Ocean Engineering 198 (2020) 106936

Contents lists available at ScienceDirect

Ocean Engineering

journal homepage: www.elsevier.com/locate/oceaneng

Data mining approach to shipping route characterization and anomaly detection based on AIS data

H. Rong, A.P. Teixeira, C. Guedes Soares *

Centre for Marine Technology and Ocean Engineering (CENTEC), Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001, Lisboa, Portugal

ARTICLE INFO

Keywords:
AIS data
Data mining
Maritime traffic
Traffic characterization
Anomaly detection

ABSTRACT

A data mining approach is presented for probabilistic characterization of maritime traffic and anomaly detection. The approach automatically groups historical traffic data provided by the Automatic Identification System in terms of ship types, sizes, final destinations and other characteristics that influence the maritime traffic patterns off the continental coast of Portugal. The approach consists of identifying relevant waypoints along a route where significant changes in the ships' navigational behaviour are observed, such as changes in heading, using trajectory compression and clustering algorithms. This provides a vector-based representation of the ship routes consisting of straight legs and connecting turning sections that facilitates route probabilistic characterization and anomaly detection. The maritime traffic is characterized probabilistically at the identified route legs and waypoints in terms of lateral distribution of the trajectories and speed profile, which allows the characterization of the typical behaviour of a group of similar ships along a particular route. In the proposed approach heading changes are automatically detected using the Douglas and Peucker algorithm and clustered by the density-based spatial clustering of applications with noise algorithm. The proposed method is applied to the characterization of southbound maritime traffic from the traffic separation scheme off Cape Roca to the ports of Lisbon, Setúbal and Sines. Finally, an example of ship trajectory anomaly detection based on the developed maritime traffic probabilistic models is provided.

1. Introduction

The maritime transportation is responsible for approximately 90% of the world's trade and more than 50,000 ships sail on the ocean each day (AGCS, 2018). Surveillance of maritime traffic is therefore of great importance for maritime safety and security.

The need to ensure the safety of navigation has led to the implementation of Automatic Identification System (AIS) base stations, which receive and maintain records of messages with dynamic and static information transmitted by ships. Although several types of errors in AIS messages have been identified (Norris, 2007), it is obvious that the system provides plenty of data for maritime traffic analysis and modelling tasks as information on the ships' behaviour and traffic conditions is important in many maritime related studies (Tu et al., 2018).

The AIS data provides navigational information of ships, including the ship's name, its particulars as well as ship's position, speed and heading. Such information about ship's navigation behaviour has been used as input to many maritime risk models, as reviewed by Chen et al. (2019). Mou et al. (2010) developed linear regression models of Closest Point of Approach (CPA) and Time to Closest Point of Approach (TCPA) between collision-involved ships based on AIS data to evaluate the risk of ship collisions in the North Sea off the Port of Rotterdam. Qu et al. (2011) applied collision risk indices such as the speed dispersion, degree of acceleration and deceleration, and a number of fuzzy ship domain overlaps to assess quantitatively the collision risk at the Singapore Strait. Silveira et al. (2013) proposed a method to determine the number of collision candidates based on available Automatic Identification System (AIS) data with the concept of collision diameter defined by Pedersen (1995). The concept of "ship domain" has been used to identify near collision scenarios by using simulation models of ship navigation in restricted waters and AIS data (Rong et al., 2015a,b), which allowed the characterization of near ship collision scenarios off the coast of Portugal in terms of ship types and dimensions, crossing angle, relative velocity, among others (Rong et al., 2016). Zhang et al. (2015) proposed the Vessel Conflict Ranking Operator (VCRO) model to rank the risk level of various ship encounter situations based on maritime expert judgements. The model considers the interaction between the two vessels, namely the

* Corresponding author.
E-mail address: c.guedes.soares@centec.tecnico.ulisboa.pt (C. Guedes Soares).

https://doi.org/10.1016/j.oceaneng.2020.106936
Received 24 August 2019; Received in revised form 8 January 2020; Accepted 9 January 2020
Available online 28 January 2020

relative speed, the angle and the distance between the ships. This model has been improved by adding the ship size and the Minimum Distance to Collision (MDTC) concept that is related to the urgency of evasive manoeuvring for better assessing the risk of the ship encounters (Zhang et al., 2016).

The spatial analysis of near collisions identifies potentially dangerous locations for the maritime transportation and provides relevant information for traffic monitoring and control. In this context, Wu et al. (2016) have investigated features of vessel conflicts in the Southeast Texas waterway in terms of vessel size, spatial distribution, time-of-day and the risk level of vessel collisions defined by Vessel Conflict Ranking Operator (VCRO) model developed by Zhang et al. (2015). Yoo (2018) has estimated the near collision density in coastal areas and has identified high collision risk locations based on AIS data. Along the same line, Rong at al. (2019a) have adopted collision risk indicators, including the DCPA, TCPA and the relative distance between ships, to assess the collision risk of near collision scenarios identified from historical AIS data. Spatial density maps of the collision risk are then derived using the Kernel Density Estimation method.

Even though the analysis of near collision scenarios can be used for maritime traffic risk assessment, a further insight into the characteristics of local maritime traffic is of great importance to maintain the operational efficiency and safety of the maritime transportation. The characterization of the maritime traffic patterns supports maritime surveillance as it provides information on the typical trajectories of ships along the main traffic routes. The ship motion data provided by AIS can be grouped based on the similarity behaviour of the ships' trajectories to provide an overview of the general motion patterns and can be used to model the ship routes.

Various traffic motion pattern extraction methods are available for converting raw AIS data into effective traffic patterns including, computer vision techniques (Aarsæther and Moan, 2009), statistical analysis techniques (Etienne et al., 2012) and clustering methods (Pallotta et al., 2013; De Vries et al., 2012). Traffic motion pattern extraction methods aim at grouping ship trajectories following the same itinerary. The traffic motion pattern once extracted can be exploited to characterize the ship behaviour within the motion pattern in a macro level. Kang et al. (2018) applied the weighted least square approach to study the ship traffic speed-density relation of the Singapore Strait traffic motion patterns. Wu et al. (2018b) studied the travel behaviour of ships in narrow waterways using AIS data in terms of distributions of flow speeds and densities of ships, as well as their relationships with the high vessel conflicts areas identified by Wu et al. (2016), and have assessed the collision risk in these locations. Chen et al. (2018) developed a ship trajectory classification model by adopting Least-squares Cubic Spline Curves Approximation (LCSCA) technique to represent the ship trajectories. AIS data from the port of Rotterdam was analysed and clustered by Zhou et al. (2019). Based on the characterization of ship behaviour clusters, it was possible to detect and rank collision risk between encounter ships within each cluster.

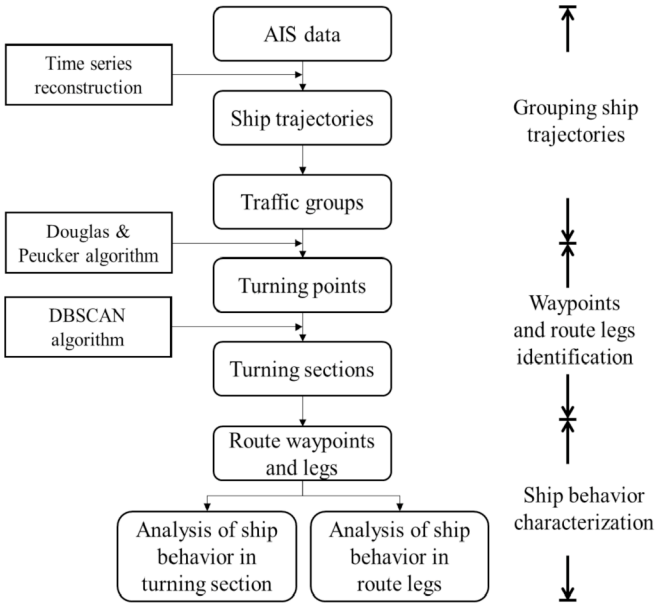


Fig. 1. Flowchart of maritime traffic characterization.

