A practical AIS-based route library for voyage planning at the pre-fixture stage

Jie Cai \(^{a, *}\), Gang Chen \(^{b, a}\), Marie Lützen \(^{c}\), Niels Gorm Maly Rytter \(^{a}\)

\(^{a}\) Department of Technology and Innovation, University of Southern Denmark, Campusvej 55, DK 5230, Odense M, Denmark
\(^{b}\) World Maritime University, Malmö, Sweden
\(^{c}\) Department of Mechanical and Electrical Engineering, University of Southern Denmark, Campusvej 55, DK 5230, Odense M, Denmark

**A B S T R A C T**

In tramp shipping, a preliminary route is required for voyage planning at the pre-fixture stage (before a chartering contract is agreed). Such routes are conventionally designated by using pilot charts or software considering long-term statistical weather. However, it has been experienced by tramp operators that such route solutions often poorly estimated sailing distances for long journeys and thereby cause inappropriate cost estimation and bad voyage plan. To fill this gap, a data-driven methodology is proposed in this paper to establish a practical route library with the consideration of ship sizes, load conditions and seasonality. In this method, it first requires a dividing of ship trajectories into local sea passage and open sea passage. The voyage trajectories made of AIS points are then simplified to pattern nodes based on a speed-weighted geolocation method. Afterwards, the KMeans algorithm is deployed to properly classify these pattern nodes, identifying the most representative nodes (routes) in open sea passages. Simultaneously, the connection points are identified by DBSCAN algorithm, representing local sea passages. Combining the representative routes in open sea passages and the connection points in local sea passages, the most navigated routes between two ports are obtained. Finally, case studies are conducted for the Pacific Ocean and the Atlantic Ocean respectively using global AIS data from tanker vessels to demonstrate the feasibility and effectiveness of this methodology. The proposed route library is capable of providing reliable route references to support the decision-making at the pre-fixture stage.

**1. Introduction**

Nowadays, shipping is the most carbon-efficient form of transportation, delivering 80% of all goods globally (Metcalf et al., 2018). In tramp shipping, when a cargo is to be transported, the shipping company must estimate the voyage cost based on a preliminary route and bid for a contract. A misguided route may substantially impair their competitiveness as a consequence. Conventionally, reference routes are designated by using pilot charts or software considering long-term statistical weather (Brooks and Faust, 2018). In this way, uncertainties will always be introduced due to human factors such as the insufficient experience of staff in the chartering department, especially for route planning crossing oceans with unknown regions. As a consequence, a poor estimation of sailing distances may be produced for long journeys and thereby cause problems such as extra sailing time and fuel consumptions when executing further operation, according to tramp operators. Therefore, an alternative method is required to facilitate the preliminary route planning at the pre-fixture for shipping industry.

The historical voyage trajectories have great potential for improving shipping transportation, but due to lack of data, they have not been fully utilized. Thanks to the Automatic Identification System (AIS), the voyage data have been constantly accumulated. The AIS is originally developed as a maritime safety and traffic surveillance system (Ball, 2012; Loretta, 2016; Svanberg et al., 2019). According to the International Association of maritime Aids to Navigation and Lighthouse Authorities (IALA), the original purpose of the system was to “improve the maritime safety and efficiency of navigation, safety of life at sea and the protection of the maritime environment” (IALA, 2003). Requirements and plans for implementation of AIS system are outlined in the International Convention for the Safety of Life at Sea (SOLAS) (IMO, 2012). The regulation requires the AIS system to be fitted aboard all ships of 300 gross tonnage and upwards engaged on international voyages, cargo ships of 500 gross tonnage and upwards not engaged on international voyages and all passenger ships irrespective of size. Thus,
a massive quantity of AIS data are accumulated which are able to facilitate not only practical vessel operations but also research activities in the maritime domain (Kotovirta et al., 2009; Le Tixerant et al., 2018). Each AIS ping consists of features to describe the current status, which is normally categorized as static information, semi-static information and dynamic information (Ou and Zhu, 2008). Static information – i.e. ship name, ship size, build year and others – includes information programmed in advance into the AIS system when commissioning the vessels, while dynamic information – e.g. longitude, latitude, speed, course, time stamp and others – is the real time information based on sensors. The semi-static information is the voyage related data such as the sailing distance, loading condition, destination and others, which may be manually added and modified.

