The Journal of Navigation (2023), **76**:4–5 506–524
doi:10.1017/S0373463323000206

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH ARTICLE

Navigation pattern extraction from AIS trajectory big data via topic model

Iwao Fujino,[1]* and Christophe Claramunt[2]

[1] School of Information and Telecommunication Engineering, Tokai University, Tokyo, Japan
[2] Naval Academy Research Institute, French Naval Academy, Lanvéoc-Poulmic, France.
*Corresponding author: Iwao Fujino; Email: fujino@tokai.ac.jp

Received: 4 March 2023; **Accepted:** 28 May 2023; **First published online:** 10 July 2023

Keywords: automatic identification system; trajectory; ship behaviour; navigation

Abstract
This paper introduces a novel approach for extracting vessel navigation patterns from very large automatic identification system (AIS) trajectory big data. AIS trajectory data records are first converted to a series of code documents using vector quantisation, such as k-means and PQk-means algorithms, whose performance is evaluated in terms of precision and computational time. Therefore, a topic model is applied to these code documents from which vessels' navigation patterns are extracted and identified. The potential of the proposed approach is illustrated by several experiments conducted with a practical AIS dataset in a region of North West France. These experimental results show that the proposed approach is highly appropriate for mining AIS trajectory big data and outperforms common DBSCAN algorithms and Gaussian mixture models.

1. Introduction

The automatic identification system (AIS) is an automatic electronic tracking system based on transceivers used by vessels to broadcast their position and identification messages. This system originated in Sweden in the early 1990s for navigation safety and collision avoidance. It was adopted and recommended by the International Maritime Organisation (IMO) in 1998 (IMO, 1998). Therefore, in 2000, the International Convention for the Safety of Life at Sea was amended to enforce all large cargo and passenger vessels to be equipped with AIS (IMO, 2000). In 2002, AIS became compulsory even for bigger fishing vessels in the European Union Member States (CEC, 2008). In 2005, the U.S. Coast Guard mandated that all commercial marine vessels must continuously transmit AIS signals while transiting U.S. inland waterways and ports (USCOAST, 2000). Coastal stations and ships can receive messages from ships anywhere in the world. Moreover, AIS is not only a terrestrial infrastructure, but from 2008, satellite-based systems were also able to pick up AIS messages from oceangoing vessels far away from the coastline (Hoye et al., 2008; Best, 2011). Nowadays, AIS receivers are commonly connected to computer networks and data centres so that AIS messages can be daily stored and reused for secondary purposes. AIS data have emerged as a dominant contributor to maritime big data, thus offering useful resources for discovering maritime trends and patterns, moreover offering valuable support for maritime transport and surveillance operational decisions.

Navigation rules have been defined to ensure the safety of vessels in a port and its vicinity, but they may not be always applied so these rules cannot completely guarantee the safety of a given maritime situation. To implement successful AIS-based surveillance systems, there is still a need to develop real-time algorithms to derive navigation patterns, identify outliers and anticipate incidents

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

on the sea. With the massive increase of AIS data resources, it becomes feasible to extract navigation patterns by mining real AIS trajectory big data. Over the past few years, text mining and natural language processing have promoted novel solutions for the extraction of domain knowledge. Topic models extract specific topics from document collections (Blei, 2012). A topic model assumes that documents can be thought of as outward manifestations of topics, which are a small set of latent variables corresponding to the underlying themes, causes, or some kind of influences or forces behind the observed documents.

The goal of this study is to apply the topic model to derive navigation patterns from AIS trajectory data. The latent factor is expected to be acquired from the vessel's trajectory. Actually, a topic model needs to be supplied by a document collection in the form of a 'bag-of-words'. Certainly, AIS data records are not documents, they contain not only longitude and latitude location data, but also heading, speed and additional vessel attributes. To fit AIS trajectory data to the topic model, a vector quantisation procedure is introduced, that is, AIS trajectory data are transformed into a code document, and it can be presented in the form of a 'bag-of-codes'. In this way, the so-called topics are extracted by the topic model, and each topic is decoded back to longitude, latitude, heading and speed. However, the bottleneck issue we have to deal with is the large amount of incoming AIS data. As it has been shown that the k-means vector quantisation approach does not perform well in the setting of large datasets, we use a high-performance PQk-means algorithm recently introduced (Matsui et al., 2017). Overall, a combination of topic model and extended vector quantisation is applied to then extract specific navigation patterns for a given maritime region of interest. In fact, this approach breaks a timestamped series of AIS trajectories into a code distribution (i.e. approximately appearance frequency), without any regard for the sequence between individual data samples. Because the details of the original data are ignored, this approach has the potential to deal with a very large amount of data as well as providing a different view of the data and potentially identifying patterns.

The main contributions of this paper are summarised as follows.

(i) A PQk-means algorithm of vector quantisation is introduced to convert AIS trajectory data to a code document collection. The superiority of the PQk-means algorithm over the k-means algorithm is evaluated as applied to AIS data.

(ii) The topic model is applied to an AIS code document collection so that specific topics are extracted as an expression of codes. Then, codes can be decoded into longitude, latitude, heading and speed vector, so that these specific topics are drawn back on a map, from which emerges vessel navigation patterns. A peculiarity of the approach is that such navigation patterns can be observed in a maritime map at a relevant scale.

(iii) Several experiments are conducted to compare the topic model with conventional methods such as the density-based spatial clustering of applications with noise (DBSCAN) algorithm and Gaussian mixture model. The experimental results show that our proposed approach is highly appropriate for mining AIS big data when compared to alternative methods.

(iv) Vessels' sails are combined with different proportions of navigation patterns and similar sails are ranked according to cosine similarity.

(v) The performance of the topic model increases with the number of topics, but it becomes saturated when its value becomes too large. Moreover, the computation time of the topic model is proportional to the number of topics.

The rest of the paper is organised as follows. Section 2 briefly reviews the current work oriented to AIS data clusterings and pattern extractions. Section 3 describes the overall methodology and fundamental principles considered in the paper. Section 4 presents the experimental results of vector quantisation, DBSCAN algorithm, Gaussian mixture model and topic model, using a reference AIS dataset. Finally, Section 5 concludes the paper and provides a few directions for future works.

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

2. Related work

Rule-based expert systems have been long applied to monitor and analyse AIS-based maritime trajectory data (Edlund et al., 2006; Ray et al., 2013). Recent progress and research challenges in maritime data integration and analysis have been surveyed (Claramunt et al., 2017). These can be categorised as follows: maritime data integration and management, event pattern detection and trajectory analysis, and maritime decision support and forecasting systems.

An unsupervised and incremental learning algorithm has been suggested for extracting maritime movement patterns to convert large amounts of AIS data into decision-supporting information (Pallotta et al., 2013). The proposed methodology, so-called TREAD (traffic route extraction and anomaly detection) can cluster relevant events based on the temporal and spatial properties of vessel navigation, as well as provide up-to-date and high-level contextual information.

Over the past few years, AIS data have been combined with machine learning approaches to explore maritime trends and patterns (Piciarelli et al., 2008; Vries and van Someren, 2012; Dobrkovic et al., 2015). Machine learning methods have been applied to identify clusters using DBSCAN from positional data extracted from streaming AIS data (Dobrkovic et al., 2015). This supports the discovery of vessel waypoints in a given maritime traffic lane. Another research extracts sequential waypoint by using a genetic algorithm and then extract sailing patterns and transform them into a directed graph (Dobrkovic et al., 2018). Automatic methods for route patterns exaction have been developed by introducing a vector quantisation and topic model (Fujino et al., 2017). Furthermore, this work has been extended to event detection identified as anomalous navigation patterns and warning anomaly events in almost real-time (Fujino et al., 2018).

A first noticeable learning algorithm oriented to anomaly detection, from AIS data for sea surveillance, applied a Gaussian mixture as a cluster model and greedy version of the expectation-maximisation (Laxhammer, 2008). By using a real recorded dataset, the proposed model was trained and evaluated. Observed records that deviated from the regular model are detected as anomalous according to a presetting threshold of anomaly rate.

Another related work, from real-time and historic AIS data, applied data mining techniques to extract motion patterns and derive motion anomalies (Ristic et al., 2008). This anomaly detection work is carried out under the framework of adaptive kernel density estimation. They proposed an algorithm for vessel motion prediction (i.e. location, velocity) within a specified time window using a Gaussian sum tracking filter.

A current work using a deep learning approach has been reported (Zhang et al., 2023). This study proposed a real-time ship anomaly detection method driven by AIS data. The method uses the DBSCAN algorithm to cluster ship trajectories to identify a normal model and a deep learning algorithm as an anomaly detection tool.

3. Methodology

This section introduces the principles behind our methodological approach, the different procedures applied and their evaluation measures.