A clustering-based method for extracting maritime traffic patterns from historical AIS data has also been presented by Sun et al. (2015). The extracted maritime traffic pattern enabled constructing a normalcy model that can be used for anomaly detection. The vessel traffic and motion information can also be exploited to perform ship route prediction at a given time. By incorporating both the traffic pattern and ship motion behaviour knowledge, Xiao et al. (2017) proposed a ship traffic prediction method capable of predicting maritime traffic 5, 30 and 60 min ahead. Rong et al. (2019b) have adopted Gaussian Process models derived from historical AIS data for probabilistic ship trajectory prediction within motion patterns and for collision probability evaluation (Rong et al., 2020).

Previous research works on ship traffic learning from AIS data have shown that valuable knowledge about the ship behaviour can be extracted through the analysis of historical data. However, the existing studies are mainly focused on the behaviour of ships in specific ship routes. A new trend in the field of maritime traffic characterization is to adopt a vector-based representation of ship routes defined as a set of straight legs connecting turning sections. Rong et al. (2018) proposed a data mining method on AIS data to automatically identify maritime traffic junctions and applied a multinomial logistic regression model for predicting the ships destination based on a set of characteristics of the ships' behaviour at the junction. Zhang et al. (2018) applied data-driven approaches, including ship trajectory compression and clustering methods, to determine the general behaviour and spatial patterns of trajectories following the same route. As a result of the characterization of the maritime traffic patterns, applications and services such as ship behaviour recognition, trajectory prediction and abnormal traffic behaviour can be developed.

In this study, a data mining approach is proposed for probabilistic characterization of the maritime traffic off the continental coast of Portugal based on Automatic Identification System data that combines trajectory compression and clustering algorithms to characterize routes as a set of straight legs and connecting turning sections. This way probabilistic models of the ship trajectories can be easily derived along a vector-based representation of the ship routes to support ship abnormal behaviour detection.

The probabilistic characterization of the maritime traffic and the definition of operational profiles and specific routes off the continental coast of Portugal will reduce the uncertainty of each operation, thus increasing the overall safety level and the efficiency of the current maritime operations. This can benefit many maritime stakeholders for optimizing their operations such as port operators, pilotage services, ship owners and ship operators. Moreover, these routes defined in terms of waypoints derived from historical AIS data have the potential to become one of the key enablers for future autonomous vessels traffic operations (e.g. Xu et al., 2019).

Moreover, the main characteristics of traffic parameters can be used as input for the traffic simulation, which is useful for risk management (Rong et al., 2015a, b; van Dorp et al., 2011; Goerlandt et al., 2011), and channel capacity assessment with increased traffic volume (Xin et al., 2019).

Finally, the probabilistic characterization of the ship trajectories along a given route enables real-time anomaly detection, which is an important feature for developing alerts to support traffic supervision and control tasks.

The approach is applied to the southbound traffic passing the Traffic

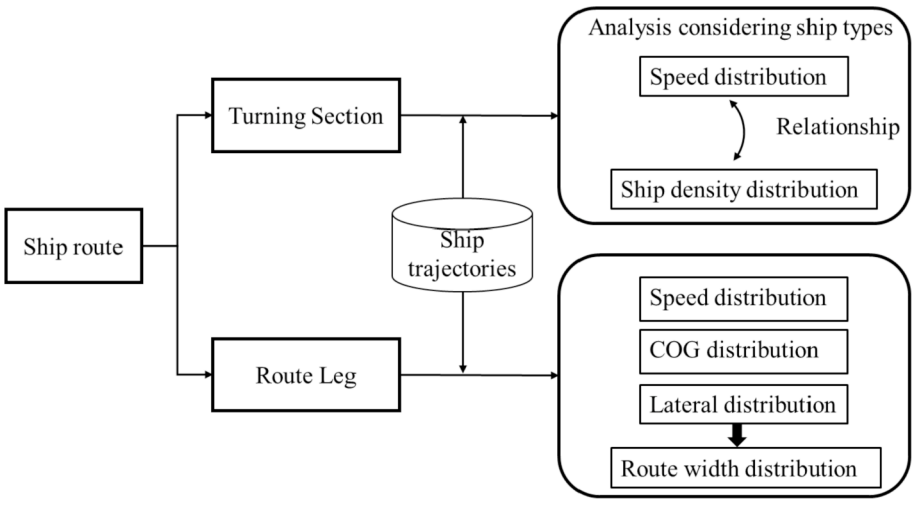


Fig. 2. Framework for characterization of ship routes.

Separation Scheme off Cape Roca toward the ports of Lisbon, Setúbal and Sines. Finally, an example of online ship trajectory anomaly detection is provided. The approach uses the normalcy models of the maritime traffic derived from historical AIS data to identify ship trajectories that deviate probabilistically from the normal behaviour of a class of ships in a particular route.

2. Methodology

The aim of this study is to characterize the motion patterns and shipping routes based on historical AIS data. The analysis of traffic patterns is essential as it provides the possibility to integrate and enrich maritime traffic services including route modelling, trajectory prediction and detection of abnormal traffic behaviour. The extraction of traffic patterns starts by grouping ship trajectories following the same itinerary, by clustering static or entry and exit points corresponding to starting and ending locations of the motion pattern.

Fig. 1 shows the flowchart of the proposed approach for maritime traffic characterization, consisting of three main steps:

1) Define traffic patterns by grouping ship trajectories according to final destinations;
2) Waypoints and route legs identification;
3) Ship behaviour characterization at waypoints and route legs.

A shipping route is assumed as a set of straight segments or legs connecting waypoints, corresponding to locations of significant changes in the navigational behaviour of the ships, which allows a compact representation of ships' motion.

In order to characterize the maritime traffic along a shipping route, ship trajectories obtained from AIS Data are first grouped according to the final destination of the ships. The ship trajectories within a particular motion pattern or route are then analysed so as to identify the waypoints, where significant changes in the ships' dynamic behaviour, such as changes on heading or velocity, are observed.

In particular, waypoints corresponding to turning sections at specific locations along the route are identified. Turning point detection and clustering techniques are applied to detect maritime turning sections. Turning points of ship trajectories are detected based on Douglas and Peucker (DP) algorithm that allows detecting ship directional changes at the trajectory level based on spatial information (Douglas and Peucker, 1973), which is more reliable than the use of the Course Over Ground (COG) for turning point detection. A turning section corresponds to an area where a significant density of turning points is observed. For this purpose, the Density-Based Spatial Clustering of Applications with Noise algorithm (Ester et al., 1996) is adopted for clustering the turning points.

Then, the maritime traffic is characterized probabilistically at the identified waypoints and at route legs connecting the waypoints, in terms of lateral distribution of the trajectories, speed profile and ship traffic density, which allows the characterization of the typical behaviour of a group of similar ships along a particular route, as shown in Fig. 2.

2.1. Route turning section identification

2.1.1. Douglas and Peucker (DP) algorithm

A route turning section is automatically identified by clustering turning points detected in ship trajectories. A turning point in ship trajectory is where a significant directional change of the ship trajectory is observed. The Douglas and Peucker (DP) algorithm typically used for ship trajectory compression can also be used for turning point detection. The DP algorithm is an approach for reducing the number of points in a curve that is approximated by a series of points. In the field of moving object analysis, this algorithm has been used to compress trajectories data (Etienne et al., 2012). This algorithm, together with a suitable threshold parameter can be used to identify the relevant turning points along the ship trajectory.

According to the DP algorithm, the first and last points of a given trajectory are kept automatically. Then, the farthest point from the line defined by the first and last points of the trajectory is found. If the

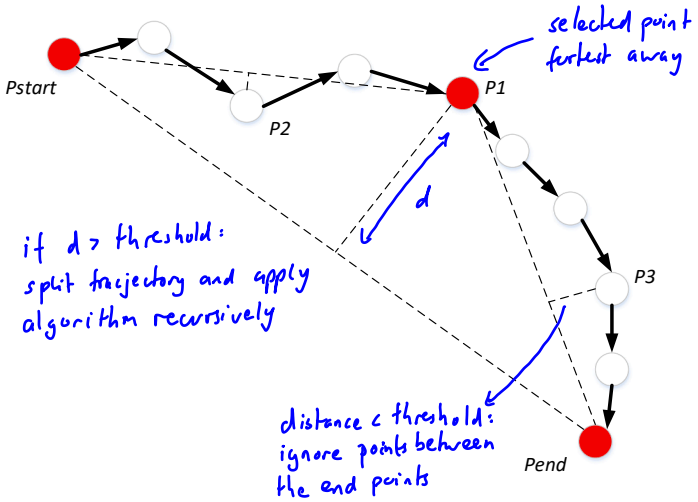


Fig. 3. Illustration of the Douglas and Peucker (DP) algorithm.

