematical verification of the clustering correctness is used.

5.3 Clusters' homogeneity coefficient

As mentioned in the previous subsection, a need to mathematically validate the correctness of clustering the AIS messages has arisen. It is worth reminding that the main goal of the clustering was to separate messages originating from one vessel from those received from any other vessel, i.e. to put all messages from one ship into one cluster. Therefore, one of the possibilities to measure whether the clustering was successful or not is to ascertain if all datapoints gathered in a cluster originate from one particular ship, i.e. all the messages from one cluster share the same number in MMSI field.

To elaborate such a coefficient, each cluster has been examined using the following procedure:

1 Firstly, decide which MMSI is the most common in the cluster.
2 Secondly, calculate how many messages the cluster consists of.
3 Then, calculate how many messages in the cluster have the same MMSI as the one indicated in point 1.

4 Finally, calculate the percentage of how much the cluster is filled with the messages with modal MMSI value by dividing the value from point 3 by the value from point 2.

Some exemplary results are presented in Table 3. Once each cluster has been examined this way, a weighted average for those calculated percentages of all clusters has been computed with the volume of each cluster being a weight. The computed average may be called tentatively “the clusters’ homogeneity coefficient”, as it indicates the average fraction of each cluster that consists of messages originating from one ship, i.e. to what extent all clusters are homogeneous. It takes values in the range from 0 to 1. The values close to 0 can be interpreted as a result of clustering where nearly every single datapoint in each cluster has been received from a different vessel. On the opposite, results close to 1 indicate that messages from mostly one ship are gathered in each cluster.

The analysis of Table 3 provides the conclusions that in most cases, each cluster consists only of messages originating from one vessel (the value of 100% is present in most of the fields in the third row in Table 3), which is the desired behaviour of the clustering. However, several clusters (1 out of 22 in this case) include datapoints from more than one ship. The clusters’ homogeneity coefficient for the first dataset is 0.99751. It equals 0.982299 and 0.98535 for the second and third datasets, respectively.

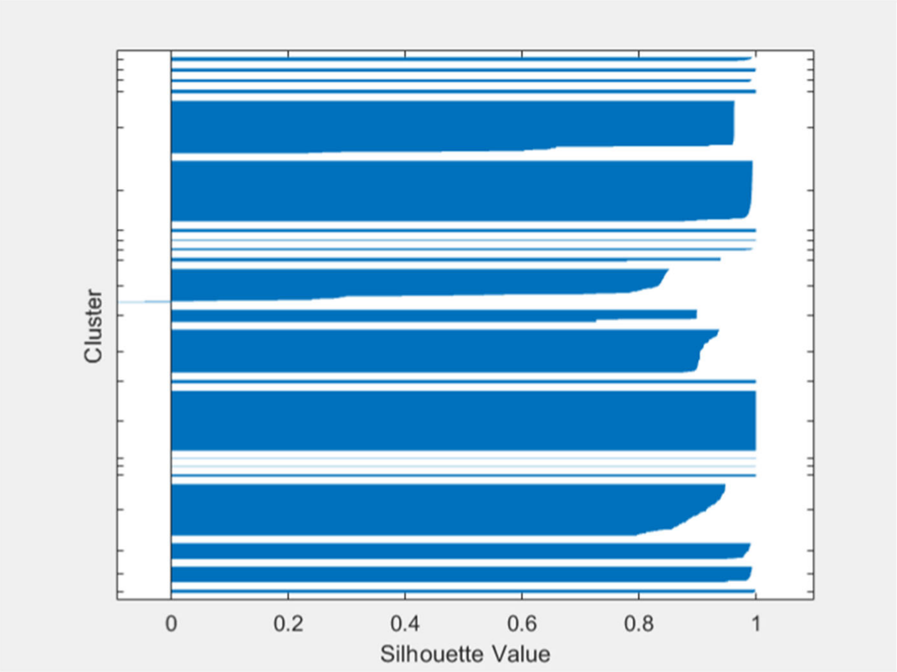

Fig. 7 Values of silhouette coefficient for each datapoint for clustering the first dataset into 22 groups — the average value equals 0.927456