The utilizing of AIS data on maritime research has largely proved itself in past decades (Le Tixerant et al., 2018; Svanberg et al., 2019). The majority of research has focused on maritime traffic analysis in narrow and congested waterways since the AIS from the beginning was based on VHF (very high frequency) signals — not satellites. Schelmedine (Shelmerdine, 2015) conducted an initial analysis of AIS data within one year in Shetland region. Zhang et al. (2019) extracted turning nodes and then built navigation routes network using these turning points in the Qiongzhou Strait based on AIS data. Chen et al. (2015) extracted a principal fairway in Taiwan Strait. Altan and Otay (2017) studied the maritime traffic of ships at the Strait of Istanbul based on data over one year. Different data features were statistically analysed, including the number of ship distribution, ship types, speeds and courses. Several researchers have investigated trajectory similarities based on AIS data using different techniques (Gonzalez et al., 2007; Chen et al., 2011) such as machine learning (Granlund and Knutsson, 2013; Jelinek, 1997; Bishop, 2006; LeCun et al., 2015; Agamennoni et al., 2010; Shalev-Shwartz and Ben-David, 2014) and Dynamic Time Warping (DTW) (Myers et al., 1980) to identity similarities. Ismail and Vigneron (Ismail and Vigneron, 2015) proposed the Merge Distance (MD) algorithm for the measuring of similarity between route trajectories. Li et al. (2018) measured the trajectory similarity through MD, and then re-sampled a representative dataset based on the similarity matrix. The new representative dataset was clustered by DBSCAN (Ester et al., 1996; Schubert et al., 2017) to obtain points for navigated routes. Han and Yang (2020) calculated the route similarities accounting for voyage directions and vessel geographical locations.

However, the investigation of spatial voyage routes in open seas with long legs (larger than 3000 nm) is rare. The voyage route selections related to satellite-based AIS data are quite new. Meanwhile, the shipping industry has been searching for different ways of selecting a proper navigated route at the pre-fixture stage for a better commercial bidding. An option is to simply use the shortest route in a region, but it may result in a route which is most likely too short and thereby causes too low cost estimation. Alternatively, the most navigated routes, representing typical historical voyage patterns, should be proposed for practical utilizing according to the needs of shipping companies. Therefore, in this paper, a route selection method is developed together with ML techniques based on AIS data in order to select the most navigated routes, aiming to facilitate the decision-making at the pre-fixture stage. A route library is established based on global AIS data of tankers within a five years period.

The structure of this paper is organized as follows. In Section 2, we construct a methodology for the identification of the most navigated routes using AIS data. In Section 3 we do scoping, cleaning, preparation and detailed investigation of five years of tanker shipping AIS data. Based on the analysis, we in Section 4 propose grouping of ships and voyages to improve the model performance. In Section 5 we present and discuss the final results of two cases and the performance of our models, and derive a statistical distribution of the most navigated routes for specific subgroups of tankers and voyages depending on seasonality, size and loading condition. Finally we draw the conclusions and propose suggestions for further research.

2. Methodology for constructing an AIS based route library

In order to construct an AIS based route library for practical use, a data-driven methodology is developed in this paper. Due to the route structures suggested by the proposed method, all voyages are firstly divided into two subsets, namely, the open sea passage and the local sea passage (it can be further divided into departure and destination local sea passages). Details will be elaborated in Section 3.2 using real AIS datasets. To extract the open sea passages, a speed-weighted geolocation method is proposed, simplifying voyage trajectories into pattern nodes. The KMeans algorithm is then deployed to classify the pattern nodes so as to find the representative routes in open sea passages. Simultaneously, for the local sea passages, the AIS data are processed by DBSCAN algorithm to represent the route pattern with connection points. The most navigated routes between ports are identified by combining the open-sea representative routes and the local connection points. Therefore, the data-driven methodology and its general application steps for the identification of the most navigated routes are first described in Section 2.1 under this logic. The detailed descriptions of respective techniques adopted in this method are then presented from Section 2.2 to Section 2.5.

The terminology used in this paper is described as follows. The pattern nodes denote the extracted nodes which can represent voyage trajectories. The connection points are defined as merging points of multiple voyage trajectories, which can not only connect all open sea passages in common but also can be used for local distance estimate.