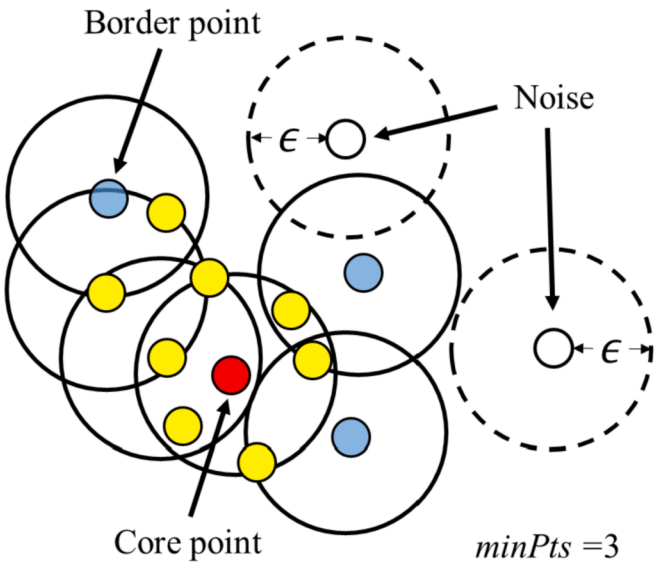


Fig. 4. Illustration of DBSCAN algorithm.

transverse distance of the point to this line calculated by Eq. (1) exceeds the threshold, then the point is retained, the trajectory is then split at that position and the algorithm is recursively applied to both sub-trajectories. Otherwise, any point between first and last points of the trajectory (or sub-trajectories) can be discarded. See Fig. 3 for an illustration of the DP algorithm. In the figure, the solid lines represent the original trajectory, and the dashed lines represent a DP simplified trajectory. P_1 is the furthest point to the line defined by P_{start} and P_{end}. The distance from P_1 to the line defined by P_{start} and P_{end} exceeds the threshold meaning that P_1 will be retained. However, distance from P_2 to the line defined by P_{start} and P_1 and distance from P_3 to the line defined by P_1 and P_{end} are smaller than the threshold, and therefore, the remainder points of both original sub-trajectories are discarded.

$$distance((P_a,P_b),p) = \frac{|(y_b-y_a)^*x_p-(x_b-x_a)^*y_p+x_by_a-y_bx_a|}{\sqrt{(x_b-x_a)^2+(y_b-y_a)^2}} \quad (1)$$

where $distance((P_a,P_b),p)$ is the distance from point $p=(x_p,y_p)$ to the line defined by two points $P_a=(x_a,y_a)$ and $P_b=(x_b,y_b)$.

x, y as lat, lng or what?

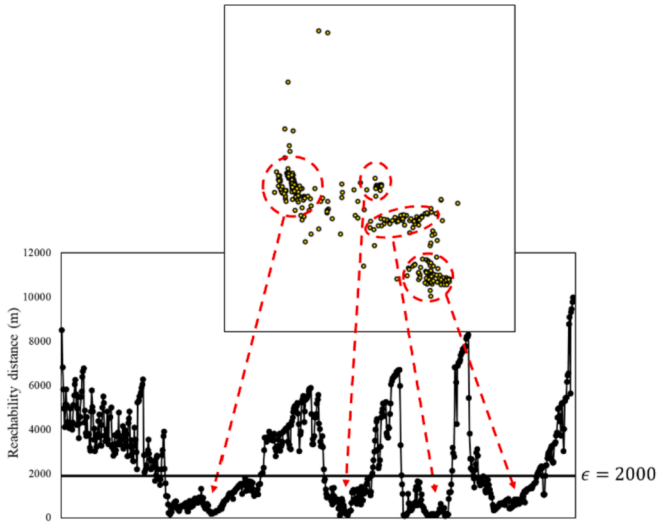


Fig. 5. Illustration of reachability plot.

2.1.2. Clustering turning points based on DBSCAN algorithm

In the context of maritime traffic, the retained points detected by the DP algorithm can be considered as turning points, and the clustering of those turning points defines the turning sections along the ship route.

In this paper turning points clustering is based on the DBSCAN algorithm. This algorithm forms clusters of elements based on the density of points in their neighbourhood. In the clustering procedure, the points are classified as core points, density-reachable points and noise, as shown in Fig. 4. Given a set of turning points $\mathscr{D}$, DBSCAN algorithm initially looks for core points, i.e. points $p_i \in \mathscr{D}$ such that $|N_\varepsilon(p_i)| \geq MinPts$, where $N_\varepsilon(p_i)$ is the ε-neighbourhood of p_i. Once a core point is found, such neighbourhood points are density-reachable from p_i and belong to the same cluster. Then, the points that are not density-reachable from other points are considered as noise. The pseudocode of DBSCAN algorithm is expressed in Algorithm 1 (Ester et al., 1996).

Algorithm 1. DBSCAN ($\mathscr{D}, \varepsilon, MinPts$)

set of turning points

```
C=0
for each unsinged point p_i in dataset D
    mark p_i as signed
    Neighbor(p_i) = RegionQuery(p_i, ϵ)
    if size(Neighbor(p_i)) < MinPts
        mark p_i as Noise
    else
        C=next cluster
        ExpandCluster(p_i, Neighbor(p_i), ϵ, MinPts)
    End if
End for

Algorithm ExpandCluster(p, Neighbor(p), ϵ, MinPts, C)
add p to cluster C
For each point p' in Neighbor(p)
    if p' is not unsinged
        mark p' as singed
        Neighbor(p')= RegionQuery(p', ϵ)
        if size(Neighbor(p')) >= MinPts
            Neighbor(p) joined with Neighbor(p')
            if p' is not yet member of any cluster
                add p' to cluster C
            End if
        End if
    End if
End for

Algorithm RegionQuery(p, ϵ)
return all points within p's ϵ -neighborhood (including p)
```

The DBSCAN algorithm does not require the specification of the number of clusters in the data *a priori*, however, there is no general way of choosing *minPts* and the radius ε. The setting of *minPts* requires some knowledge on the data density on study area. However, in the present case the algorithm has shown to be robust with respect to *minPts*, as similar clustering structures are obtained for a wide range of *minPts* values (between 5 and 10). The value of the radius ε is determined by a reachability plot of the ordered points produced by the OPTICS algorithm (Ankerst et al., 1999) that is a plot showing on the x-axis the ordering of the points as processed by OPTICS and the reachability distance on the y-axis. The reachability plot of the ordered points enables the visualization of the clustering structure of the data set. Fig. 5 shows a graphical representation of density-based clustering structure of a data set. Since points belonging to a cluster have a low reachability distance to their nearest neighbour, the clusters correspond to valleys in the reachability plot.