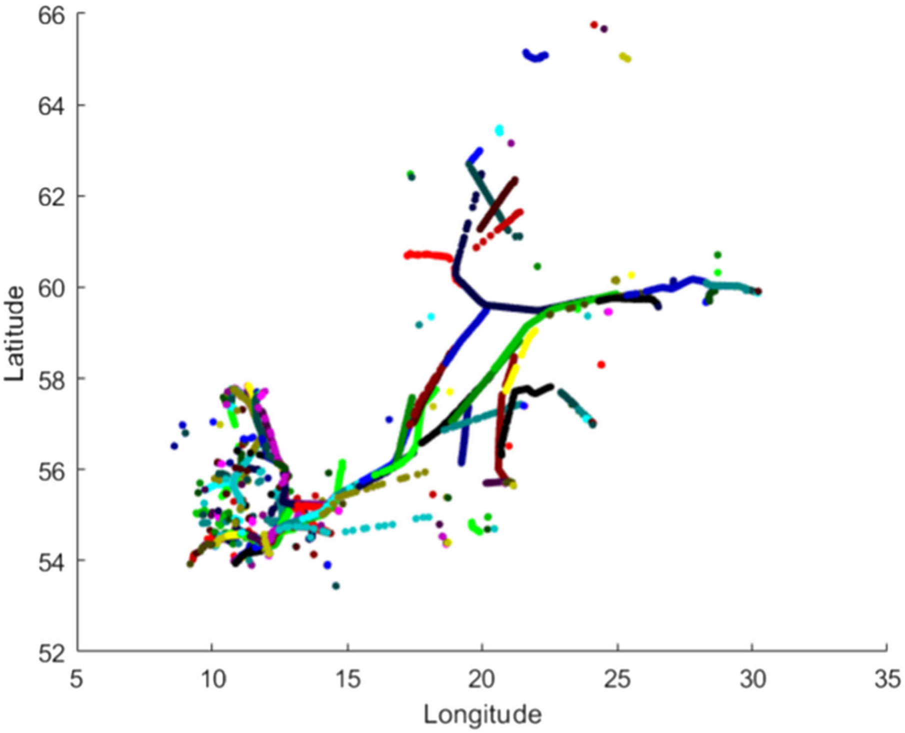

Fig. 8 Recorded AIS data from 387 ships clustered into 387 groups (each marked with a different colour)

5.4 Vessels' homogeneity coefficient

The method of calculating a so-called clusters' homogeneity coefficient of the AIS data clustering described above is, however, not the only one that can be elaborated. In fact, the clusters' homogeneity coefficient does not provide information whether packets from one vessel have been divided into several clusters. Those clusters may still be characterized with the high clusters' homogeneity coefficient (each of them might consist of packets from only one ship), although this behaviour is not desirable.

Table 3 Numerical values necessary for computing the correctness of clustering the first dataset according to the percentage of different MMSI identifiers prevalent in each cluster: cluster — the number of the cluster, volume — the amount of messages assigned to the corresponding cluster, % — percentage of the occurrence of modal MMSI value in the corresponding cluster

Cluster	1	2	3	4	5	6	7	8	9	10	11
Volume	110	129	183	26	128	91	5	4	34	7	7
%	100	100	100	100	100	100	100	100	100	100	100
Cluster	12	13	14	15	16	17	18	19	20	21	22
Volume	8	1	6	7	2	8	2	6	1	33	7
%	75	100	100	100	100	100	100	100	100	100	100

To elaborate such a coefficient, packets from each vessel have been examined using the following procedure:

1 Firstly, decide which clusters contain messages from the particular ship.
2 Secondly, designate the most common cluster that the messages have been assigned to.
3 Calculate how many messages have been received from that ship.
4 Then, calculate how many messages from that vessel have been assigned to the cluster indicated in point 2.
5 Finally, calculate the percentage of how many of the messages from the corresponding ship have been assigned to the modal cluster by dividing the value from point 4 by the value from point 3.