3.1. Overall solution

First, to employ the topic model to explore AIS vessel data, vector quantisation is introduced, that is, trajectories are encoded into a set of codes so that a document collection in ‘code’ is obtained. Furthermore, the topic model can be applied to the code document collection and achieve a code distribution for each topic. Finally, the representative codes for each topic are reconstructed to longitude, latitude, heading and speed, from which emerges the vessels’ navigation pattern.

The step-by-step procedures that are incrementally and algorithmically implemented are as follows.

(i) Prepare moving vessel dataset:

   (a) remove data samples of anchored or moored vessels from the original dataset;
   (b) segment AIS data series into sails at longer temporal intervals, where the word ‘sail’ denotes a continuous vessel sailing without stops.

(ii) Prepare code document collection for the topic model:

   (a) an AIS data record is a vector associated with longitude, latitude, heading and speed data. By introducing vector quantisation to all these vectors of AIS data, a codebook can be generated;
   (b) by referencing the codebook, each AIS data record of a sail can be converted to a code, thereby a sail can be converted to a code document. Moreover, by converting all sails in the dataset, a code document collection can be obtained from the original AIS dataset.

(iii) Apply the topic model to a code document collection. By applying the LDA method of the topic model to the code document collection, a document-topic matrix and a topic-code matrix are obtained. Then:

   (a) with the document-topic matrix, split the matrix to a topic distribution vector, then similar sails can be identified in relation to the proportion of topics by calculating cosine similarity;
   (b) with the topic-code matrix, choose the top-ranked codes as a topic representative, and draw back the top codes on the map of the region of interest, then vessel navigation patterns are likely to emerge.

3.2. *Vector quantisation: PQk-means algorithm*

Scalar quantisation is a one-dimensional signal processing technique to represent a signal sample by absorbing each value to linearly divided discrete gradations and then valued by a code (mostly from a finite set of integer numbers). When applied to vector data, if one uses scalar quantisation to each dimension of the vector, it may be difficult to achieve efficient precision even after setting a very large value of gradation. Vector quantisation is a lossy algorithm for data compression used in speech and image processing. The basic idea of vector quantisation comes from Shannon’s rate-distortion theory, which is only to quantise the observed data vectors rather than the whole space expanded by each dimension of a possible vector. It has been shown that better performance can be achieved by coding vectors rather than scalars (Linde et al., 1980; Gray, 1984). Essentially, vector quantisation is implemented in two steps. The first step is to generate a codebook, which contains a finite set of code and its reference information. The second step is to quantise each vector by comparing its value with the reference information in the codebook and then find its corresponding code.

Many implementations of vector quantisation have been proposed so far (Linde et al., 1980; Lloyd, 1982; Likas et al., 2003). Most common approaches are based on the k-means clustering algorithm. With the k-means algorithm, vectors in a given dataset are partitioned into K clusters and a code is assigned to each cluster. The code and centroid pair of each cluster (reference information) is stored as a codebook. Accordingly, each given vector can be expressed by the code of the cluster to which it belongs. For decoding, the code can be replaced by its centroid in the codebook. As a result, one can use a code number instead of a vector with many elements and there are only a finite number of codes for a specific application.

However, when using the k-means clustering algorithm for quantising large-scale vector datasets, the problems of vast memory consumption and prohibitive runtime costs become remarkable. Recently, a new approach that improved this problem significantly was introduced. The idea of product-quantisation (PQ) was originally developed for source coding (Gray, 1984) and further applied to approximated nearest neighbour search (Jégou et al., 2011). In the later work, each vector was compressed into a short code, called PQ-code, and the search was conducted over the PQ codes using lookup tables. More precisely, an intermediate code is introduced to compress the original vector dataset to a lower dimension and short-length dataset. Based on this concept, a billion-scale memory-efficient clustering algorithm has been developed, called the PQk-means algorithm (Matsui et al., 2017).

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

Assuming a dataset of D-dimensional vectors $\mathbf{p}_n = \{x_1, x_2, \ldots, x_D\}$ is given, where $n \in \{1, 2, \ldots, N\}$, first, split the dataset into M disjoined subvectors $\mathbf{p}_n = \{s_1, s_2, \ldots, s_M\}$. Then for each D/M-demensional vector, the closest codeword from pre-trained L codewords is determined. Finally, vector $\mathbf{p}_n$ can be expressed with a code vector, whose element is the code determined by its subvectors, as follows:

$$\mathbf{c}_n = \{c^{s1}, c^{s2}, \ldots, c^{sM}\} \tag{3.1}$$

For two given code vectors, the squared Euclidian distance is given by

$$d(\mathbf{c}_i, \mathbf{c}_j) = \sum_{m=1}^{M} (c^{im} - c^{jm})^2 = \sum_{m=1}^{M} \sum_{l=1}^{L} (\bar{\mathbf{p}}_l^{im} - \bar{\mathbf{p}}_l^{jm})^2 \tag{3.2}$$

and then cluster the resultant intermediate codes with the k-means algorithm, that is, the cluster to which each vector belongs is determined by assigning the vector to K clusters to minimise the following cost function:

$$W(C) = \sum_{k=1}^{K} N_k \sum_{C(i)=k} d(\mathbf{c}_i, \mathbf{c}_k) \tag{3.3}$$

where $N_k = \sum_{i=1}^{N} I(C(i) = k)$ is the number of vectors of the kth cluster, and $d(\mathbf{c}_i, \mathbf{c}_k$ is the distance between the ith code vector and kth code vector. With this algorithm, clustering a one billion dataset with 100,000 clusters is achieved in 14 h on a 32GB RAM personal computer (Matsui et al., 2017). Unlike other existing large-scale clustering methods, such as Bk-means (Gong et al., 2015) and IQk-means (Avrithis et al., 2015), the original vector sample can also be reconstructed approximately from a given code, this being particularly efficient for large AIS trajectory data.

3.3. Topic model

The topic model provides an effective approach for extracting conceptually coherent topics from a massive document dataset towards a considerably small set of latent variables. To the best of our knowledge, one of the most useful implementations of the topic model is the latent Dirichlet allocation (LDA) algorithm (Blei et al., 2003), which is an unsupervised machine learning technique to identify latent topic information from a massive document collection.

Let us give a brief description of the LDA algorithm. In a topic model, a document is described as a bag of words, i.e. attention is not given to the order of these words, just care about the appearance and frequency of words. In this way, the topic model can focus on solving problems with a probability distribution of words. Furthermore, the topic model assumes that a document is composed of several topics. This assumption means that the distribution of words depends on the topics selected for an article. It is also assumed that all documents in a data collection can be composed of a limited number of topics.

From this standpoint, the probability distribution of words in a given document collection **W** under the condition of topic proportion matrix θ can be expressed as follows:

$$P(\mathbf{W} \mid \theta) = \prod_{n=1}^{N} \sum_{z} P(w_n \mid z) P(z \mid \theta) \tag{3.4}$$

where $P(z \mid \theta)$ is the topic distribution on θ and $P(w_n \mid z)$ is the word distribution on topic z. In an LDA implementation, the topic proportion is assumed as a Dirichlet prior distribution, so that Equation (3.4) is reduced to the following equation:

$$P(\mathbf{W}) = \int P(\mathbf{W} \mid \theta) P(\theta)\, d\theta = \int P(\theta) \prod_{n=1}^{N} \sum_{z} P(w_n \mid z) P(z \mid \theta)\, d\theta \tag{3.5}$$

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

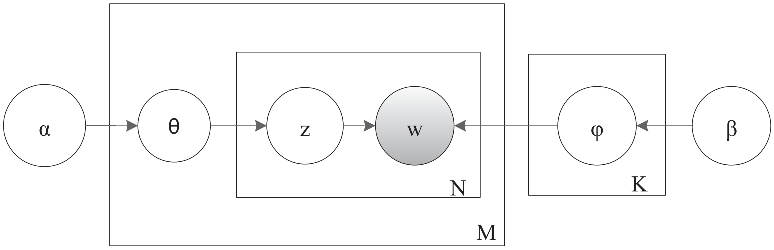


Figure 1. *Graphical model representation of LDA.*

where $P(\theta)$ is a Dirichlet prior distribution. Based on Equation (3.5), the LDA model can be schematised and depicted by a graphical representation as shown in Figure 1. In Figure 1, θ is a topic distribution of documents and ϕ is a word distribution of topics. Here, α is an hyper parameter for generating θ, β is an hyper parameter for generating ϕ, z is a topic matrix, and w is a word count matrix for each document and each word. Furthermore, K stands for the number of topics, M stands for the number of documents and N stands for the number of unique words in the given document collection.

According to this graphical model representation, a document collection can be generated by the following procedures.