2.2. Characterization of ship traffic behaviour along the route

After identifying the route waypoints corresponding to turning sections or significant changes on the ships' velocities, the maritime traffic at the waypoints and route legs between waypoints is characterized, as shown in Fig. 2. In particular, at the turning sections of the route, ship speed distribution and traffic density are statistically analysed. In order to investigate the relationship between ship speed and traffic density,

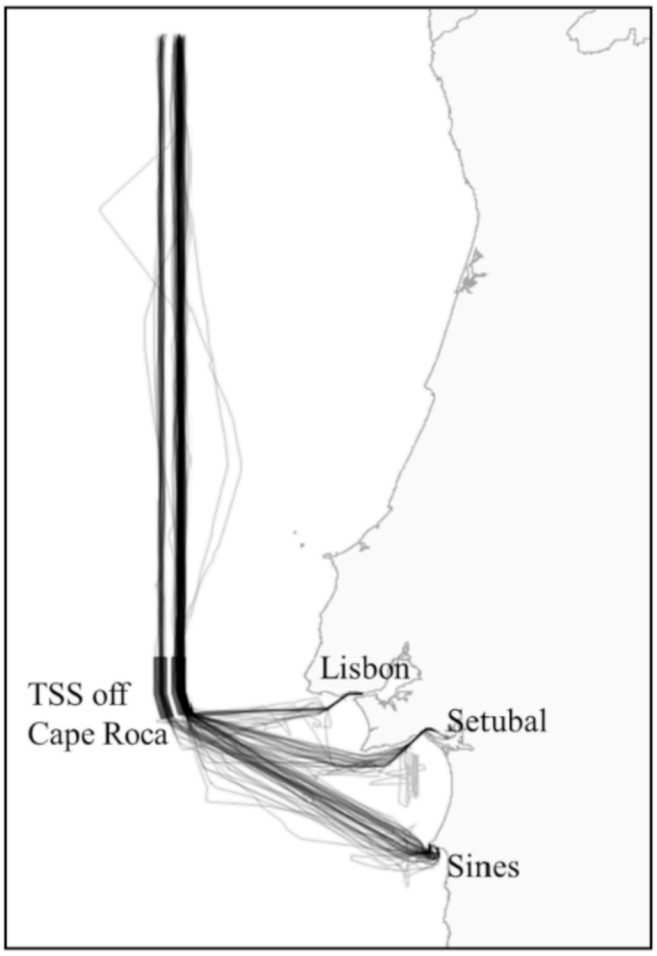


Fig. 6. Southbound ship traffic in TSSs off Cape Roca to the ports of Lisbon, Setúbal and Sines.

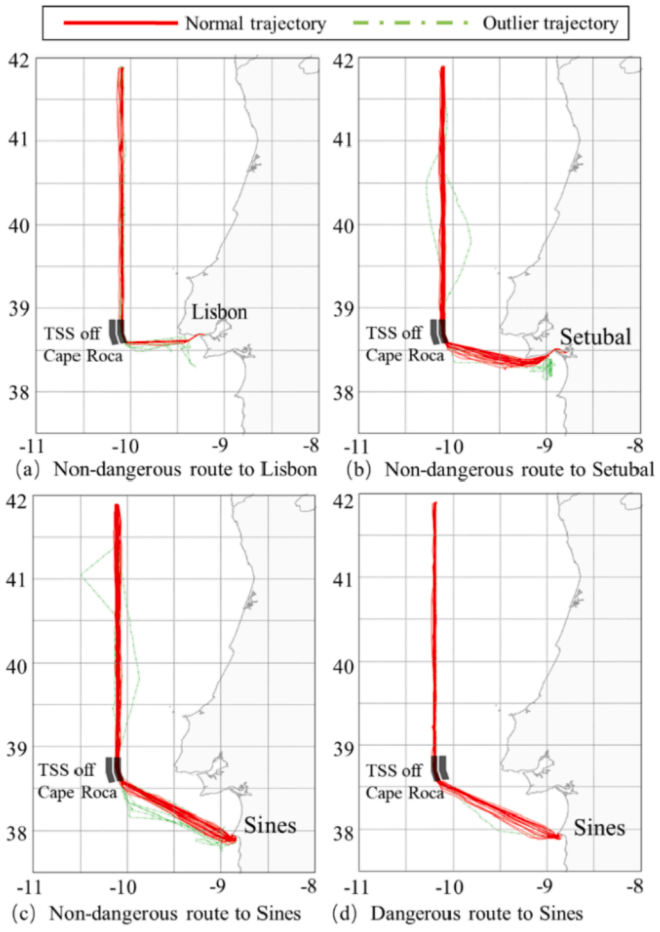


Fig. 7. Traffic-groups by final destination.

ship speed profiles under different traffic densities are studied. The maritime traffic at the route legs is also characterized in terms of ship speed distribution, ship heading distribution and lateral distribution of ship trajectories.

3. Case study

The maritime areas off the continental coast of Portugal are crossed by a complex network of routes where routes connecting northern Europe and the Mediterranean Sea meet vessels bound to and leaving from national ports. Two Traffic Separation Schemes (TSSs) located off Cape Roca and off Cape San Vincente organize the Northbound and Southbound Traffic of dangerous and non-dangerous cargo off the coast of Portugal and influence significantly the routes of the ships entering and leaving the national ports of Lisbon, Setúbal and Sines, as shown in Fig. 6.

In this paper, the suggested approach to define and characterize shipping routes according to the ships' final destination is applied to the southbound maritime traffic data passing the TSS off Cape Roca to the ports of Lisbon, Setúbal, and Sines over a period of three months (from 1st October to December 31, 2015). Fig. 6 shows the ensemble of the southbound ship trajectories to the ports of Lisbon, Setúbal and Sines.

3.1. Maritime traffic groups

First, the southbound maritime traffic passing the TSS off Cape Roca is grouped into motion patterns or main routes defined by the ship trajectories corresponding to the different ports of destination. In particular, the maritime traffic in this area is divided into four motion patterns which are: 1) Non-dangerous traffic lane to Lisbon Port; 2) Non-dangerous traffic lane to Setúbal Port; 3) Non-dangerous traffic lane to Sines Port and 4) Dangerous traffic lane to Sines Port. The motion patterns generally show a certain geographic clustering of the ship trajectories, as the individual ships follow a similar path to their final destinations. Fig. 7 shows the southbound maritime traffic passing the TSS off Cape Roca extracted from three months of AIS data grouped in terms of the different final destinations. The number of ship trajectories in each group is shown in Table 1. One can see that the number of ship trajectories to the three ports is similar.

Fig. 7 shows some outlier trajectories in each traffic group identified in green. An outlier is a data point that stands out in contrast to the other data points around it. The task of outlier detection consists of inferring whether the data point deviates significantly from the data. These trajectory outliers are removed before the characterization of the ship route as explained below.

The definition of the spatial distribution of trajectories along the ship route starts by estimating the route centreline. According to Rong et al. (2019a, b), each AIS position record of a ship trajectory within each traffic group is aligned based on the Dynamic Time Warping algorithm,

Table 1
Number of trajectories in each traffic-group.

Traffic group	No. of trajectories
Lisbon Port	25
Setúbal Port	31
Sines Port	58
Sum	114

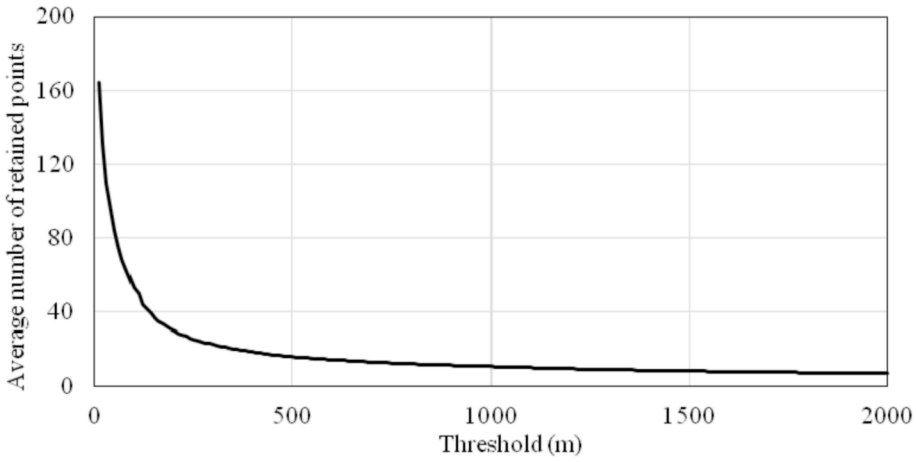


Fig. 8. The relationship between the distance threshold of DP algorithm and average number of trajectory retained points.