Another exemplary result of the aforementioned calculations is presented in Table 4. Again, after each source vessel has been examined this way, a weighted average for those calculated percentages of all ships has been computed with the number of messages originating from one particular ship being a weight. This time the name of the computed average might be "the vessels' homogeneity coefficient" as it indicates the average fraction of messages from each ship that have been grouped into one cluster. Similarly to the previous coefficient described, it takes values in the range of 0–1. Results close to 0 indicate that almost all messages from the ship have been

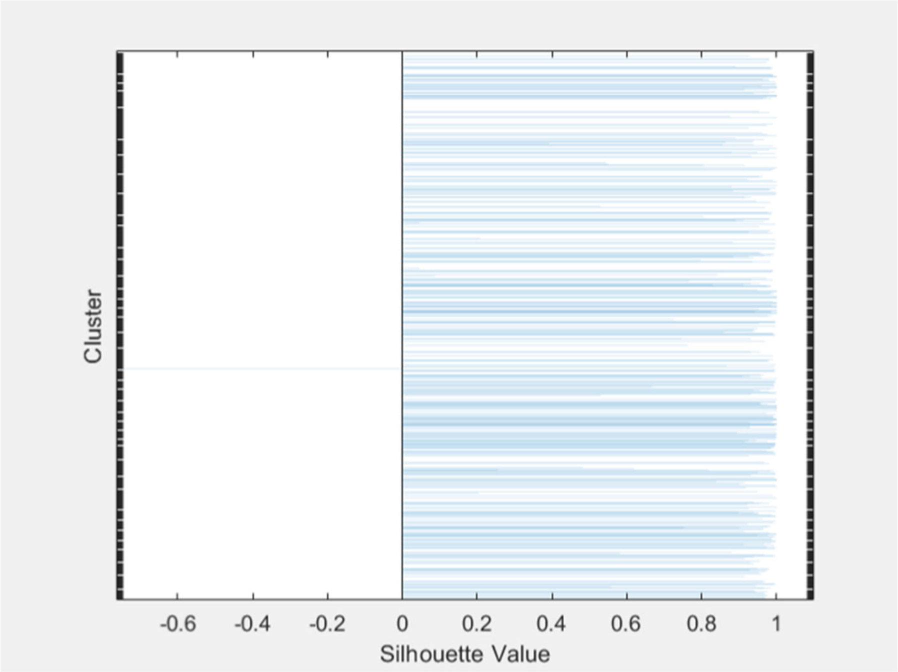

Fig. 9 Values of silhouette coefficient for each datapoint for clustering the second dataset into 387 groups — the average value equals 0.9325

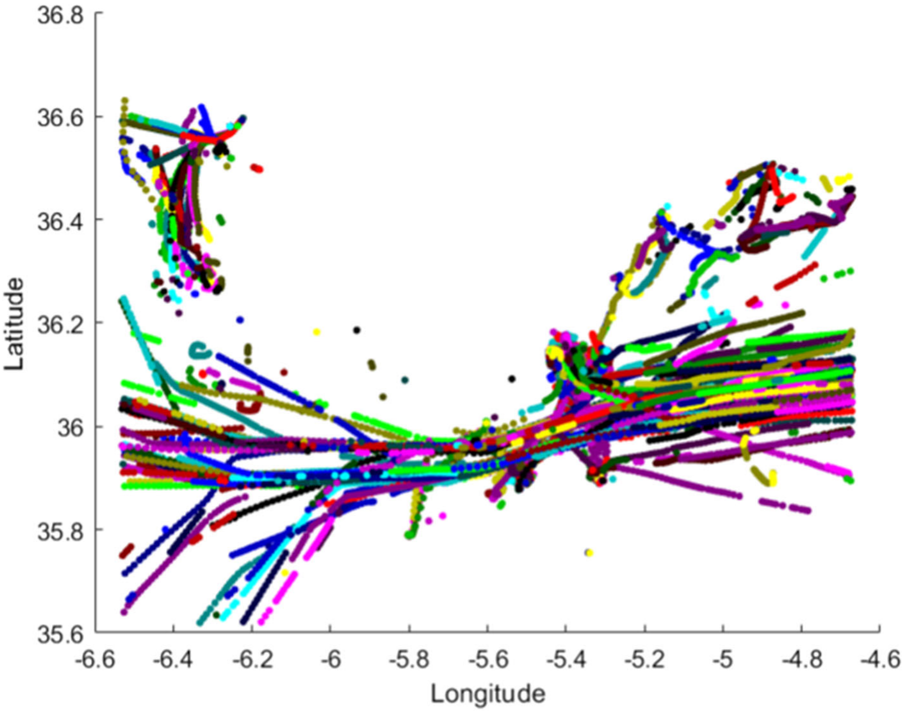

Fig. 10 Recorded AIS data from 524 ships clustered into 524 groups (each marked with a different colour)

clustered into different groups. Results close to 1 can be interpreted as a situation where every message from a particular vessel has been clustered into one group.