(i) For each topic $k = 1, 2, \ldots, K$:

(a) draw word distribution,

$$\phi_k \sim Dir(\beta) \tag{3.6}$$

(ii) For each document $d = 1, 2, \ldots, M$:

(a) draw topic distribution,

$$\theta_d \sim Dir(\alpha) \tag{3.7}$$

(iii) For each word $n = 1, 2, \ldots, N_d$:

(a) draw topic,

$$z_{dn} \sim Mult(\theta_d) \tag{3.8}$$

(b) draw word,

$$w_{nm} \sim Mult(\phi_{z_{dn}}) \tag{3.9}$$

where ϕ_k denotes the distribution of codes in topic k, and θ_d denotes the topic proportion in document d. Here, $Dir(\cdot)$ denotes the Dirichlet prior distribution and $Mult(\cdot)$ denotes the multinomial distribution. After the inference of LDA to fit the given document collection, the results are represented as a distribution over codes in which top probability codes form a semantically coherent concept, and each document can be represented as a distribution over the discovered topics. As a result, the LDA provides two matrices, one is $\mathbf{\Theta}$, called a topic distribution of documents, and the other is $\mathbf{\Phi}$, called a word distribution of topics, as shown in Equations (3.10) and (3.11):

$$\mathbf{\Theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_M \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{12} & \ldots & \theta_{1K} \\ \theta_{21} & \theta_{22} & \ldots & \theta_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{M1} & \theta_{M2} & \ldots & \theta_{MK} \end{pmatrix} \tag{3.10}$$

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_K \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \dots & \phi_{1N} \\ \phi_{21} & \phi_{22} & \dots & \phi_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{K1} & \phi_{K2} & \dots & \phi_{KN} \end{pmatrix} \tag{3.11}$$

The matrix $\mathbf{\Theta}$ of Equation (3.10) gives a topic distribution for each document, which means a document can be represented by a combination of topics and the elements of the θ matrix are the weights for each topic here. However, the number of topics K is quite small, as is the number of words N, so the document can be converted to a very small dimension vector by the θ matrix. The matrix $\mathbf{\Phi}$ of Equation (3.11) gives a word distribution for each topic, which means that each word has a specific appearance probability. In other words, some words with high probability may appear very frequently for a specific topic, this yields that the top-ranked words in the distribution are a kind of representative words for the topic.

3.4. *Performance measure for vector quantisation*

To evaluate the result of vector quantisation, a performance measure is introduced. For each element x of a given vector, its relative mean root squared error (RRMSE) is defined as follows:

$$\text{RRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - \tilde{x}_n)^2}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}} = \sqrt{\frac{\sum_{n=1}^{N}(x_n - \tilde{x}_n)^2}{\sum_{n=1}^{N} x_n^2}} \tag{3.12}$$

where x is the original value of the element and $\tilde{x}$ is the reconstructed value of the element from its code.

3.5. *Performance measure of the topic model*

Because the topic model intends to separate the latent factor into its basic components, a valuable topic model should reflect sufficient differences between any two topics. Considering the Jessen–Shannon divergence gives the distance or unsimilarity between two probability distributions, an average of the Jessen–Shannon divergence between any two different topics is chosen as the total performance measure for a specific topic model, whose definition is given as follows:

$$\overline{D_{JS}} = \frac{1}{K(K-1)} \sum_{p \neq q} D_{JS}(\phi_p, \phi_q) \tag{3.13}$$

where $D_{JS}(\phi_p, \phi_q)$ is the Jessen–Shannon divergence between topic p and q. Also, ϕ_p and ϕ_q are the code distribution of topic p and q, respectively.

4. Experiments and evaluations

4.1. *Preliminary experiment: moving vessels dataset*

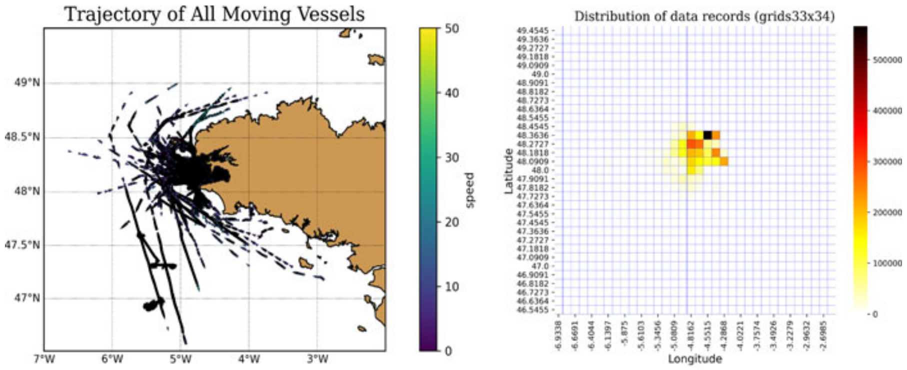
The original AIS dataset used for our experiments is the one distributed by the 2nd Maritime Big Data Workshop (MBDW) (Ray et al., 2018; MBDW, 2020), collected from Brest Bay in North West France from October 1, 2015 to March 31, 2016. The dataset attributes are sourcemmsi, navigationalstatus, rateofturn, speedoverground, courseoverground, trueheading, lot(longitude), lat(latitude) and t(timestamp). The original dataset contains 19,035,630 AIS records, which come from 3,928 ships. However, there are 8,203,254 records whose speed equals 0 knot/s in this dataset.

For a preliminary exploration of the AIS trajectory of the moving vessels, a moving vessels' dataset is first prepared by the following procedures.

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

Table 1. *Descriptions of the original dataset and two derived moving vessels' dataset.*

Item	Original dataset	MovingVessels dataset
Number of attributes	9	6
Number of records	19,035,630	1,146,372
Number of ships	3,928	76
Number of sails	–	209
Upper left longitude	−9 · 71	−7
Upper left latitude	50 · 88	49 · 5
Lower right longitude	−0 · 01	−2
Lower right latitude	45	46 · 5



(a) All trajectory in MovingVessels dataset

(b) Distribution of MovingVessels dataset

Figure 2. *Trajectory and distribution of the MovingVessels dataset. (a) All trajectory in the MovingVessels dataset. (b) Distribution of the MovingVessels dataset.*

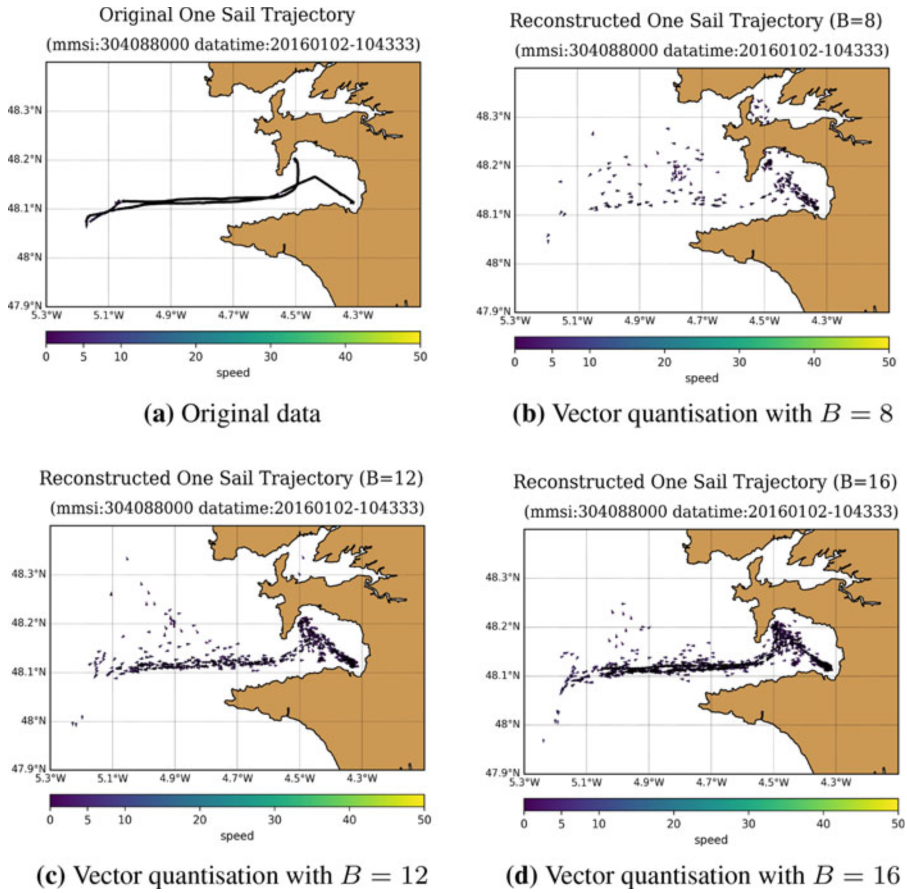
(i) A region of interest is defined in Brest Bay and its vicinity, which gives a rectangle area of (longitude = −7 · 0, latitude = 49 · 5) and (longitude = −2 · 0, latitude = 46 · 5). Data records out of this coverage are ignored.
(ii) To clean the dataset, records without valid heading values are removed.
(iii) To select the moving vessels, records with a speed value equal to 0 knot/s are removed.
(iv) Only retain the attributes of sourcemmsi, lon, lat, speedoverground, trueheading and t, others are removed.