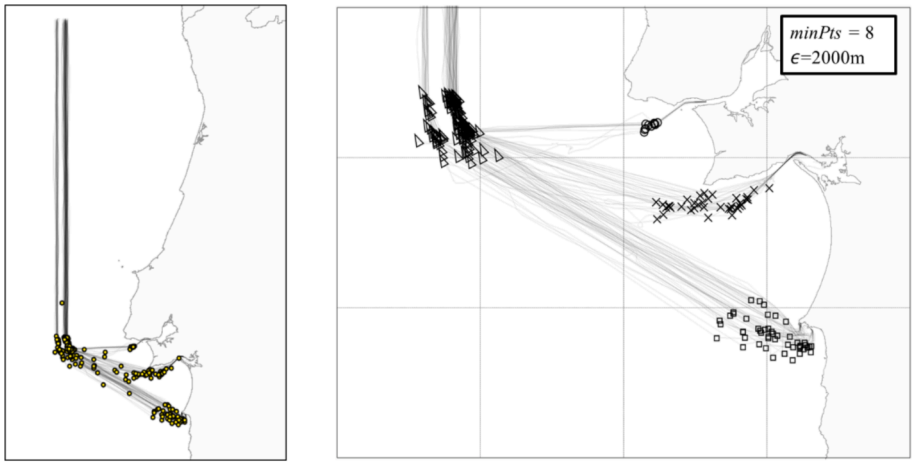


Fig. 9. Clustering of turning points.

and then the route centreline is derived from the arithmetic mean of the aligned elements.

Let $p = (lat_p, lng_p)$ be a point on the geographic coordinate system and let $Cl = \{cl_1, \ldots, cl_n\}$ represent the centerline of the ship route, where $cl_i = (lat_{cli}, lng_{cli})$. p' is the point from object p projected on the centreline CL, and located in two adjacent centerline points (cl_i, cl_{i+1}). Then, the longitudinal and lateral distance of point $p = (s_p, u_p)$ on the ship route can be calculated by:

$$s_p = \sum_{j=1}^{i-1} ||cl_{j+1} - cl_j|| + ||p' - cl_i|| \tag{2}$$

$$u_p = ||p - p'|| \tag{3}$$

In this study, a normal distribution is adopted to describe the spatial distribution of the ship lateral position. This way, outliers can be considered to be points that lie three or more standard deviations from the mean of the normal distribution. Therefore, outlier trajectories can be identified if the lateral distance of trajectories points is significantly large.

3.2. Turning points, sections

Knowing where a ship changes its heading is important to characterize the normal behaviour of the traffic along the route and, therefore, the DP algorithm described before is used for ship trajectory compression and turning point detection. According to the relationship between threshold of DP algorithm and average number of retained points (Fig. 8), it can be seen that the trajectory points reduce as the tolerance increases and tend to stabilise when the threshold is greater than 500m. The aim of trajectory compression based on the DP algorithm is to decrease the number of retained points keeping an adequate trajectory representation, therefore, the threshold distance is set to 500 m in this study.

The turning points detected along the ship routes passing the TTS off Cape Roca to the three ports are shown in Fig. 9a (black points). Then, turning sections are identified by the spatial distribution of the coordinates of the turning points detected.

To investigate the spatial distribution of the turning points, the Kernel Density Estimation (KDE) method is applied to estimate the spatial density distribution of the turning points on the study area. The KDE method constructs the spatial density distribution by placing a kernel function on the turning points. The spatial density distribution is given by:

$$g(x,y) = \frac{1}{nh^2} \sum_{i=1}^{n} \phi\left(\frac{\sqrt{(x-x_i)^2 + (y-y_i)^2}}{h}\right) \tag{4}$$

where n is the number of turning points, h is the bandwidth of the kernel function. In this study, the most commonly used squared exponential kernel function is adopted:

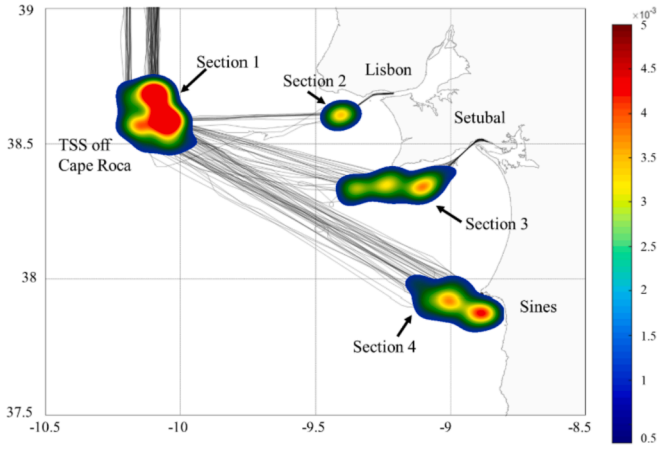


Fig. 10. Turning sections identified based on the KDE.

Table 2
Ship type distribution at each route.

Route	Cargo	Tanker	Passenger
ND to Lisbon	16	4	2
ND to Setúbal	26	0	1
ND to Sines	30	5	0
D to Sines	1	20	0

*ND: Non-Dangerous; D: Dangerous.

$$\phi(x) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{5}$$

where σ is the standard deviation of the samples. If the Gaussian basis kernel function is used to approximate samples, the underlying density being estimated is Gaussian, then the optimal choice for bandwidth h is (Silverman, 1986):

$$h = \left(\frac{4\sigma^5}{3n}\right)^{1/5} \approx 1.06\sigma n^{-1/5} \tag{6}$$

Fig. 10 shows the spatial density distribution of the turning points. The study area is discretized in a 600*500 grid and the spatial density distribution of the turning points at each cell is calculated using the KDE method and depicted using the RGB (Red Green Blue) colour code. Once the density value of each grid cell is calculated, the colours of the grids are determined according to the corresponding turning points.

3.3. Characteristics of the maritime traffic in each turning section

A statistical analysis of the maritime traffic shows that the southbound traffic passing the TSS off Cape Roca consists mainly of cargo ships and tankers, which account for 69.5% and 27.6% respectively.

Table 3
Ship speed distributions at each turning section.

Turning section	Mean	Stand deviation	p-value
Section 1 (All ships)	12.59	3.19	0.59
Section 1 (Tankers)	13.09	2.38	0.50
Section 2	8.94	2.83	0.41
Section 3	9.28	3.34	0.44
Section 4	8.17	6.42	0.06

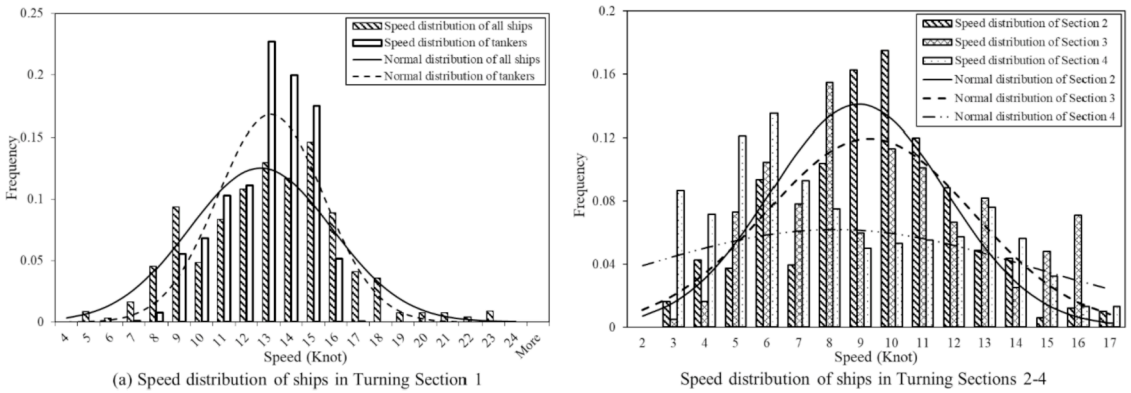
Different routes are used by different ship types. Most of the ships that visit the ports of Lisbon and Setúbal are cargo ships, 72.7% and 96.3% respectively, whereas 44.6% of the ships toward the port of Sines are tankers, as shown in Table 2. Passenger ships only navigate in routes to the port of Lisbon and Setúbal. On the other hand, 86.2% of the tankers sail to Sines Port. The characteristics of the ship type distributions of each motion pattern imply that ship route was used by different ship types.

Fig. 11 shows the speed distributions of ships in each turning section and a statistical summary of the speed distributions can be found in Table 3. It is seen that the average ship speed at turning Section 1 is higher than at the others (12.59 versus 8.94, 9.28 and 8.17, respectively). Turning Sections 2-4 correspond to the entrances of Ports of Lisbon, Setúbal and Sines, where the ship speed is limited due to the geophysical characteristics of restricted waterways. From the speed distribution of tankers in Turning Section 1 (Fig. 11a), one can see that the average speed of tankers is slightly higher than the average speed of general ship types.