2.1. Data-driven methodology and its application procedures

The proposed data-driven methodology for route identification consists of a speed-weighted geolocation method to form pattern nodes (see Section 2.2) and two types of ML techniques (DBSCAN and KMeans, see Section 2.3). Fig. 1 is the flowchart to illustrate the methodology and its application procedures. The deployed techniques are highlighted in red. Note that, in this methodology, we simply apply the ML techniques without any improvements of the algorithms themselves since the current patterns are sufficient to identify routes from the standpoint of practice. All dataset should be divided into local sea passages and open sea passages before training (see Fig. 6 in Section 3.2 for concrete cases). Detailed description of the flowchart is presented as follows.

- Connection points in local sea passages: Train connection points in local sea passages by DBSCAN and update the points when needed;
- Navigated routes in open sea passages: Compute pattern nodes based on speed-weighted geolocation method, then extract navigated routes by KMeans, finally find the nodes (voyages) with the minimum distances close to cluster centroids;
- The most navigated routes: Combine the navigated route in open sea and its corresponding connection points in local sea between journey departure and destination;

Based on the most navigated routes, relevant voyage information such as journey distance can be accurately estimated. Note that the training of connection points in local sea passages and the identification of navigated routes in open sea passages can be conducted simultaneously.

2.2. Speed-weighted geolocation method

Since each voyage trajectory consists of a series of AIS data points, a pattern node should be extracted in order to standardize trajectory representation. Unlike other existing methods such as DTW and MD to directly trace trajectory similarity, the proposed method first simplifies the trajectory (a two-dimensional (2D) spatial line) into a node (one-dimensional spatial point), accounting for the geographical positions
Fig. 1. The data-driven methodology and its general application procedures for the most navigated routes.

Fig. 2. Schematic image of the speed-weighted geolocation method.

of each AIS data point \((X_{ij} \text{ and } Y_{ij})\) and vessel speeds \((V_{ij})\). The coordinates of pattern nodes \((Z_{iY} \text{ and } Z_{iX})\) are calculated to represent the \(i\)th 2D voyage trajectory, as expressed by Eq. (1). \(X_{ij}\) denotes the normalized \(j\)th AIS position of vessels in terms of latitude in the \(i\)th single voyage, while \(Y_{ij}\) denotes the normalized \(j\)th AIS position of vessels in terms of longitude in the \(i\)th voyage. The overall number of AIS points in the \(i\)th voyage is \(N_i\). Note that geographical locations are normalized by their extreme values within the dataset. The aim of using speed as a weight is to account for the speed variation during a voyage and also to increase the data size in terms of distribution in the same sea region so as to have a better identification by KMeans. As illustrated in Fig. 2 for example, a ship can sail from connection point 1 (CP1) to connection point 2 (CP2) with four different options, indicated by different colours. Using this method, the corresponding pattern nodes of four voyages are A, B, C and D, respectively.

\[
\begin{align*}
Z_{iY} &= \frac{\sum_{j=1}^{N_i} (Y_{ij} \cdot V_{ij})}{N_i} \\
Z_{iX} &= \frac{\sum_{j=1}^{N_i} (X_{ij} \cdot V_{ij})}{N_i}
\end{align*}
\]  

(1)

In this way, all the pattern nodes of ship trajectories in an open-sea area can be computed, forming a new dataset representing all possible routes. Afterwards, with the new dataset, the KMeans algorithm (Shalev-Shwartz and Ben-David, 2014) is adopted to find similar clusters and the node close to the centre of each cluster would be selected as one of the navigated routes. Euclidean distance is used to find the centre points.

It should be noted that a limitation exists for the application of this method. For instance, the number of AIS points for routes in the same open-sea regions are not often identical with varied speeds. Theoretically, excessively large speeds may separate such pattern nodes into different clusters and thereby cause bad or wrong route selections. However, in practice, it rarely happens since the average sailing speeds are quite similar. One remedy is to remove all speed outliers during data cleaning. Meanwhile, we further divide original dataset into several subsets so that such discrepancies may be reduced, as described in Section 3.2.

2.3. Unsupervised machine learning algorithms

The unsupervised machine learning (ML) algorithms KMeans and DBSCAN are deployed together with the speed weighted geolocation method to find the most navigated routes. A brief description of these two algorithms are presented in this section.