After these procedures, the dataset shrinks to 4,980,363 records of 3,928 ships. Then AIS data records are categorised according to vessels' MMSI and segmented to sails when the interval of its timestamp is longer than one hour. Furthermore, based on the following two assumptions, a minimum length of data records is introduced for making this dataset.

(i) Because navigation patterns are composed of a series of AIS data records, longer trajectories are more likely to reflect representative patterns.
(ii) Because computational costs for vector quantisation depend on the total number of data records, short-distance sails are not retained.

Without loss of generality, a minimum length of 3,000 is experimentally chosen. This yields a subset of the original dataset, so-called as the MovingVessels dataset in the rest of the paper. The descriptions of basic items for the original dataset and MovingVessels dataset are shown in Table 1.

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press


Original One Sail Trajectory
(mmsi:304088000 datatime:20160102-104333)
Reconstructed One Sail Trajectory (B=8)
(mmsi:304088000 datatime:20160102-104333)
Reconstructed One Sail Trajectory (B=12)
(mmsi:304088000 datatime:20160102-104333)
Reconstructed One Sail Trajectory (B=16)
(mmsi:304088000 datatime:20160102-104333)
48.3°N
48.2°N
48.1°N
48°N
47.9°N
5.3°W
5.1°W
4.9°W
4.7°W
4.5°W
4.3°W
0
5
10
20
30
40
50
speed

(a) Original data

(b) Vector quantisation with $B = 8$

(c) Vector quantisation with $B = 12$

(d) Vector quantisation with $B = 16$

Figure 3. *Trajectory of AIS data reconstructed from codes of k-means vector quantisation. (a) Original data, (b) vector quantisation with* $B = 8$*, (c) vector quantisation with* $B = 12$ *and (d) vector quantisation with* $B = 16$.

As a whole image of the dataset, the sail trajectories of the MovingVessels dataset are shown in Figure 2(a). Moreover, by dividing the coverage into 33×34 grids, where each grid area is approximately 10 km × 10 km, distribution of all data records is generated by counting the data records in each grid. The result of the experiment for the MovingVessels dataset is shown as a heatmap in Figure 2(b).

It is clear that data records occur intensively only in a few central grids, whereas only a very small number of data records occur in the other grids. Actually, this leads us to consider a four-dimension vector for our research, as scalar quantisation generates a huge number of data points and leads to expensive computation time and resources. This experimental result further motivates the introduction of vector quantisation when processing AIS trajectory data.

4.2. Experiments for vector quantisation

4.2.1. Experiment 1: vector quantisation via k-means algorithm

Vector quantisation is applied to these four-dimension data records obtained from previous procedures, with the codebook length L assigned in a style of $L = B^4$, although each dimension is not equally divided into B intervals, where the number B denotes the precision of a vector quantisation setting. For applying the k-means algorithm of vector quantisation, experiments with $B = 8$, 12, 16 have been conducted. As an example of the results, a trajectory of original data and three reconstructed trajectories from

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

Table 2. *Performance and elapsed time of k-means algorithm for different codebook lengths.*

		RRMSE (%)				
B	Codebook length	Longitude	Latitude	Heading	Speed	Elapsed time
8	4,096	1 · 294	0 · 112	2 · 85	0 · 551	2:46:14
12	20,736	0 · 739	0 · 059	1 · 655	0 · 287	8:17:54
16	65,536	0 · 464	0 · 037	1 · 072	0 · 159	16:45:32

different codebook lengths are shown in Figure 3. From these results, it can be confirmed that a bigger value of B gives better precision of reconstruction. However, there still remain some sparse data away from the trajectory. These sparse data arise because the encoding method is not a completely reversible process. Therefore, when reconstructing the initial data, that is, longitude, latitude, heading and speeds, the centroid of each code is used, leading to a scattering of the data as compared to the initial data. This trend is improved by the choice of longer codebooks as revealed by Figure 3. For different codebook lengths, the RRMSE of each attribute and the elapsed time are shown in Table 2. These results show that a good precision of reconstructed data can be achieved as the codebook length is efficiently long, but it takes unaffordable computation time.

4.2.2. Experiment 2: vector quantisation via PQk-means algorithm

For applying the PQk-means algorithm of vector quantisation, experiments with $B = 8$, 12, 16, 20 and 24 are conducted. As an example of the results, four reconstructed trajectories when $B = 12$, 16, 20 and 24 are shown in Figure 4, whereas their original trajectory is the same one as in Figure 3(a). From these results, it can be confirmed that a bigger value of B gives better precision of reconstruction, similar to Figure 3. Furthermore, the case of $B = 24$ in Figure 4 provided a better reconstruction precision and less sparse data than the case of $B = 16$ in Figure 3.

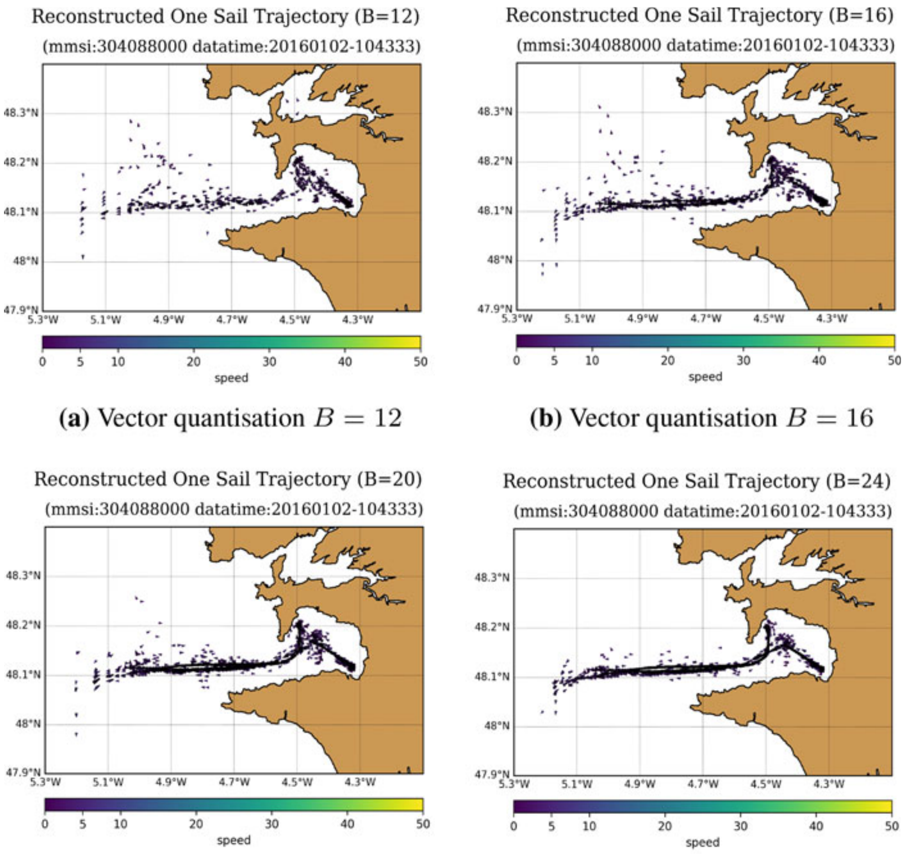
The RRMSE of each attribute and the elapsed time are shown in Table 3. These results show that a good precision of reconstructed data can be achieved as the codebook length is efficiently long, but its computation time becomes longer according to the codebook length. However, compared with the k-means algorithm, this result shows a great improvement of approximately 127 times in computation speed when $B = 16$. This provides a significant benefit when processing large amounts of data in a limited time. On the other side, suppose $B = 16$, when assigning the minimum length to 2,000 and 1,000, the total number of codes in the MovingVessels dataset becomes 1,898,973 and 3,590,309. The elapsed time for each case is 0:13:56 and 0:26:35, respectively. This implies that the computational time increases almost linearly when the minimum length decreases.

4.3. Experiments for validating the topic model

4.3.1. Experiment 3: clustering moving dataset – DBSCAN algorithm

The DBSCAN algorithm has been reported to extract waypoints from AIS data (Birant and Kut, 2007). To make a comparison with the topic model, a trajectory clustering experiment with the DBSCAN algorithm is conducted. As the DBSCAN algorithm directly clusters data records, data records of all sails are gathered together at first. In this experiment, the parameter of maximum distance is assigned to 0 · 2 and the number of samples in a neighbourhood is assigned to 5. With this setting, all the 1,146,372 trajectory data records in the MovingVessels dataset are clustered in 227 clusters, whereas 1,625 data records are judged as noise, so do not belong to any cluster. From the results, the clustering results appear as considerably imbalanced, the maximum number of data records is 1,105,352, whereas the number of data records in most clusters is less than 200. Because DBSCAN is a density-based algorithm, this

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press



Reconstructed One Sail Trajectory (B=12)
(mmsi:304088000 datatime:20160102-104333)

Reconstructed One Sail Trajectory (B=16)
(mmsi:304088000 datatime:20160102-104333)

Reconstructed One Sail Trajectory (B=20)
(mmsi:304088000 datatime:20160102-104333)

Reconstructed One Sail Trajectory (B=24)
(mmsi:304088000 datatime:20160102-104333)

(a) Vector quantisation $B = 12$

(b) Vector quantisation $B = 16$

(c) Vector quantisation $B = 20$

(d) Vector quantisation $B = 24$

Figure 4. *Trajectory of AIS data reconstructed from codes of PQk-means vector quantisation. (a) Vector quantisation* $B = 12$*, (b) vector quantisation* $B = 16$*, (c) vector quantisation* $B = 20$ *and (d) vector quantisation* $B = 24$.

result can be explained easily by the distribution of data records shown in Figure 2(b). To provide the typical trajectories of the data records, these clusters with length ≥ 200 are shown in Figure 5.