In turning section 1, a *t*-test showed that the speed distribution of all ship types and tankers is well approximated by a normal distribution (*p*-value equal to 0.59 and 0.5, respectively). In addition, the standard deviation of the tankers' speed is smaller than that of all ships. It should be mentioned that the speed distribution at Turning Section 4 is not normally distributed. The southbound ship route to the port of Sines is dominated by tankers, and near the Port of Sines there exists an anchoring area, the travel behaviour of tankers in Section 4 is more complex compared with that of ships in turning Sections 2 and 3.

The revealed ship trajectory data makes it possible to study the density of ships in the corresponding route. In this study, the "density" of a ship at a section simply refers to the number of ships at this section. It should be mentioned that the estimation of density takes all collected historical AIS data into consideration.

Fig. 12 shows the ship density distribution at each turning section. According to Fig. 12a, it can be found that 17.9%, 21.7% and 18.9% of ships encounter 3–5 other ships in Section 1, respectively. In Section 2 and Section 3, ships are less likely to encounter more than three ships



(a) Speed distribution of ships in Turning Section 1

Speed distribution of ships in Turning Sections 2-4

Fig. 11. Histograms of SOG for ships in the turning sections.

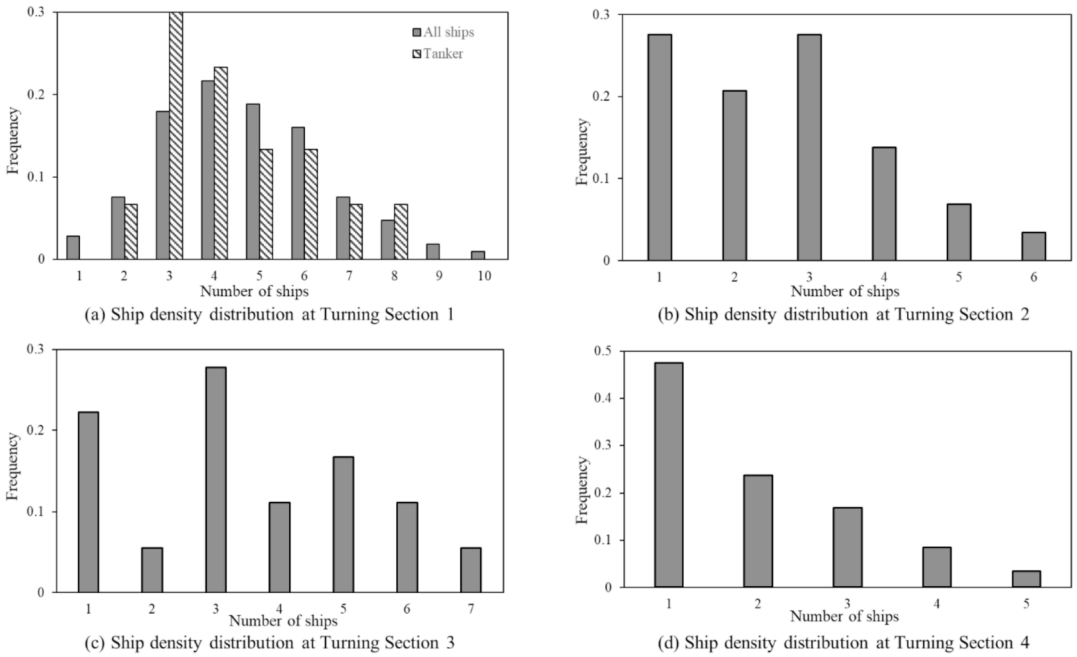


Fig. 12. Ship density distribution in each Turning section.

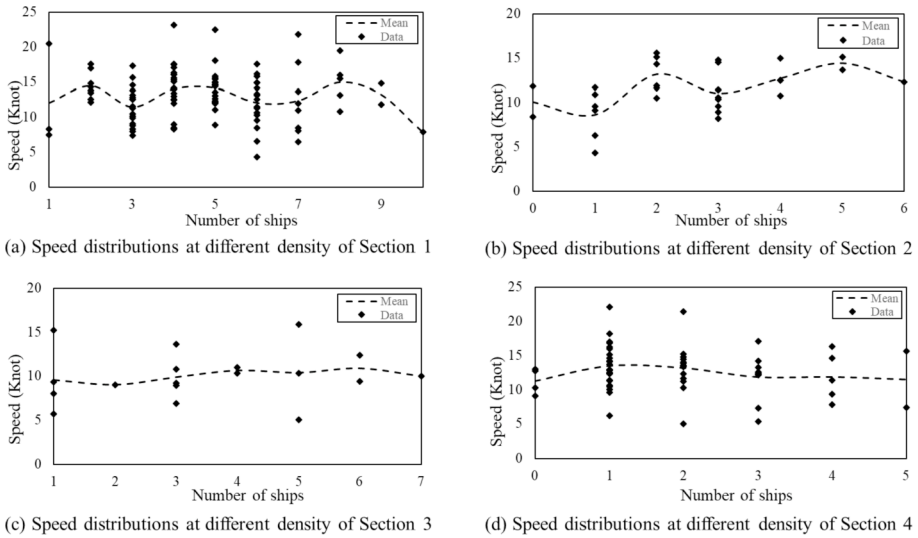


Fig. 13. Speed distributions at different density of each section.

(72.2% and 51.7% in Section 2 and Section 3, respectively).

In terms of environmental pollution, tankers pose a high environmental risk as they carry large amounts of oil or toxic chemicals. In this study, the density distribution when a tanker is at Turning Section 1 is presented. Fig. 12a also shows that it is more likely for a tanker to encounter three or four ships (30% and 23.3%, respectively) in Turning Section 1. The inbound ship traffic to the Port of Sines is dominated by tankers (see Table 2), 71.2% of ships encounter only one or two ship in Section 4 (entrance of Port Sines), which indicates that the entrance in the Port of Sines seems is less congested.

The speed distributions and density distribution of each section is presented above, respectively. In the following, for the general traffic flow of all types of ships and traffic flow of tankers, the mean speed under different traffic density level were investigated, as shown in Fig. 13. It is seen that the density level did not significantly affect ship speed. In addition, the difference between minimum and maximum speed of mean speed distribution in Section 2 and Section 4 (Fig. 13) are 1.9 knot and 2.16 knot. The nearly flat curves imply that the speeds at these two sections are free-flow speeds which indicate that these two sections do not suffer from serious traffic congestion.

3.4. Characteristics of the maritime traffic in the route legs

As the Port of Sines is the main port on the Atlantic seaboard of Portugal due to its geophysical characteristics, it is the main gateway to the energy supply of Portugal: natural gas, coal, oil and its derivatives. In terms of environmental pollution, tankers pose a high environmental risk as they carry large amounts of oil or toxic chemicals. An example of an accident of this type that occurred close to the entrance of Sines is described by Sebastião and Guedes Soares (2006), while a more general description of pollution accidents in the Portuguese Coast can be found in Gouveia and Guedes Soares (2010). A statistical analysis is performed to study the maritime traffic on the route passing southbound dangerous traffic lane to Port of Sines, as tanker is the main ship type in this route.