2.3.1. KMeans

The KMeans algorithm (Shalev-Shwartz and Ben-David, 2014) attempts to minimize the average squared distance of the representative
object to all the other objects \((x)\) in the same clusters, as expressed by Eq. (2). In this objective function, the trained dataset is partitioned into sets \(C_1, \ldots, C_k\), where \(\mu_i\) is the centroid of the partitioned clusters \(C_i\), \(d(x, \mu)^2\) is the square distance of each point \((x)\) in a given set to the centroid of its cluster.

\[
L_{\text{KMeans}} = \min \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2
\]

(2)

The number of clusters \(k\) is the parameter needed in a KMeans algorithm. In practice, it is often up to the user to choose the parameter \(k\) that is the most suitable number for given clustering problems. A simple iterative algorithm is normally applied to minimize the KMeans objective cost. In the iterative procedure, the \(k\) value is first given so as to partition the overall dataset. Initialization is then performed with a randomly chosen centroids \(\mu_1, \ldots, \mu_k\). For any points in given clusters, the distance between points and their centroids are calculated. Update of the centroid is iteratively performed based on the summation of all the distance until the solution is converged. In this paper, KMeans algorithm is used to select the representative navigated routes by finding clusters and their centres.

2.3.2. DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) algorithm is a density-based unsupervised clustering algorithm initially proposed by Ester et al. (1996). It has been widely implemented and is available in popular toolkits such as scikit-learn (Pedregosa et al., 2011). A simple density level estimation is used in DBSCAN based on a threshold for the number of neighbours, \(\text{minPts}\), within a given radius \(\epsilon\). Three basic concepts including core points, border points and noise are defined through the points interaction relationship such as direct density reachability, density reachability and others, as illustrated in Fig. 3. In this case, the parameter \(\text{minPts}\) is 4, and the radius distance \(\epsilon\) is indicated by all the circles. Point N is a noise point because it is not density reachable from any other points. Point A is a core point, while points B and C are border points. The black arrows indicate direct density reachability. Points B and C are only density connected, since both of them are density reachable from A.

Based on such concept definitions, the basic algorithm of DBSCAN can be simply described: the database is first linearly scanned for objects not been processed. The non-core points are assigned to noise. When a core point is discovered, its neighbours are iteratively expanded and added to a cluster. When encountered later by the linear scan, objects that have been assigned to a cluster will then be skipped. Border points are normally assigned to the first cluster they are reachable from in the original algorithm. Thus the result of applying such DBSCAN algorithm is deterministic, but may change a little if the dataset is permuted. DBSCAN is used in this paper to train the connection points in local sea passages.

In order to evaluate the training accuracy of model clustering by ML techniques, the metric of Silhouette coefficient (Rousseeuw, 1987) is used, which is a measurement of how similar an object is to its own cluster compared to other clusters. The score ranges between \(-1\) and \(1\), where a high value indicates that the clustering configuration is appropriate. From the standpoint of practice, the score approaching 0.4 is considered acceptable in this paper.

3. AIS data scoping, modelling and analysis

Two original AIS datasets of tankers are explored in the present analysis. It includes one with daily recordings from January 2014 to March 2019 and the other with hourly recordings from December 2018 to September 2020. Data details containing AIS static information (e.g., ship name and ship size), dynamic information (e.g., speed, longitude and latitude) and semi-static information (voyage distance and destinations) are available for use. The datasets are first described, cleaned and visualized. Then, dividing strategy of data is elaborated, preparing for further route selections by the methodology introduced in Section 2. The analyses of influential factors of shipping are conducted in Section 3.3.

3.1. Data scoping, cleansing and preparation

In order to better explore data, the entire datasets are visualized, as seen in Figs. 4 and 5. Fig. 4 illustrates the density map of all the tanker voyages through daily AIS dataset within 5 years. In this route library, the voyages with long leg larger than 3000 nm are the only focuses. Thus, any voyages less than 3000 nm will be filtered in advance. There are main regions with high voyage density, such as voyages from US to Europe and voyages from Asia to Gulf. More than 500 000 voyages from 4600 ships are finally included in the dataset, which consists of six million daily positions and their corresponding information. The number of terminals/ports are 3288 in total, which have been roughly classified into 518 main port and 25 main regions. Meanwhile, the high-frequency AIS dataset with hourly-based recordings of a selected tanker fleet is available, as seen in the density map in Fig. 5. This dataset has for all voyages 776 000 hourly positions and is only used to identify connection points of local sea passages. For the sake of accuracy, no filter of voyages is needed.