The conclusions that can be drawn from Table 4 can be spelled as follows: only several ships (5 out of 22) had transmitted messages that have been divided into 2 or more clusters, while the common behaviour of our clustering algorithm is to put those messages into one group. Therefore, the vessels' homogeneity coefficient for

Table 4 Numerical values necessary for computing the correctness of clustering the first dataset according to the amount of different clusters that include messages from particular vessel: vessel — the pseudo-identifying number of the particular ship, volume — the amount of messages originating from the corresponding vessel, % — percentage of how many of the messages from the corresponding ship have been assigned to the modal cluster

Vessel	1	2	3	4	5	6	7	8	9	10	11
Volume	110	129	183	26	128	91	5	4	34	7	7
%	100	100	61,2	100	100	100	100	100	100	100	100
Vessel	12	13	14	15	16	17	18	19	20	21	22
Volume	8	1	6	7	2	8	2	6	1	33	7
%	100	100	100	100	100	100	100	100	100	100	100

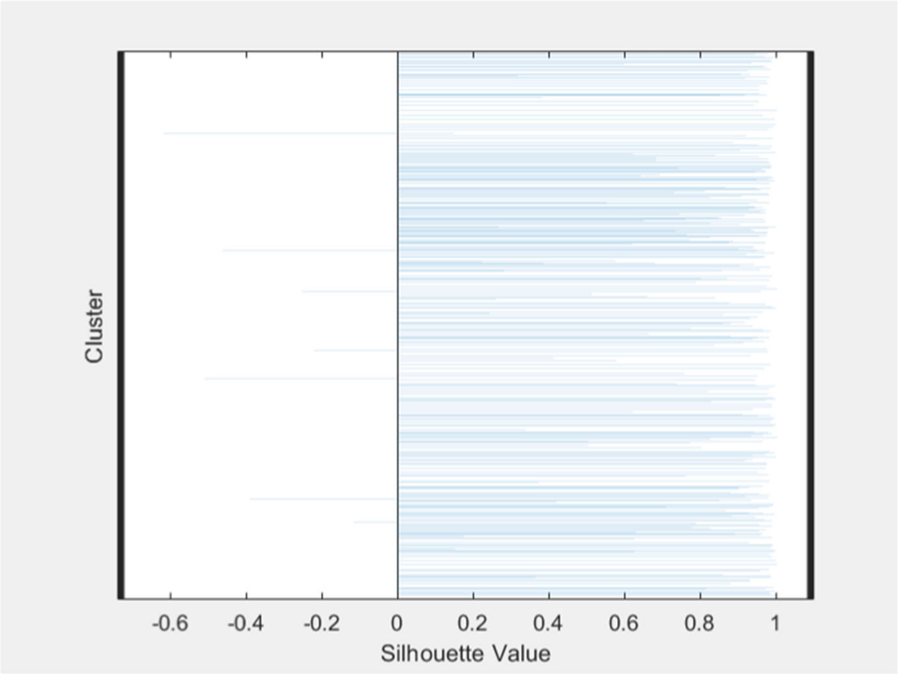

Fig. 11 Values of silhouette coefficient for each datapoint for clustering the third dataset into 524 groups — the average value equals 0.8696

the first dataset is 0.9118. It equals 0.957698 and 0.921396 for the second and third datasets, respectively.

5.5 Correctness coefficient

In previous sections, two possible measures of the correctness of the clustering have been proposed. They are similar to each other, yet both focus on two different aspects of clustering of the AIS data. If the clustering model works correctly, both of them should produce satisfactory, high results. However, assessment based on relying on two independent values might be either uncomfortable or impossible. Ideally, there could be one, uniform coefficient that would provide the answer to what extent the clustering is acceptable or not.