4.3.2. Experiment 4: clustering moving dataset – Gaussian mixture model

Gaussian mixture model is reported to detect an anomaly in sea traffic (Laxhammer, 2008). To compare with the topic model, a trajectory clustering experiment with the Gaussian mixture model is conducted. As the Gaussian mixture model cluster data records directly, the data records of all sails are first gathered together. Without loss of generality, the number of components is assigned to 20. With this setting, all the 1,146,372 trajectory data records in the MovingVessels dataset are categorised into each one of 20 clusters with different belonging probabilities. To provide the typical trajectories of the data records, the top 300 records for each cluster are shown in Figure 6.

4.3.3. Experiment 5: topic model – vessels' navigation patterns extraction

After vector quantisation to the MovingVessels dataset with parameter $B = 24$, 209 code documents are obtained. Then, according to the previous procedure, the LDA algorithm of the topic model is applied to these code documents.

The results of the experiment yield two matrices, document-topic matrix $\boldsymbol{\Theta}$ and topic-code matrix $\boldsymbol{\Phi}$. The topic-code matrix $\boldsymbol{\Phi}$ can be used to extract the navigation patterns. Because each row of the

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

Table 3. *Performance and elapsed time of the PQk-means algorithm for different codebook lengths.*

		RRMSE (%)				
B	Codebook length	Longitude	Latitude	Heading	Speed	Elapsed time
8	4,096	1 · 371	0 · 116	3 · 025	0 · 706	0:01:12
12	20,736	0 · 771	0 · 062	1 · 721	0 · 544	0:03:02
16	65,536	0 · 48	0 · 041	1 · 145	0 · 472	0:07:59
20	160,000	0 · 351	0 · 029	0 · 808	0 · 542	0:20:28
24	331,776	0 · 25	0 · 021	0 · 671	0 · 387	1:20:25

$\mathbf{\Phi}$ matrix gives a probability distribution on all the codes in the codebook, the top-ranked codes are used to represent the topic. However, a topic denotes a vessel navigation pattern. Therefore, these top-ranked codes can be selected as a representative of a navigation pattern. Furthermore, by referencing the codebook of vector quantisation, these representative codes can be reconstructed to their corresponding approximate data of longitude, latitude, heading and speed. Then, these data are depicted on the map of Brest Bay, so that the navigation patterns of vessels can be clearly visualised.

According to these considerations, for all rows of the $\mathbf{\Phi}$ matrix, these reconstructed codes give us all the navigation patterns for a specific topic set. Moreover, to investigate the effect of the number of topics, this work is applied to all $\mathbf{\Phi}$ matrices of experiments with $K = 20$, 40, 60, 80, 100, 120, 140, 180, 200, 220 and 240. For an overview of the results of this extraction, the navigation patterns when $K = 20$ are shown in Figure 7. Although $K = 20$ is not the best performance case among all of these experiments, considering space constraints, the case of $K = 20$ is chosen here. Additional details are presented in Section 4.4.3.

To provide a better understanding of vessel navigation patterns, an example of comparing the results between different K is shown in Figure 8. Figure 8(a) illustrates the reconstructed trajectory of topic #16 from the experiment results of the number of topics $K = 20$. For comparison, Figures 8(b), 8(c) and 8(d) illustrate the reconstructed trajectory of topics #9, #19 and #115 in the experiment results for the number of topics $K = 120$. By observation, it appears clearly that the pattern in Figure 8(a) is decomposed to the patterns in Figures 8(b), 8(c) and 8(d). Therefore, this confirmed that higher resolution of vessel navigation patterns can be achieved by assigning higher topic numbers.

4.4. Further discussions on the topic model

4.4.1. Discussion: comparison of topic model with DBSCAN algorithm and Gaussian mixture model

The experimental results of the topic model and two conventional methods, the DBSCAN algorithm and the Gaussian mixture model, have been shown in previous subsections. By comparing Figures 5–7, even $K = 20$ is not the best choice for the number of topics when using the topic model, and the following three points can be observed.

(i) The DBSCAN algorithm gives 227 clusters but there are approximately 96 · 4% of records in one cluster, and only 8 · 8% of clusters have more than 200 records. As this is an extremely imbalanced clustering result, it appears that the DBSCAN algorithm does not provide satisfactory results for this AIS trajectory dataset.

(ii) Most of the components of the Gaussian mixture model form flocks rather than line-shape trajectories of moving vessels. In fact, the topic model identifies much more navigation patterns than the Gaussian mixture model, even though the top 100 codes are chosen for the topic model and the top 200 records for the Gaussian mixture model.

(iii) The topic model provides an abstract representation of navigation patterns, this is not a set of real footprints of vessels such as obtained from the DBSCAN algorithm and Gaussian mixture model.

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

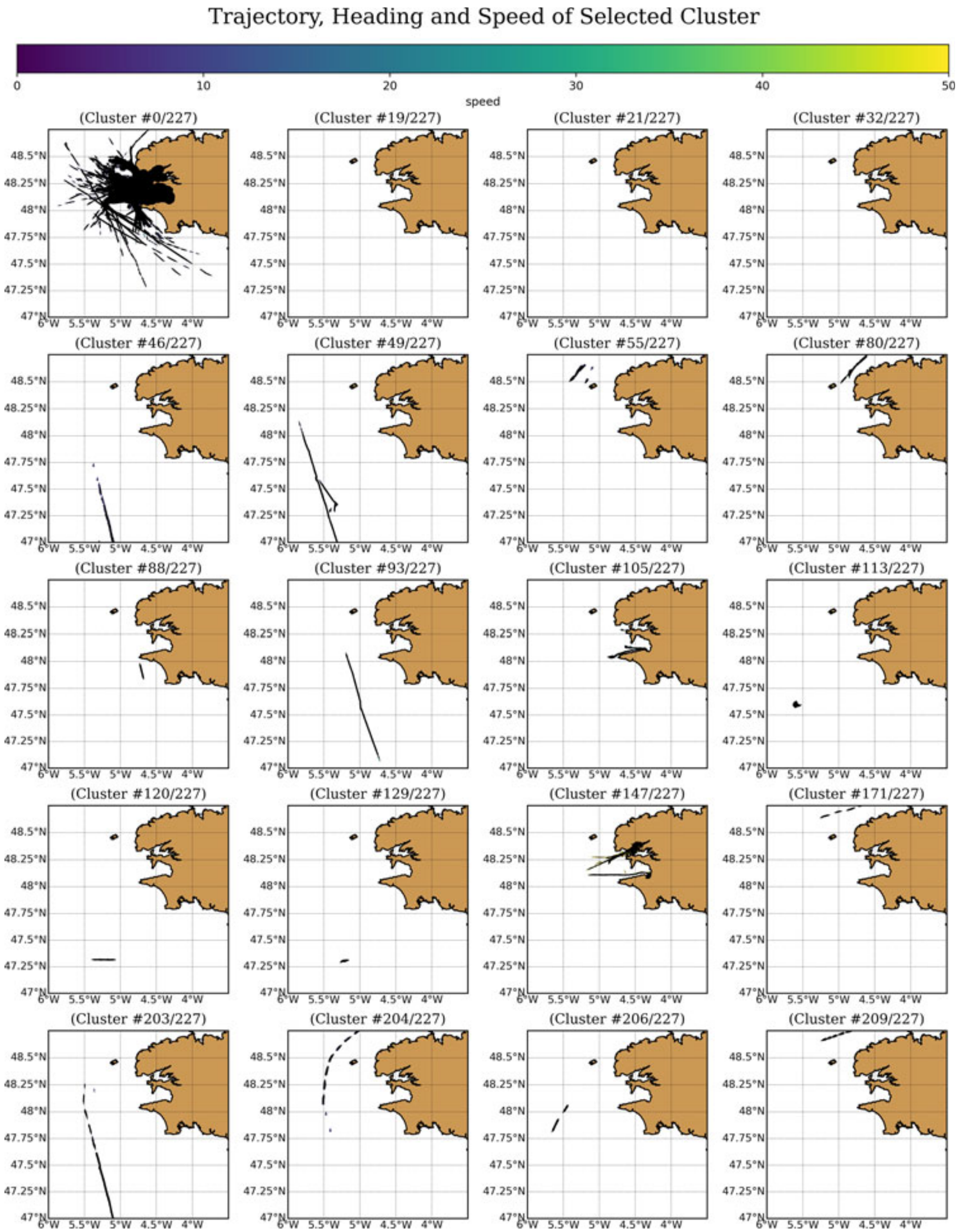


Figure 5. *Navigation pattern results extracted by DBSCAN algorithm (selected cluster (length of cluster* ≥ 200*)).*

> This is because the topics are reconstructed from the centroid of their representative codes, whereas the Gaussian mixture model represents the components with the clustered original data records directly.