Missing data such as the journey distance are also fixed by recalculation based on the voyage geographical locations. Furthermore, wrongly assigned terminal names are observed due to the mistaken input by crews. Duplicates terminals have been found in different countries. For instance, the terminal name of “Damen Ship Repair” belongs to either the Netherlands or France, which has introduced ambiguity in the data grouping. All these observed data problems are fixed before further utilizing.

3.2. Data division into local versus open sea passages

For every major region, the AIS data are manually divided into three subsets as described in Section 2, namely, the departure local sea passage, the open sea passage and the destination local sea passage. Specifically, the subset of open sea passages is only used for the selection of the navigated routes. The subsets of local seas passages are used to train common connection points. Note that the data in local sea passages are not the daily AIS positions but the hourly high-frequency data in order to identify connection points with high accuracy. As a result, the two datasets should have the same cutting boundary so that the final selected voyages would not overlap. The determination of the cutting boundary is flexible based on the observation of data distribution and the shape of local territories, since it will not affect too much on the method application.
As seen in Fig. 6, two cases are presented to elaborate the data division strategy, including the left one between Gulf and East & Northeast Asia and the right one between Gulf and West Europe. The local passages are indicated by the black boxes, while the open sea passages are the rest of the voyages. It has been observed that the voyages in local passages have regular patterns with common connection points determined by the density of voyages distribution, as shown by the red dots. Combining the selected open sea route with the extracted connection points, the most navigated routes in a given region are found. Examples of them are also highlighted, as seen from route A to route D.

3.3. Analysis of parameters influencing ship routes

In this section, we investigate the impact of seasonality, ship size and loading condition on executed routes of voyages based on AIS data in open sea passages. The clarifying of their effects on the voyage execution facilitates a better data grouping to identify the most navigated routes.

3.3.1. Vessel seasonality analysis

As stated by Shelmerdine (2015), processing AIS data on a monthly basis is found to be highly beneficial as it allowed greater flexibility for future outputs and provided a manageable quantity of data for analysis.
Hence, the dataset is also analysed monthly to consider the seasonality effect, searching for the tendency of route variations.

Fig. 7 shows an example of the seasonal distribution of voyages from the Gulf region to the East & Northeast Asia region across the Pacific Ocean. The AIS data in local sea passages are intentionally removed. Four temporal periods are aggregated by months, including December to February, March to April, May to August and September to November. It is found that crews chose more routes towards the south in winter time in the Pacific Ocean (e.g., voyages from December to February) compared with summer time (e.g., voyages from May to August). Voyage distribution in the period from March to April presents a pattern similar to the one from December to February. Routes from September to November gradually moved to the northern part of the ocean, presenting heavy density in the middle region close to Hawaii islands. Therefore, it can be briefly concluded that the route choice pattern varies with seasonality. Dataset should be grouped based on such voyage differences in different temporal periods.

3.3.2. Vessel loading condition analysis

The change of loading conditions during vessel operations will affect their seakeeping ability (Faltinsen, 1993). The operational draft of vessels varies during sailing, affecting the voyage routes in seas when encountering different sea states. However, the AIS dataset only roughly indicates whether a voyage is in a laden condition or a ballast condition. Hence, the effect of draft variation will not be considered during sailing. Instead, only the two loading conditions are used, indicated by “1” for laden condition and “0” for ballast condition in the AIS dataset, respectively. Note that the loading condition is a parameter which crews manually put into system during sailing, therefore data errors are unavoidable as a consequence.

Fig. 8 illustrates the distribution of route patterns between Gulf and East & Northeast Asia region in different loading conditions, where the orange points denote all voyages in laden condition and the red points denote all voyages in ballast condition. It is found that, the majority of data are labelled as laden condition, for instance, 68% of voyages from East & Northeast Asia to Gulf and 99.9% of voyages from Gulf to East & Northeast Asia. Therefore, no further data grouping is conducted for voyages in ballast condition due to limited number of data.

3.3.3. Ship size analysis

The size of vessels in terms of cargo carrying capacity may affect the route selections in practical journey, especially when encountering bad weather. In practice, there are two basic types of oil tankers (Hsu et al., 2015): crude tankers and product tankers. Based on the further recommendation of shipping companies in this project, seven standard ship types (SMALL, MR1, MR2, HANDY, LR1, LR2 and AFRAMAX) are adopted without considering the specific vessel functions, whatever crude or product tankers. Each type corresponds to a specific size range based on their carrying capacities, for instance, LR1 type is between 60 000 DWT and 84 999 DWT.