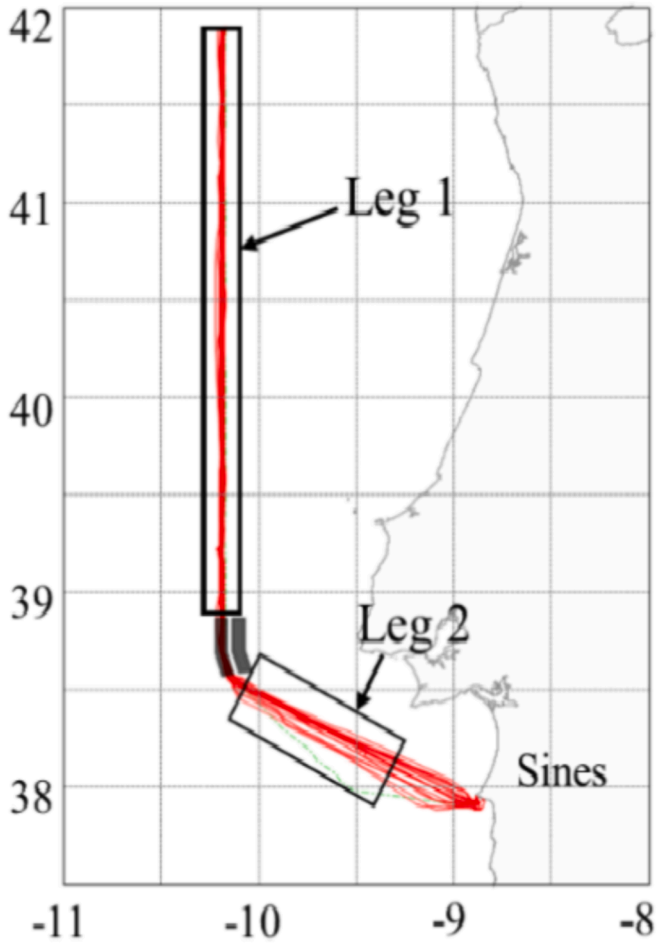


Fig. 14. Two legs of the route passing southbound dangerous traffic lane to the Port of Sines divided by the turn section.

As shown in Fig. 14, the ship route that passes the Southbound dangerous traffic lane to Port of Sines is divided by two turning sections derived above. In Leg 1 ships move southward toward the TSS off Cape Roca's Southbound dangerous traffic lane. Then, in Turning Section 1 ships turn east toward Sines Port in Leg 2.

Fig. 15 illustrates the speed and COG distributions of ships in the two legs. In terms of speed distributions, it is seen that most ships in this route sail at 12–15 knots (70.9% in Leg 1 and 66.2% in Leg 2, respectively). Fig. 15b shows COG distributions at each leg. Obviously, there are significant differences between the traffic in each leg. One may notice two peaks in the histogram of COG, which correspond to the two directions of the legs.

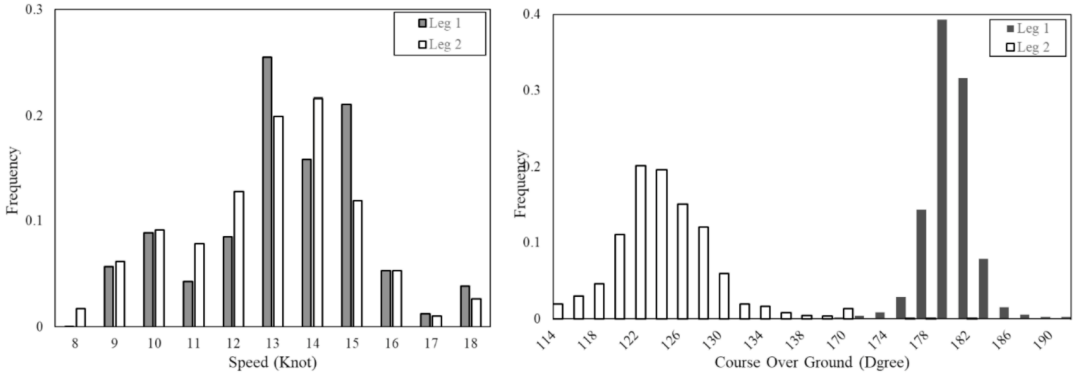
The lateral distribution reflects the spatial extension of the ship routes. Fig. 16 shows the lateral distributions of the ship trajectories in the two legs. On the *x-axis*, the zero value corresponds to the calculated route centreline. A positive value for x means that the vessel sails to the starboard side of the route centreline. The comparison of lateral distribution of Leg 1 and Leg 2 shows that the deviation of lateral distribution in Leg 2 is larger, which means that ship trajectories in Leg 1 are more concentrated on the route centreline. It is found that route width changes along the route (Fig. 14), for example at the north part of Leg 2 the route is narrow whereas at other locations the distribution over the route is relatively wide. This indicates that the lateral distribution along the waterway needs to be investigated. The route width is estimated according to trajectory spatial distribution along the ship route. Once the route centreline is derived, the lateral distributions along legs 1 and 2 are calculated from the available AIS data at every 16 and 5 nautical miles (NM), respectively, as shown in Fig. 17. The leg lateral distribution is approximated by a normal distribution, and then the 95% probability interval of the distribution is adopted as the route width. The grey box indicates the 95% confidence interval estimated by the series of route widths in a given section. According to Fig. 17a, the route width estimated by AIS data varies considerably. The reason is that the uncertainty in ship trajectories varies along the route. The route width distribution in Leg 2 (Fig. 17b) indicates that the route become wider along the route from Turning Section 1 to Turning Section 4. An obvious change is observed from the first 5NM-section to the second 5NM-section. Restricted by the TSS off Cape Roca traffic lane, ships are convergence in the traffic route; therefore, the route width at Leg 2 is narrow (only 2800m in the first 5NM-section). The width of ship route increases steady from 20 nautical miles to 60 nautical miles along Leg 2 and finally reaches at 18500 m.

3.5. Anomaly detection

The characterization of ship behaviour within motion patterns based on historical AIS data provides important information for maritime traffic surveillance. In this study, ship routes are characterized probabilistically and the motion of the ships is described by a series of Gaussian distributions. This normalcy representation supports anomaly detection of ship trajectories that deviate from the route behaviour.

Ship trajectories can be outliers for different types of reasons: ships may sail to irregular place, turn to destination ahead or delayed of schedule and steering to the left or to the right of the main route followed by most of ships. Anomalies could also be the result of accidental situations of ships that lose control due to failure of its propulsion or course keeping equipment (e.g. Wu et al., 2017, 2018a).

As shown in Fig. 18a, a new ship is first associated to the route passing southbound dangerous traffic lane to the Sines Port based on the



(a) Comparison of speed distributions at the two route legs

(b) Comparison of COG distributions at the two legs

Fig. 15. Speed distributions and Course Over Ground (COG) distributions at the two legs.

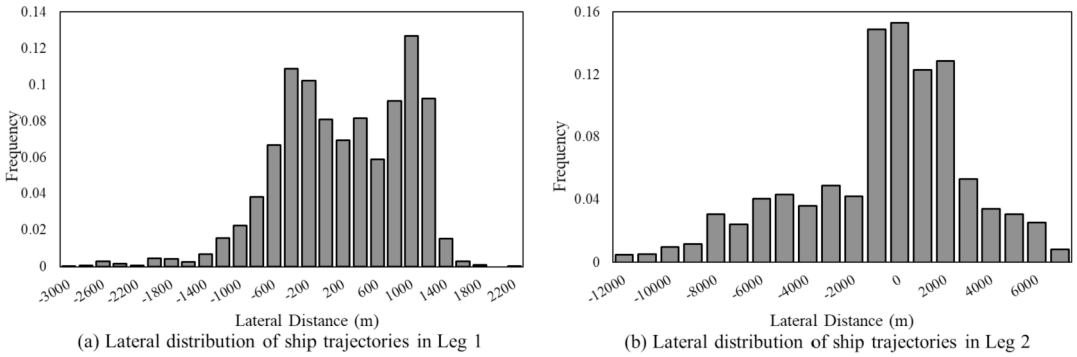


Fig. 16. Lateral distribution of ship trajectories in the two legs.

ship position. It is assumed that the ship lateral position is normal distributed along the ship route, and therefore, the probability that the ship belongs to the route is calculated by the normal probability density distribution:where u_p is the lateral distance of point p and σ is the standard deviation of the lateral distribution at the current section.