To create such a measure, a method similar to calculating F-score in classification problems has been developed. The coefficient called “the correctness coefficient” (CC) of the clustering is a harmonic average of the clusters’ homogeneity coefficient and vessels’ homogeneity coefficient. It has been computed using the following formula:

$$CC = \frac{2 \cdot clusters' \quad homogeneity \quad coefficient \cdot vessels' \quad homogeneity \quad coefficient}{clusters' \quad homogeneity \quad coefficient + vessels' \quad homogeneity \quad coefficient}$$

Not only does the aforementioned coefficient combine and compromise all of its components, but it also retains the trait of taking the range from 0 to 1. If any of its components (either clusters' homogeneity coefficient or vessels' homogeneity coefficient) produces 0, the correctness coefficient will drop to 0, while if all component coefficients take the value of 1, correctness coefficient will also result in 1. Therefore, the closer to 1, the better results of clustering have been obtained and, on the opposite, the closer to 0, the worse results of clustering.

For the considered datasets, the correctness coefficient equals 0.952734, 0.9728 and 0.9479 for first, second and third datasets, respectively.

5.6 Clustering of damaged messages

Apart from mathematical methods of evaluating the correctness of the clustering results, another way of verifying whether the algorithm works in an acceptable manner is to check its work in practice. As mentioned before, the aim of clustering is to prepare a background for further anomaly detection analysis. To predict the missing parts of AIS messages, damaged or incorrect packets received from one ship should still be clustered with all other messages from the same vessel, despite the fact that some of their fields contain abnormal values.

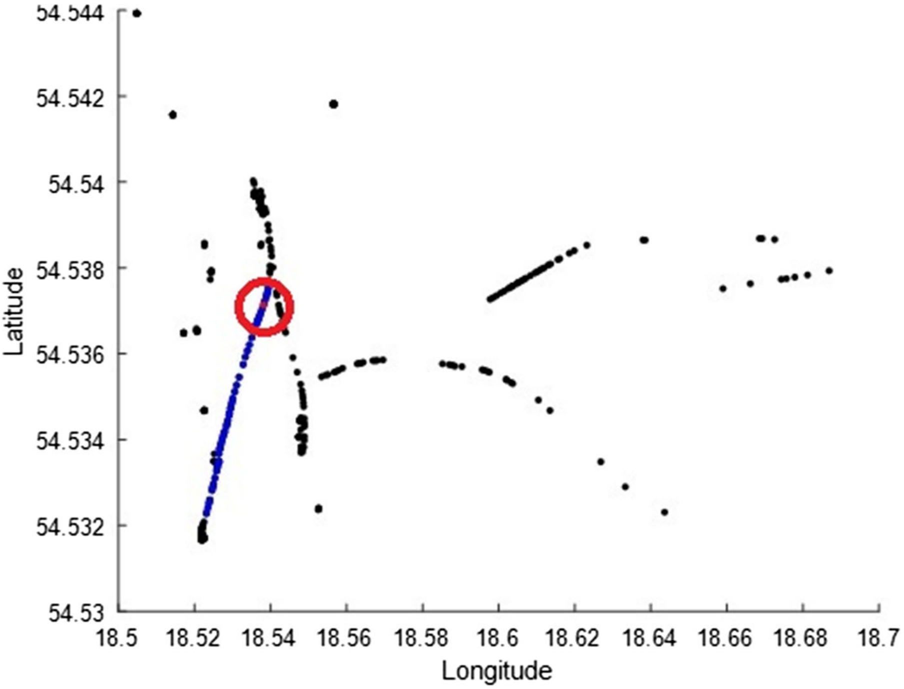

Fig. 12 Visualization of damaging the AIS message and repeated clustering — single datapoint randomly selected to form an outlier (red) and its cluster (blue)

To verify whether the proposed model assigns the damaged messages to the same cluster as it would do with the intact ones, additional experiments have been conducted. For the need of a clear analysis of this part of experiments, they have been carried out using only the first dataset.