Consequently, it is confirmed that the topic model provides more visually realisable navigation patterns than the results of the DBSCAN algorithm and Gaussian mixture model.

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

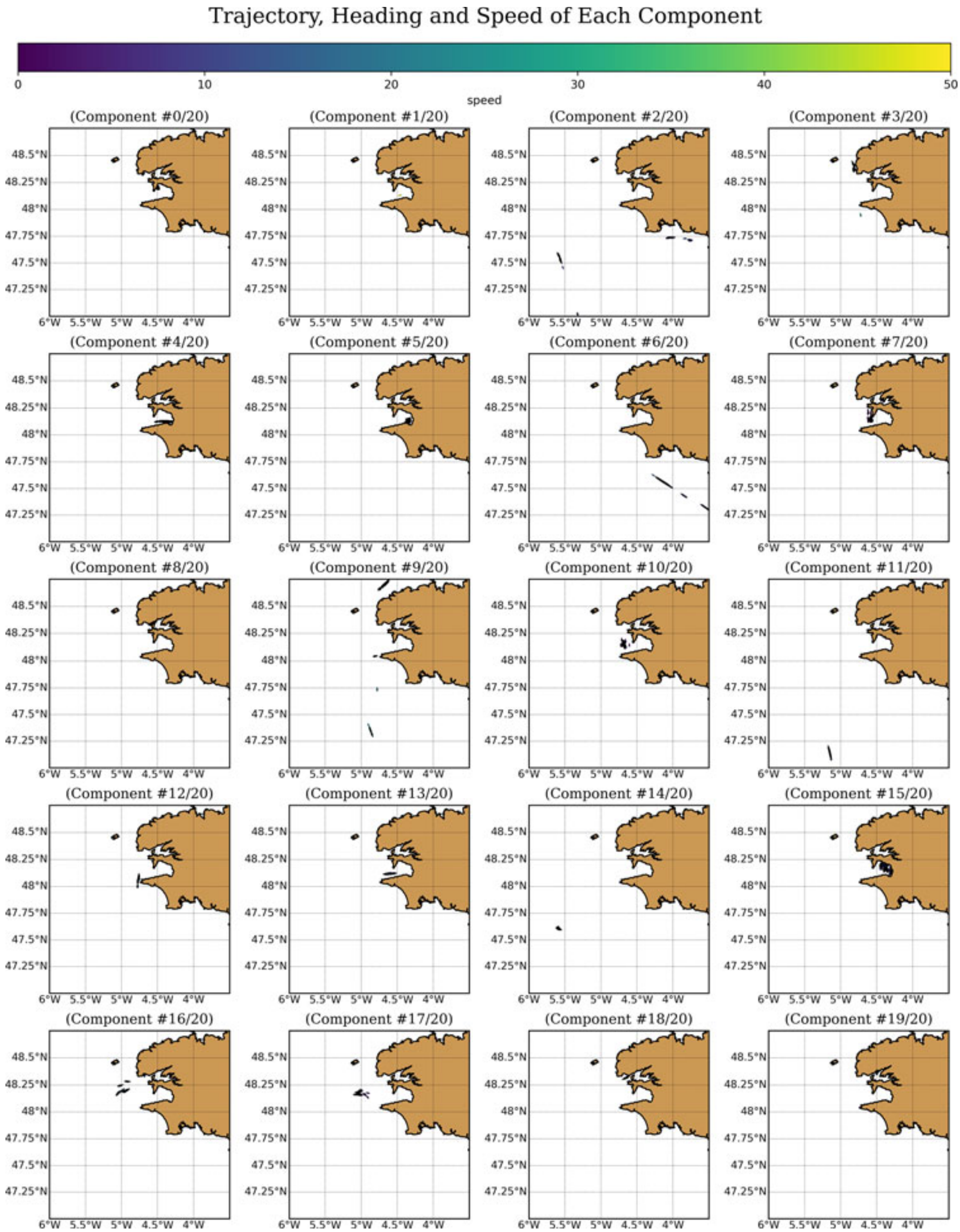


Figure 6. *Navigation pattern results extracted by Gaussian mixture model (the number of components = 20, constructed from the top 300 records).*

4.4.2. Experiment 6: topic model – find similar sails

The experiment results give document-topic matrix **Θ** and topic-code matrix **Φ**. The usages of topic-code matrix **Φ** have been described in Section 3.3 In this subsection, the usages of document-topic matrix **Θ** will be described.

As mentioned in Section 3.3, document-topic matrix **Θ** gives a set of proportions on topics for all sails. this means that similar sails can be derived by calculating the cosine similarity between their topic proportion vector. For illustration purposes, examples of similar sails regarding the topic proportion are shown in Figure 9. By comparing these results, it can be easily confirmed that proportions have top

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

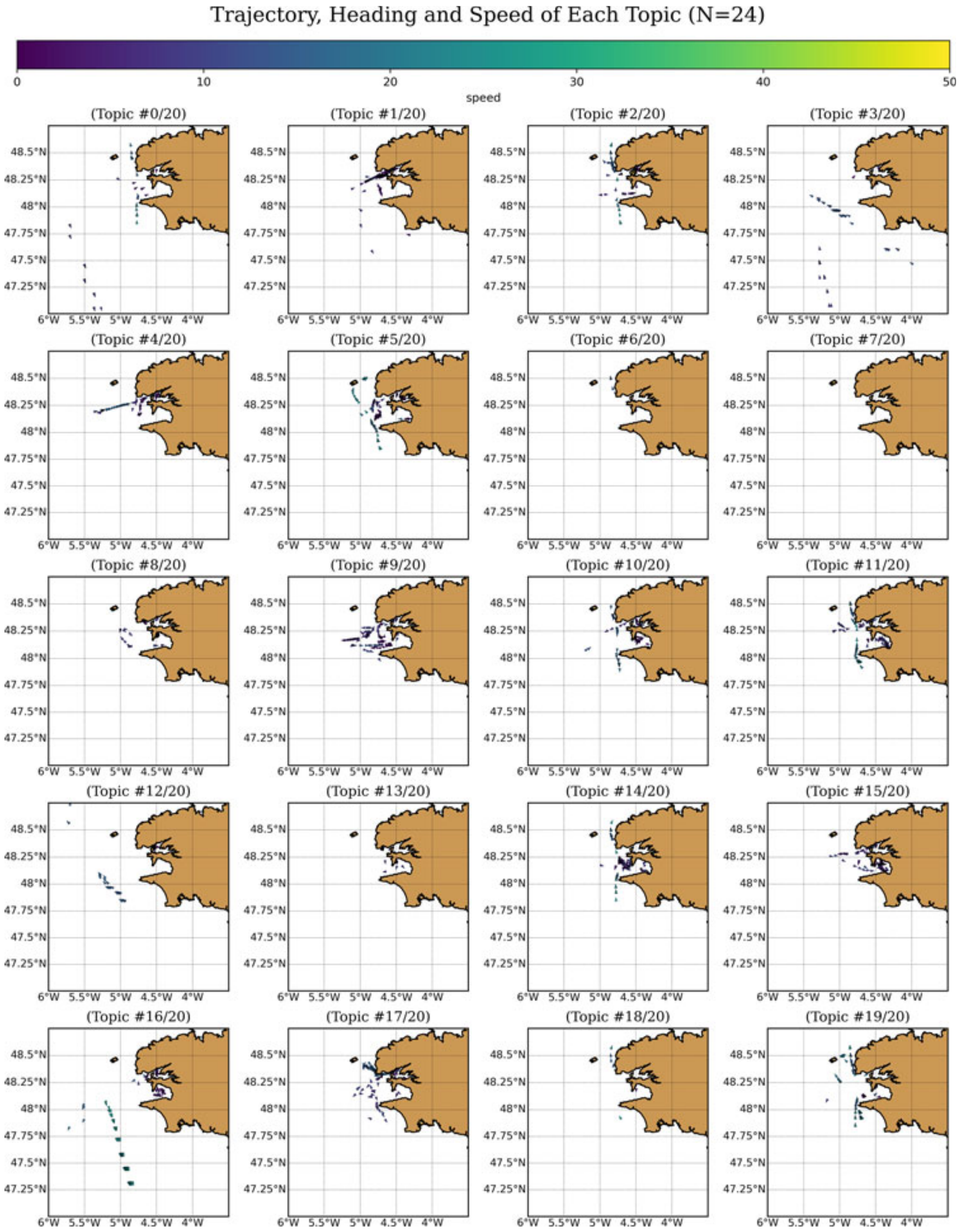
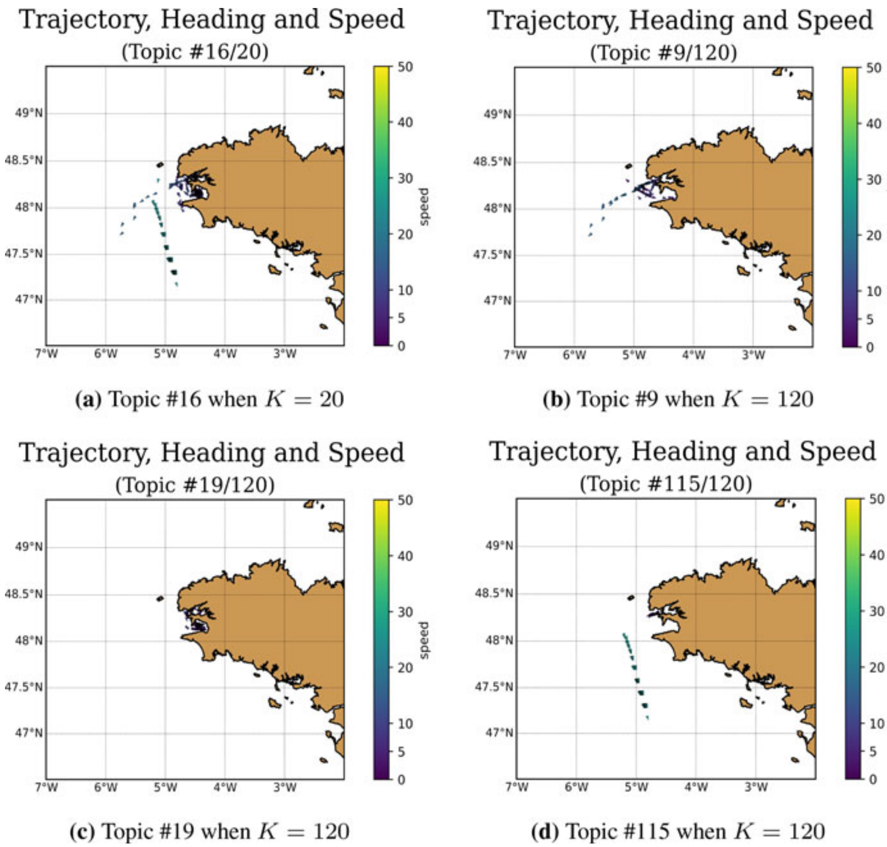