In the dataset, the overall tanker size ranges between 25 000 dead weight metric ton (DWT) and 120 000 DWT. As seen in the second column of Table 1, the raw information of ship types from dataset is listed, including 21 different types in total. In order to consider ship size in a manageable way so that the identified results can be generalized in practice, the original vessel types have been updated into the corresponding standard types. Note that, during this process, the vessel types with few number of data are manually changed into other similar type with sufficient number of data, e.g., PANAMAX MT (DIRTY) to LR1 standard type for the sake of model accuracy. Due to the limited data scale in each region, we further aggregate the standard vessel categorizes based on data analysis. As seen in Fig. 9, it illustrates the effect of ship size on the variations of voyage selections. Three types (LR2, MR2 and SMALL) are visualized for comparison in the Atlantic Ocean. It is observed that the larger vessels (e.g., LR2) are more prone to use stable routes in regions close to the northern ocean compared with the smaller ones (e.g., SMALL). Some of the routes in SMALL and MR2 types are gradually moving to the southern region, presenting larger scatters in distribution compared with the LR2 type. This phenomenon demonstrates the weak seakeeping performance and relatively well manoeuvrability of vessels in smaller size, trying to following a more suitable routes when encountering adverse factors in seas, such as bad weather. Therefore, it is reasonable for us to recategorize the vessel type. Due to limited data scale, only two new categories of ship type are defined for further route selections in this paper, namely, Large type (incl. AFRAMAX, LR2 and LR2, roughly between 60 000 DWT and 125 000 DWT) and Small type (incl. SMALL, HANDY, MR1 and MR2, roughly between 25 000 DWT and 59 999 DWT).

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Original and simplified ship type categories.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>Original ship type</td>
</tr>
<tr>
<td>1</td>
<td>MR2 TNKR (Chem / Product) DW</td>
</tr>
<tr>
<td>2</td>
<td>HANDY CHEMICAL PARCEL TNKR</td>
</tr>
<tr>
<td>3</td>
<td>SMALL CHEMICAL PARCEL TNKR</td>
</tr>
<tr>
<td>4</td>
<td>LR1 TNKR PR</td>
</tr>
<tr>
<td>5</td>
<td>MR2 TNKR PR</td>
</tr>
<tr>
<td>6</td>
<td>MR2 TNKR (Chem / Product) PR</td>
</tr>
<tr>
<td>7</td>
<td>LR1 TNKR DW</td>
</tr>
<tr>
<td>8</td>
<td>HANDY CHEMICAL TNKR (Methanol)</td>
</tr>
<tr>
<td>9</td>
<td>HANDY CHEMICAL TNKR</td>
</tr>
<tr>
<td>10</td>
<td>MR2 TNKR (Parcel) DW</td>
</tr>
<tr>
<td>11</td>
<td>MR1 TNKR (Chem / Product)</td>
</tr>
<tr>
<td>12</td>
<td>SMALL CHEMICAL TNKR</td>
</tr>
<tr>
<td>13</td>
<td>MR2 TNKR DW</td>
</tr>
<tr>
<td>14</td>
<td>PANAMAX MT (DIRTY)</td>
</tr>
<tr>
<td>15</td>
<td>LR1 TNKR (Chem / Product) PR</td>
</tr>
<tr>
<td>16</td>
<td>LR2 TNKR</td>
</tr>
<tr>
<td>17</td>
<td>LR1 TNKR (Chem / Product) DW</td>
</tr>
<tr>
<td>18</td>
<td>SMALL CLEAN TNKR (Chemical / Product)</td>
</tr>
<tr>
<td>19</td>
<td>POST PANAMAX TNKR (CLEAN)</td>
</tr>
<tr>
<td>20</td>
<td>MR1 TNKR</td>
</tr>
<tr>
<td>21</td>
<td>AFRAMAX TNKR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Data grouping for Atlantic Ocean voyages.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>Temporal periods</td>
</tr>
<tr>
<td>G1</td>
<td>May to August</td>
</tr>
<tr>
<td>G2</td>
<td>May to August</td>
</tr>
<tr>
<td>G3</td>
<td>September to April</td>
</tr>
<tr>
<td>G4</td>
<td>September to April</td>
</tr>
<tr>
<td>G5</td>
<td>All</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Data grouping for Pacific Ocean voyages.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>Temporal periods</td>
</tr>
<tr>
<td>G1</td>
<td>May to August</td>
</tr>
<tr>
<td>G2</td>
<td>May to August</td>
</tr>
<tr>
<td>G3</td>
<td>September to November</td>
</tr>
<tr>
<td>G4</td>
<td>September to November</td>
</tr>
<tr>
<td>G5</td>
<td>December to April</td>
</tr>
<tr>
<td>G6</td>
<td>December to April</td>
</tr>
<tr>
<td>G7</td>
<td>All</td>
</tr>
</tbody>
</table>
4. Prioritizing voyage groups for further route investigation