The purpose of the first experiment was to visualize the clustering of a randomly chosen message in two scenarios: when the message had been damaged and not. After running a regular clustering algorithm and assigning all datapoints to their closest clusters, one message from the first dataset was randomly selected. The chosen one and its corresponding cluster are shown in Fig. 12. Then, the selected datapoint was damaged to form an artificial outlier. Again, to make the visualization comprehensible on a 2D plot, only its location features (latitude and longitude fields) were slightly changed, which simulates e.g. the GPS drift in a real environment. The clustering process was repeated to investigate whether the datapoint would be assigned to the same cluster as before or not. Experiments confirmed that in most cases, no change in clustering had occurred, which can be seen in Fig. 13.

Since changing the location features in a restricted range did not change the clustering results, further experiments were required. This time not only location features should have been corrupted, but the damage should have been spread among all 168 message bits to see if the algorithm can handle more sophisticated disruption. To

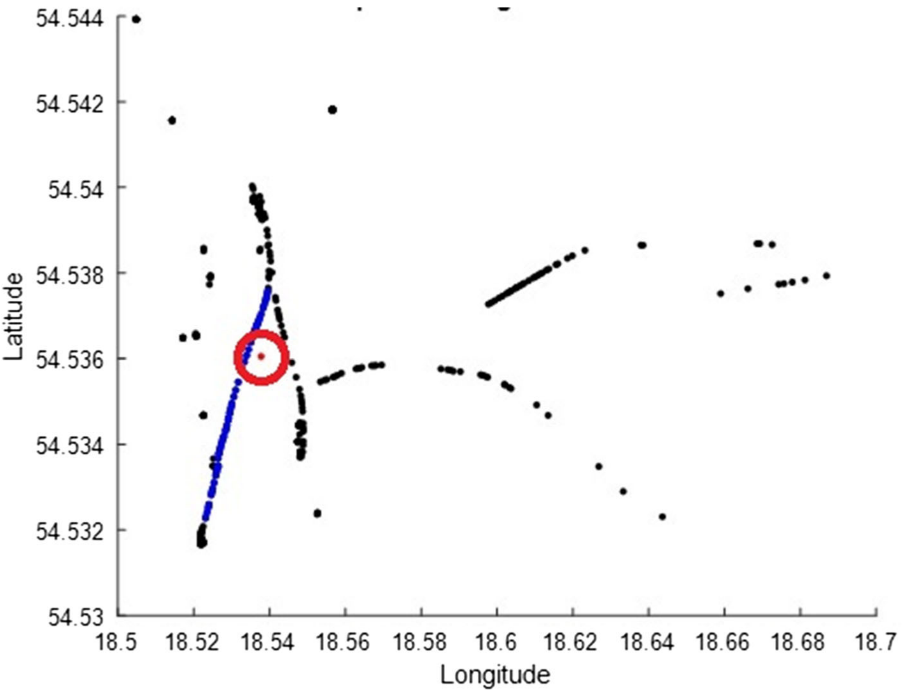

Fig. 13 Visualization of damaging the AIS message and repeated clustering — damaged message (outlier, red) and its cluster after repeated clustering (blue)