***Figure 7.** Navigation pattern results extracted by the topic model (the number of topics = 20, reconstructed from the top 100 codes).*

values for the same topic number from Figures 9(a) and 9(b), also the trajectory of the pair sails are very similar to each other from Figures 9(c) and 9(d). The significance of this fact can be stated as follows: if a set of topics have been extracted by the topic model, a sail can be represented with a topic proportion vector of the same length as K, instead of its original AIS trajectory data which have very long and uncertain lengths. This will lead to high performance for identifying a specific vessel or finding similar vessels.

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press


Trajectory, Heading and Speed
(Topic #16/20)
Trajectory, Heading and Speed
(Topic #9/120)
Trajectory, Heading and Speed
(Topic #19/120)
Trajectory, Heading and Speed
(Topic #115/120)
49°N
48.5°N
48°N
47.5°N
47°N
7°W
6°W
5°W
4°W
3°W
speed
50
40
30
20
10
5
0

(**a**) Topic #16 when $K = 20$

(**b**) Topic #9 when $K = 120$

(**c**) Topic #19 when $K = 120$

(**d**) Topic #115 when $K = 120$

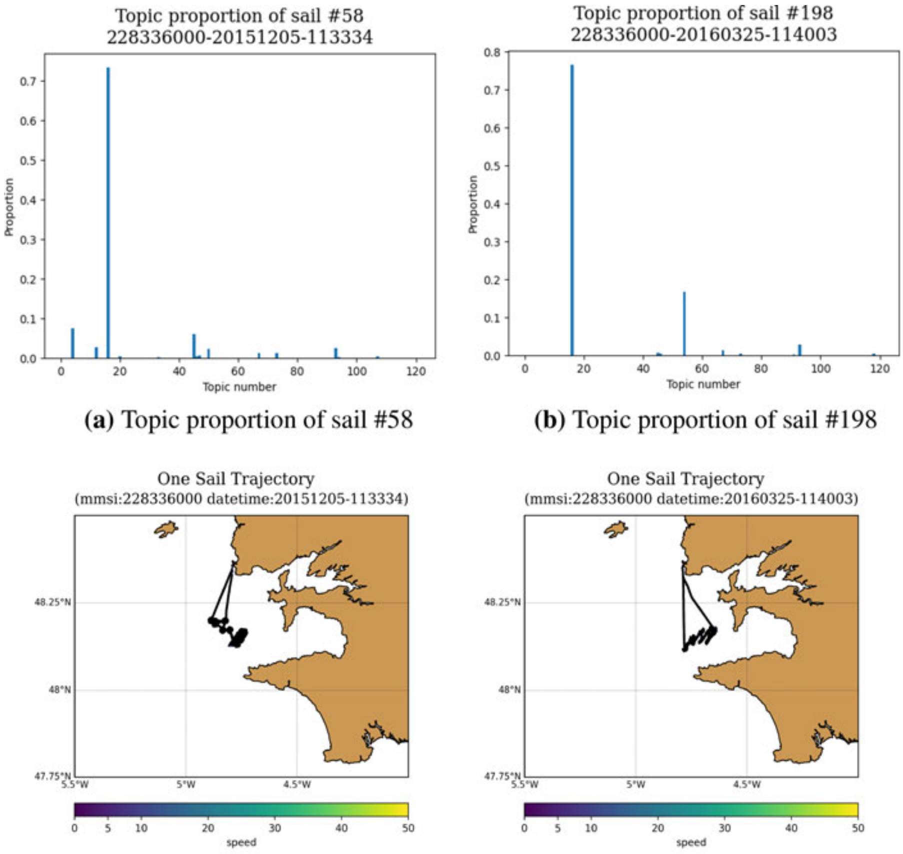
Figure 8. *Comparison of navigation pattern between different numbers of topics. (a) Topic #16 when* $K = 20$*, (b) topic #9 when* $K = 120$*, (c) topic #19 when* $K = 120$ *and (d) topic #115 when* $K = 120$.

4.4.3. Experiment 7: topic model – the number of topics

To evaluate the performance of the topic extraction, a set of topic model experiments with different numbers of topics has been conducted, where the number of topics K has been set to 10, 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 220 and 240. From the topic distribution results of each experiment, the average of Jessen–Shannon divergence between any two different topics for each setting of K has been calculated. Results of this calculation are shown in Figure 10(a), also the elapsed times for each topic model setting are shown in Figure 10(b) together. From Figure 10(a), as the number of topics increases, the average performance makes clear increases at the lower range from 10 to 80, but almost makes no difference at the higher range from 80 to 240. Because Jessen–Shannon divergence represents the differences between topics, a higher topic difference means a better model. The results of Figure 10(a) denote that the performance of the topic model is improved between $K = 10$ and 80, while no more improvements are observed between $K = 80$ and 240. The computation times of the topic model for different K are shown in Figure 10(b). From this result, one can understand that the computation time is proportional to the number of topics. Therefore, considering the computation cost, the best choice of the number of topics should be $K = 80$ for this MovingVessels dataset.