In this section, the voyages in two major oceans (the Pacific Ocean and the Atlantic Ocean) are grouped based on the data analysis results in Section 3. As a result, there are two major grouping systems for voyages in the Pacific Ocean and voyages in the Atlantic Ocean respectively. The difference is largely due to the different weather patterns observed from data analysis. For voyages in some regions without sufficient data, no grouping strategy is deployed for the sake of accuracy (defined as G0), such as the one between Gulf and South-west Europe. The rest of grouping strategies in two different oceans is listed in Tables 2 and 3, respectively.

5. Case study results and discussions

In this section, we use two specific cases in different ocean regions for the identification of the most navigated routes based on the method described in Section 2. Fig. 10 illustrates the cleaned AIS data distribution of the two cases, including the one with voyages from Gulf to East & Northeast Asia (a size of 2663 AIS points), and the other from the North Sea & West Europe to Gulf (a size of 7267 AIS points). Data in both local regions and other irrelevant domains are filtered. Data grouping strategies are applied in advance accounting for the influential parameters, as described in Section 4. The trained clusters by DBSCAN in Gulf local region are showed and discussed, from where connection points are extracted. The identified navigated routes in the open sea passages from the two cases are illustrated. Besides, the results of clustering data through KMeans are showed so as to demonstrate the effectiveness and feasibility of this selection method.

5.1. Identification of connection points

Connection points are not only served as an important role to tie the selected navigated routes in open sea passages and local regions, but also can be used for local distance calculation. Using DBSCAN, a hyperparameter tuning is first conducted in order to select the best parameters for model training. Fig. 11 illustrates the comparison diagram of Silhouette scores accounting for different combinations of minPts and eps (ε). The considered minimum number of points within a cluster (minPts) is between 1 and 250 with an increment of 1 in each training, while the eps distance is between 2 km and 14 km with an increment of 2 km in each training. It is found that the highest accuracy in terms of the Silhouette score is 0.335 in Gulf region, as indicated by the P point. Hence, the corresponding minPts value of 111 and the eps (ε) value of 14 km are selected for models’ training. The Haversine metric is adopted to calculate the great-cycle distance between two points on a sphere.
Fig. 9. Distribution of route patterns of voyages from Gulf to West Europe region in different ship sizes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 10. The cleaned AIS data distribution in two major cases for routes identification: G5 from Gulf to East & Northeast Asia in Pacific Ocean (left) and G2 from North Sea & West Europe to Gulf in Atlantic Ocean (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 11. The comparison diagram of Silhouette scores accounting for different minPts and eps (\(\epsilon\)). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 12 illustrates the clusters which have been trained for the local sea passages in Gulf through the DBSCAN. The tiny black dots in the left image denote all the noises not belonged to any clusters, while all the clusters are in colours. In the right image, the geographical locations of centroids of all clusters are extracted, considering as connection points (e.g., A, B, ..., F). The arrows indicate the possible directions local voyages could pass through. Note that the original data distribution in local region has multiple and varying densities, which compromises the training accuracy of clusters. Therefore, the route selection cannot simply rely on the training results of algorithm, instead, practical experience is needed to identify major connections (e.g., Panama canal) so as to find all effective connection points. Meanwhile, some clusters close to terminals/ports are irrelevant, which have been manually removed.