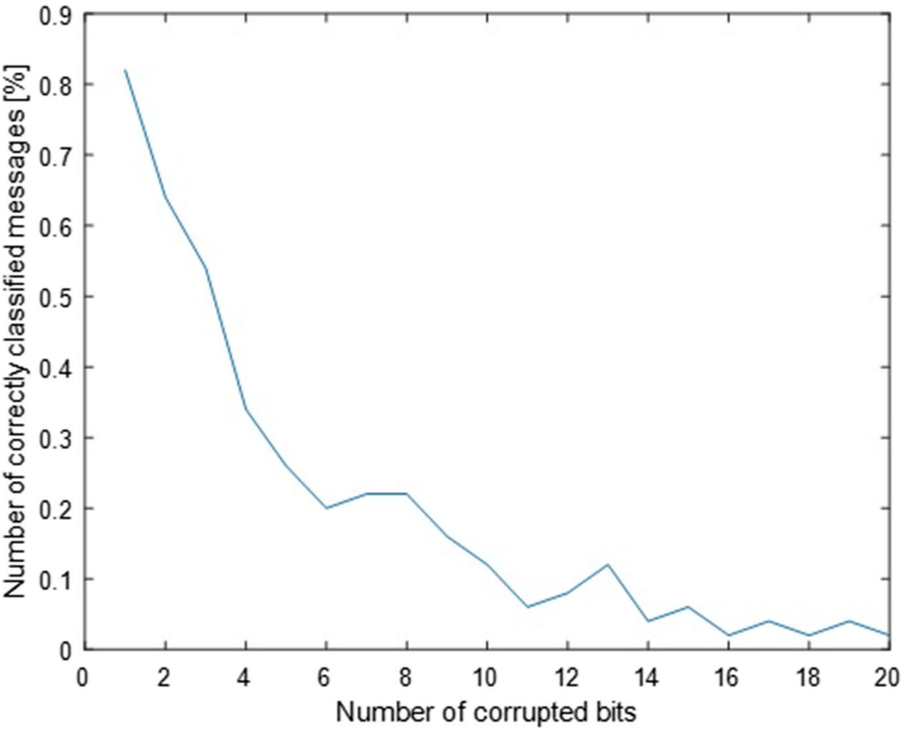

Fig. 14 The average percentage of correctly classified messages (i.e. assigned into the same groups) after the corruption versus the number of bits damaged in each examined message

accomplish this, a random noise was generated to disrupt the data. One hundred sixty-eight-dimensional binary vectors were created with 0 values on most of the positions and 1s on a certain amount of randomly selected positions with the amount depending on the level of disruption to be achieved. Bits with a value of 1 simulate bits that have been misinterpreted. Next, the XOR logical operation on a noise vector created this way and a randomly chosen message was conducted to form an outlier.

The aim of this part of the research was to determine the impact of the level of distortion on the correctness of clustering, i.e. on the number of messages that were clustered to the same groups as before despite the disruption. The work was conducted in an iterative way. At each iteration, the percentage of damaged bits (i.e. the number of 1s in a noise vector) was increased. Fifty randomly selected datapoints were not disrupted with the help of the noise vector. After that, the clustering algorithm was run and each of the artificial outlying datapoint was examined according to the result of its clustering, i.e. whether it was assigned to the same cluster as before. The percentage of correctly classified messages among the examined 50 ones has been plotted on an X axis, while the number of damaged bits (ranging from 1 to 20 out of 168) has been shown on a Y axis.

The research shows (in Fig. 14) that if the number of incorrect bits in the message does not exceed 1% (2 bits out of 168), the algorithm successfully groups the damaged data into the correct cluster as before the disruption (with over 80% accuracy).

6 Conclusions

This paper focuses on packet collision and recovery in the AIS system. The proposed approach is based on an unsupervised machine learning technique called clustering. One of its algorithms, the *k*-means algorithm, has been exploited to conduct a computational experiment to ascertain whether this approach provides sufficient results. The research proves that the clustering-based approach may be used in further works as an initial stage of AIS data analysis which aims to distinguish packets originating from one specific vessel (to enable the detection of abnormal ones between them and, consequently, predict their correct form): the CC value, authors' original coefficient which reflects to which extent one cluster consists of messages from only one ship, obtained high values between 0.94 and 0.97, with silhouette coefficient between 0.86 and 0.93. It is also worth noticing that even if some bits in an AIS message are incorrect, *k*-means still manage to find the right cluster for such a damaged packet.

The results are also very promising when we compare the results to other similar experiments. For example, using a more sophisticated TREAD method of ship trajectory extraction 80% of the traffic around Brest has been discovered (see in Pallotta et al. 2013a). For the same method, the percentage of correctly recognized routes varies from 40% (on Indian Ocean) to 95% around Gibraltar (see in Pallotta et al. 2013b), when for comparable data from Gibraltar area the method proposed in this paper managed to correctly cluster around 95% messages, regarding the CC value.

Further research will concern the subsequent improvement of clustering, i.e. studying the influence of different distance measures on the results, as well as the attempt to detect the outliers in each of the clustered data groups. Considering a character of the problem, new methods for outlier detection, as well as for data reconstruction, will be developed.

Acknowledgements Special thanks for Mr. Marcin Waraksa and Prof. Jakub Montewka from Gdynia Maritime University for sharing raw data, which have been used in the experiment.