5. Conclusion and future work

This paper introduces a novel approach that models and extracts the patterns that emerge from AIS trajectory big data. By introducing a PQk-means algorithm of vector quantisation, AIS data records can



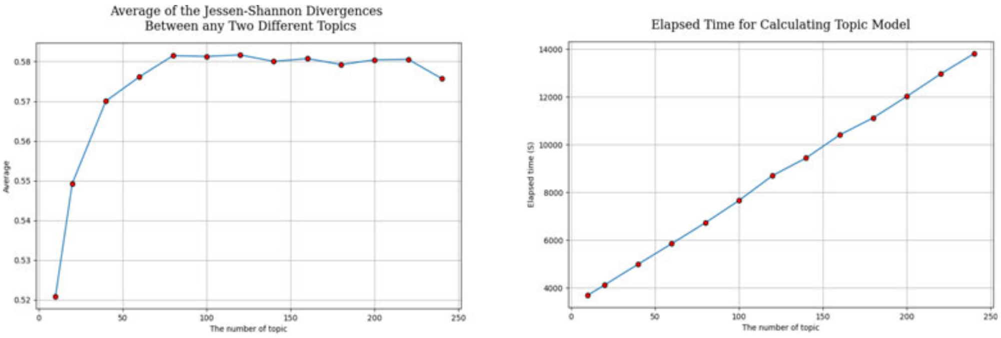
(a) Topic proportion of sail #58

(b) Topic proportion of sail #198

(c) Trajectory of sail #58

(d) Trajectory of sail #198

Figure 9. *(a) Topic proportion of sail #58, (b) topic proportion of sail #198, (c) trajectory of sail #58 and (d) trajectory of sail #198. Cosine similarity between panels (a) and (b) is 0·9677.*



(a) Average of Jessen-Shannon divergences between any two different topics

(b) Elapsed time for calculating topic model

Figure 10. *Performance of topic model according to the number of topics. (a) Average of Jessen–Shannon divergences between any two different topics. (b) Elapsed time for calculating topic model.*

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

be first converted to a set of code documents. By applying a topic model to these code documents, the following two achievements are obtained.

(i) Vessel navigation patterns of vessels are extracted from AIS trajectory data. Compared with the DBSCAN algorithm and Gaussian mixture model, our approach shows more legible and meaningful results.
(ii) Vessels' sails can be represented according to their underlying navigation patterns, which may help to identify specific vessel behaviours by ranking them according to a cosine similarity measure.

Future works will be oriented towards anomaly detection. By aggregating the navigation patterns extracted by the topic model, a knowledge-based comparison can be performed between usual and uncommon navigations. This might favour the detection of navigation anomalies in almost real-time, by comparing sail codes and their degree of deviation from normal navigation patterns. Finally, the proposed approach is not only restricted to exploring AIS big data but also can be extended to many kinds of different data from moving objects, such as aeroplanes, urban vehicles, and human and animal trajectories. More generally, this approach can be adapted to any fixed-dimension vector data recorded by sensor-based distributed frameworks, such as in environment and urban contexts. In this meaning, it can be expected that this approach has great potential for a very wide range of applications.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Number JP20K04943.

References

Avrithis Y., Kalantidis Y., Anagnostopoulos E. and Emiris I. Z. (2015). Web-Scale Image Clustering Revisited. In: *Proceedings of IEEE ICCV*. Araucano Park, Chile: IEEE Computer Society.

Best G. (2011). Satellite-based AIS system provides continuous tracking at sea. *Sea Technology*, **52**(3), 15–17.

Birant D. and Kut A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, **60**, 208–221.

Blei D. M. (2012). Probabilistic topic models. *Communications of the ACM*, **55**(4), 77–84.

Blei D. M., Ng A. Y. and Jordan M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Claramunt C., et al. (2017). Maritime Data Integration and Analysis: Recent Progress and Research Challenges. In: *Proceedings of 20th International Conference on Extending Database Technology (EDBT)*, Venice, Italy.

Commission of the European Communities. (2008). Common position adopted by the Council with a view to the adoption of a Directive of the European Parliament and of the Council amending Directive 2002/59/EC establishing a Community vessel traffic monitoring and information system, COM (2008) 310 final 2005/0239(COD); Brussels, Belgium, 11 June 2008. Available at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri{\mathsurround=\opskip$=$}COM:2008:0310:FIN:EN:pdf.

Dobrkovic A., Iacob M. E. and van Hillegersberg J. (2015). Using Machine Learning for Unsupervised Martime Waypoint Discovery from Streaming AIS Data. In: *Proc. i-KNOW'15*, Graz, Austria, Oct. 2015. Available at: http://dx.doi.org/10.1145/2809563.2809573.

Dobrkovic A., Iacob M. E. and van Hillegersberg J. (2018). Maritime pattern extraction and route reconstruction from incomplete AIS data. *International Journal of Data Science and Analytics*, **5**(6), 111–136. https://doi.org/10.1007/s41060-017-0092-8

Edlund J., Grönkvist M., Lingvall A., Sviestins E. (2006). Rule-based Situation Assessment for Sea Surveillance. In: Dasarathy, B. V. (ed.). *Proceedings of SPIE vol. 6242 Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications 2006.* Orlando, FL: SPIE, 624203.

Fujino I., Claramunt C. and Boudraa A. O. (2017). Extracting Route Patterns of Vessels from AIS Data by Using Topic Model. In: *Proc. IEEE International Conference on Big Data (BIGDATA2017).* Boston, MA: IEEE Computer Society, 4662–4664.

Fujino I., Claramunt C. and Boudraa A. O. (2018). Extracting Courses of Vessels from AIS Data and Real-Time Warning Against Off-Course. In: *Proc. 2nd International Conference on Big Data Research (ICBDR2018).* Weihai, China: Association for Computing Machinery, 62–69.

Gong Y., Pawlowski M., Yang F., Brandy L., Bourdev L. and Fergus R. (2015). Web Scale Photo Hash Clustering on A Single Machine. In: *Proc. IEEE CVPR*. Boston, MA: IEEE Computer Society.

Gray R. M. (1984). Vector Quantization. *IEEE ASSP Magazine*, 4–29.

Hoye G. K., Eriksen T., Meland B. J. and Narheim B. T. (2008). Space-based AIS for global maritime traffic monitoring. *Acta Astronautica*, **62**, 240–245.

IMO (1998). RESOLUTION MSC.74(69) (adopted on 12 May 1998) Adoption of New and Amended Performance Standards. Available at: https://wwwcdn.imo.org/localresources/en/OurWork/Safety/Documents/AIS/Resolution%20MSC.74(69).pdf.

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press

IMO (2000). RESOLUTION MSC.99(73) (adopted on 5 December 2000) Adoption of Amendments to the International Convention for the Safety of Life at Sea, 1974, as Amended. Available at: https://wwwcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/MSCResolutions/MSC.99(73).pdf.

Jégou H., Douze M. and Schmid C. (2011). Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**(1), 117–128.

Laxhammer R. (2008). Anomaly Detection for Sea Surveillance. In: *Proceedings of the 11th International Conference on Information Fusion*, Cologne, Germany.

Likas A., Vlassis N. and Verbeek J. J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, **36**, 451–461.

Linde Y., Buzo A. and Gray R. M. (1980). An algorithm for vector quantiser design. *IEEE Transactions on Communications*, 702–710.

Lloyd S. P. (1982). Least squares quantization in PCM. *IEEE TIT*, **28**(2), 29–137.

Matsui Y., Ogaki K., Yamasaki T. and Aizawa K. (2017). PQk-means: Billion-Scale Clustering for Product-Quantised Codes. In: *Proceedings of the 25th ACM International Conference on Multimedia.* Mountain View, CA: ACM Computer Society, 1725–1733.

MBDW (2020). *2nd Maritime Big Data Workshop*. https://sites.google.com/view/mbdw2020

Pallotta G., Vespe M. and Bryan K. (2013). Vessel pattern knowledge discover from AIS data: a framework for anomaly detection and route prediction. *Entropy*, **15**, 2288–2315.

Piciarelli C., Nicheloni C. and Foresti G. L. (2008). Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for Video Technology*, **18**, 1544–1554.

Ray C., Grancher A., Thibaud R., Etienne L. (2013). Spatio-Temporal Rule-based Analysis of Maritime Traffic. *Third Conference on Ocean & Coastal Observation: Sensors and Observing Systems, Numerical Models and Information (OCOSS)*, Nice, France. hal-01627352.

Ray C., Dréo R., Camossi E. and Jousselme A. L. (2018). Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance (0.1) [Data set]. Zenodo. Available at: https://doi.org/10.5281/zenodo.1167595.

Ristic B., La Scala B., Morelande M. and Gordon N. (2008). Statistical Analysis of Motion Patterns in AIS Data: Anomaly Detection and Motion Prediction. In: *Proceedings of the 11th International Conference on Information Fusion*. Cologne, Germany.

U.S. Coast Guard Navigation Center (2000). *AIS Requirements*. https://www.navcen.uscg.gov/?pageName{\mathsurround=\opskip$=$}AISRequirementsRev#.

Vries G. K. D. and van Someren M. (2012). Machine learning for vessel trajectories using compression, alignments and domain knowledge. *Expert Systems with Applications*. http://doi.org/10.1016/j.eswa.2012.05.060

Zhang B., Hirayama K., Ren H., Wang D., Li H. (2023). Ship anomalous behavior detection using clustering and deep recurrent neural network. *Journal of Marine Science and Engineering*, **11**, 763. https://doi.org/10.3390/jmse11040763

Cite this article: Fujino I, Claramunt C (2023). Navigation pattern extraction from AIS trajectory big data via topic model. *The Journal of Navigation* **76**: 4–5, 506–524. https://doi.org/10.1017/S0373463323000206

https://doi.org/10.1017/S0373463323000206 Published online by Cambridge University Press