For identifying connection points, other unsupervised learning algorithms could also be applied, which we have not discussed in detail in this paper. For instance, the KMeans algorithm could be adopted but it is obviously impractical for this case due to the requirement of prescribed number of clusters, as described in Section 2.3.1. Another simple simulation with Balanced Iterative Reducing and Clustering
using Hierarchies (BIRCH) algorithm (Pedregosa et al., 2011) has been conducted. The BIRCH is a clustering algorithm which summaries large datasets into small and dense regions (called Clustering Feature entries). In this way, it will provide an efficient running time when the size of datasets increases. Two critical parameters are needed during training, namely, the threshold which defines the maximum number of entries in each leaf node and the branching factor which is to specify the maximum number of data points in each sub-cluster. Applying this algorithm for connection points training, we conducted a hyperparameter tuning in terms of branching factor with the range between 2 and 200 and the threshold with the range between 1 and 50. However, we found that no clear clusters can be identified for this type of dataset. The major reason is that the noise points as illustrated in Fig. 12 could not be properly excluded such that the algorithm will always consider the entire group of data points as a single cluster. For the sake of clarity, no images are listed and further investigation for detailed comparisons among different unsupervised learning algorithms to identify connection points will be done in the future.

5.2. Identification of navigated routes

Fig. 13 shows the most navigated routes in two cases (as seen in Fig. 10) with the historical selection probabilities. The probability is calculated through the number of route pattern nodes in each cluster and the overall number of route pattern nodes in each data group, as seen in Fig. 14. The utilizing of such historical probability can provide an intuition of how people actually chose their ways during journeys under different weather conditions. For instance, if a ship transports goods from somewhere in Gulf region to somewhere in South Korea. No doubt, it will first pass through Panama canal (a connection point B is set here, as seen in Fig. 12), then, there are three options to choose. One route is going to the north of Pacific Ocean (probability...
of 0.08), and passing the Tsugaru Strait in Japan (a connection point is set here) to local sea region then all the way to the final destination. The second route is going to the south of Pacific Ocean close to Hawaii islands (probability of 0.62), and passing the area close to Kagoshima in Japan (another connection point) to the local sea region then all the way to the final destination. The third route is the one between the former two routes with a probability of 0.30. Therefore, such route library is capable of providing plausible voyage references without suffering risks due to unknowns and hazard sea regions. Based on these routes, the journey distances as well as other information relevant to the pre-fixture stage can be easily estimated.

Using KMeans to identify routes, the number of clusters should be manually determined in advance. The training strategy here we adopted is to compare the Silhouette scores among models with different number of given clusters. As seen in Fig. 14, the clusters of voyage pattern nodes in two cases are properly identified with respective Silhouette score of 0.55 and 0.44 when setting the number of clusters to 3. The scores reflect a reasonable categorization of the voyage patterns from the standpoint of engineering. However, there are some limitations by using this algorithm. For instance, the number of clusters have to be manually determined in advance, which may cause an inappropriate estimation due to human-factors. For the simplification of voyages, we used pattern nodes. Other methods to identify trajectory similarities could be also adopted such as MD and DTW, which will be done in the future research.

6. Conclusion remarks

In this paper, a practical data-driven methodology for route selections has been proposed. A route library is therefore obtained based on AIS data of tankers within 5 years, which can be used for voyage planning and distance estimation by shipping companies at the pre-fixture stage. The speed-weighted geolocation method is first used to obtain all the pattern nodes of voyages. Then, the KMeans is adopted to classify pattern nodes into different clusters. The nodes close to the centroids of clusters are selected as the representative routes in open sea passages. Simultaneously, connection points in local sea passages have been trained by DBSCAN. Combining the representative routes with connection points, the most navigated routes among ports are formed. Two cases have been showed to demonstrate the effectiveness of this method. The route library selected by this method with accurate journey distance information is capable of providing voyage references for shipping industry to support their decision making at the pre-fixture stage.

In this work, the accuracy of using the sophisticated ML algorithm DBSCAN to train connection points in local sea passages is not high due to uneven AIS data distribution. Expert/manual intervention is needed to further ensure a meaningful selection. Hence, a possible future work should focus on other methods which can be more effective to identify connection points in local regions. Meanwhile, the current work and findings are only based on tanker vessels. The application of the methodology on other types of vessels such as container ships is therefore could be done in the future. In addition, comparison investigation with the existing routes from pilot charts or the shortest path of the methodology on other types of vessels such as container ships work and findings are only based on tanker vessels. The application of this work reported in this paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work of this paper belongs to ShippingLab project “Digital vessel operation”. The funding of the project is provided by the Innovation Fund Denmark (IFD) under File No. 8090-00063B, Danish Maritime Fund, Lauritzen Fonden and Orient’s Fond, Denmark. Gratitude is also expressed to the shipping companies Torm for the providing of AIS data and to VPS for precious discussions and comments.

References