References

Brownlee J (2017) Why one-hot encode data in machine learning? Mach Learn Mast. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning Accessed 13 December 2019

Czarnowski I (2019) A cluster-based approach to maritime data analysis: the case of SAT-AIS data analysis. In: Proceedings of the international association of maritime Universities (IAMU) conference, pp 182–190

European Space Agency (2019) Satellite - automatic identification system (SAT-AIS) overview. European space agency website. https://artes.esa.int/sat-ais/overview. Accessed 13 December 2019

exactEarth (2015) Satellite AIS. An exactEarth Technical White Paper. exactEarth Website. https://cdn2.hubspot.net/hubfs/183611/Old%20Content/Landing_Page_Documents/Satellite_AIS_White_Paper_Final-1.pdf. Accessed 06 July 2020

He YK, Zhang D, Zhang JF, Zhang MY (2019) Ship route planning using historical trajectories derived from AIS data. Int J Marine Navig Saf Sea Transport 13(1):69–76. https://doi.org/10.12716/1001.13.01.06

International Maritime Organisation (IMO) (2019) Automatic identification systems (AIS). IMO Website. http://www.imo.org/en/OurWork/Safety/Navigation/Pages/AIS.aspx. Accessed 13 December 2019

International Telecommunications Union (2014) Recommendation ITU-R M.1371-5 (02/2014). https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.1371-5-201402-I!!PDF-E.pdf. Accessed 13 December 2019

Lane RO, Nevell DA, Hayward SD, Beaney TW (2010) Maritime anomaly detection and threat assessment. In: 2010 13th Conference on IEEE Information Fusion (FUSION). https://doi.org/10.1109/ICIF.2010.5711998

Liang M, Liu RW, Zhong Q, Liu J, Zhang J (2019) Neural network-based automatic reconstruction of missing vessel trajectory data. In: 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA). https://doi.org/10.1109/ICBDA.2019.8713215

Mieczyńska M, Czarnowski I (2019) A cluster-based approach for AIS data analysis and vessel trajectory reconstruction. In: PP-RAI'2019 polskie porozumienie na rzecz rozwoju sztucznej inteligencji conference proceedings, pp 103–106

Millefiori LM, Zissis D, Cazzanti L, Arcieri G (2016) A distributed approach to estimating sea port operational regions from lots of AIS data. In: 2016 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/BigData.2016.7840774

Pallotta G, Vespe M, Bryan K (2013) Traffic route extraction and anomaly detection from AIS data. 2COST MOVE Workshop on Moving Objects at Sea

Pallotta G, Vespe M, Bryan K (2013) Vessel pattern knowledge discovery from AIS data: a framework for anomaly detection and route prediction, Entropy. https://doi.org/10.3390/e15062218

Prevost R, Coulon M, Bonacci D, LeMaitre J, Millerioux J, Tourneret J (2012) Extended constrained Viterbi algorithm for AIS signals received by satellite. In: 2012 IEEE First AESS European Conference on Satellite Telecommunications (ESTEL). https://doi.org/10.1109/ESTEL.2012.6400111

Seta T, Matsukura H, Aratani T, Tamura K (2016) An estimation method of message receiving probability for a satellite automatic identification system using a binomial distribution model. Sci J Marit Univ Szczecin 46(118):101–107

Swetha GM, Hemavathy K, Natarajan S (2018) Overcome message collisions in satellite automatic ID systems, Informa PLC Microwave & RF. https://www.mwrf.com/systems/overcome-message-collisions-satellite-automatic-id-systems. Accessed 13 December 2019

Wawrzaszek R, Waraksa M, Kalarus M, Juchnikowski G, Gorski T (2019) Detection and decoding of AIS navigation messages by a low earth orbit satellite. In: Sasiadek J (ed) Aerospace robotics III. GeoPlanet: earth and planetary sciences. Springer, Cham. https://doi.org/10.1007/978-3-319-94517-0_4

Wierzchon ST, Klopotek M (2015) Algorithms of cluster analysis. Institute of Computer Science Polish Academy of Science, Warsaw

Zhang W, Goerlandt F, Montewka J, Kujala P (2015) A method for detecting possible near miss ship collisions from AIS data, Ocean Eng. https://doi.org/10.1016/j.oceaneng.2015.07.046