Then, off-route behaviour is detected by calculating the value of the Gaussian distribution function at the ship lateral position. If the value is below a threshold, then the ship is detected as not following the route. Fig. 18b shows the ship in-route probability corresponding to its lateral position. The grey area and red dash-line demonstrate the main route and the route centreline followed by most of the ship trajectories. The 95%-probability interval of the lateral traffic distribution is adopted to

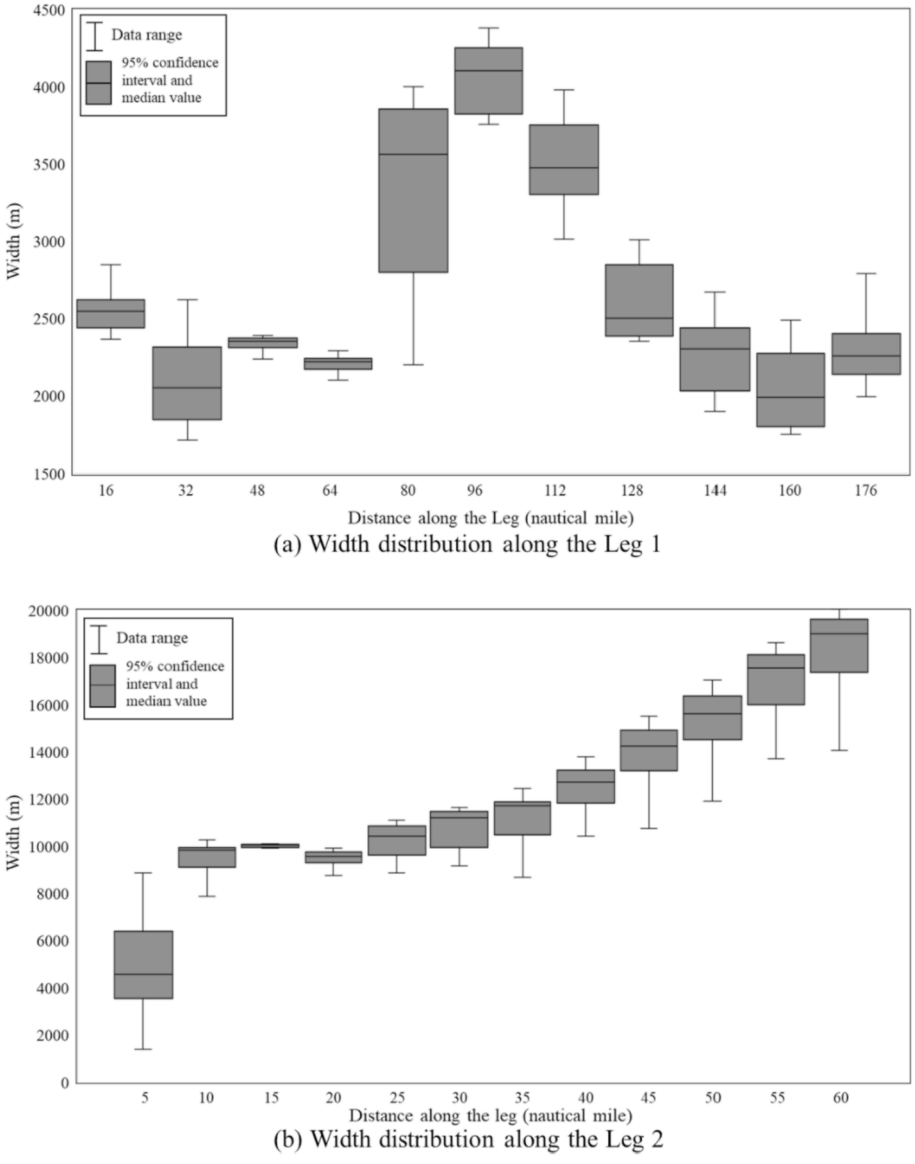
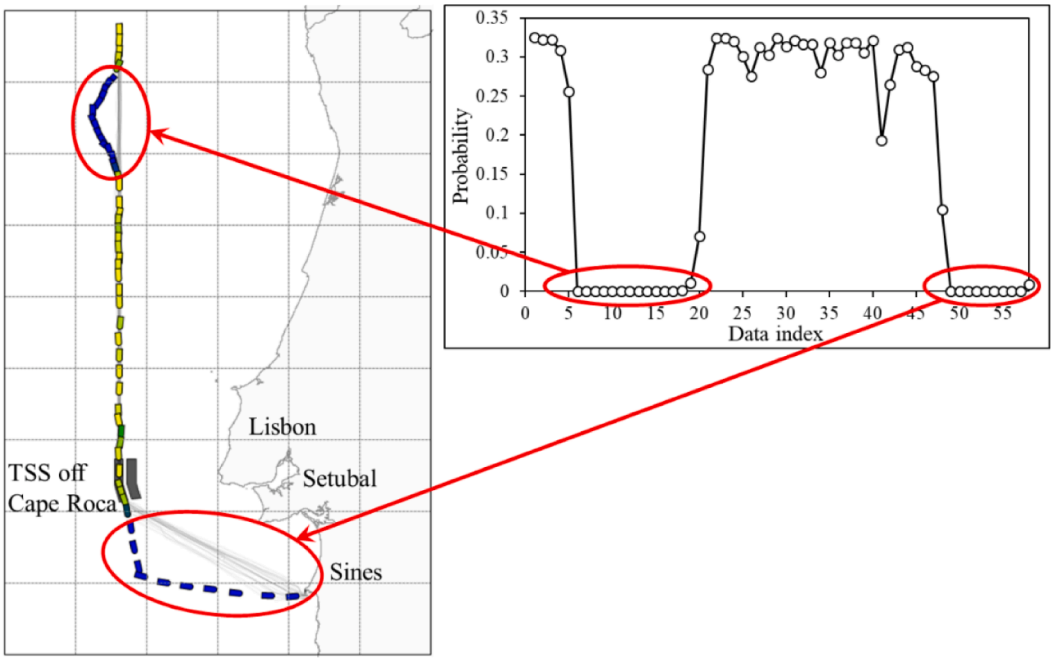
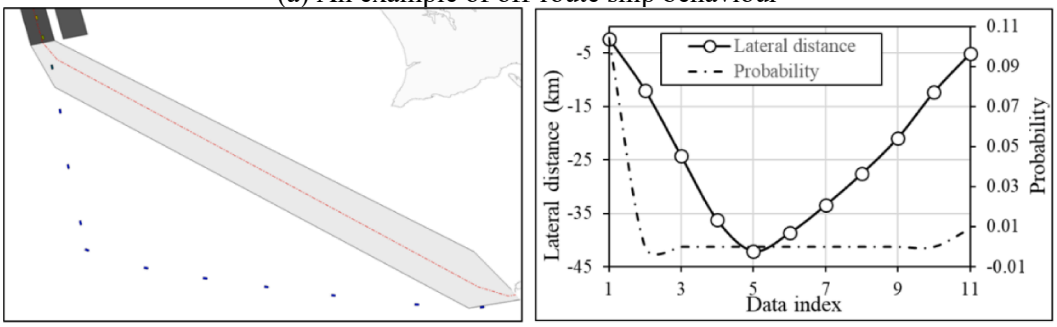


Fig. 17. Width distributions and uncertainty along the two legs.



(a) An example of off-route ship behaviour



(b) Sequential estimation of ship in-route probability and lateral position

Fig. 18. Online outlier detection.

define the boundaries of the waterway. It can be seen that the probability decreases dramatically when the ship trajectory located out of the route.

The off-route behaviour detection results suggest that the method is capable of detecting unexpected departures from regular movements. However, this detection method is a distance-based approach, which means that the velocity and acceleration of the ships are not taken into consideration.

4. Conclusions

The proposed data mining method provides a relatively straightforward and unsupervised approach to determine maritime turning sections and to characterize the maritime traffic in each turning section.

The turning section is automatically identified using ship trajectory compression and turning point detection techniques and clustering methods. In particular, the trajectory compression DP algorithm is adopted for turning point detection and a density-based clustering algorithm is applied for clustering these dispersed points into clusters. The clusters of turning points forms turning sections of the route, which benefits greatly from the previous compression step. The turning section is the geographical area where the turning behaviour is frequently observed. In this study, the Kernel Density Estimation method is applied to estimate the density area based on the identified turning points. Then, relatively high-density areas form turning sections of the ship route. This method is applied to southbound ship trajectories passing the Traffic Separation Scheme off Cape Roca to the ports of Lisbon, Setúbal and Sines. Four main turning sections have been identified at the exit of the TSS off Cape Roca and along the three routes to Lisbon, Setúbal and Sines.

A statistical analysis of the maritime traffic at the turning sections shows that the ship type distribution depends on the route destination. Most cargo ships visit the ports of Lisbon and Setúbal, whereas tankers dominate in the Dangerous traffic lane of TSS off Cape Roca and in the route to Sines. The speed distributions of the ships in each section reveal that ships sailed at higher speed in Turning Section 1 than in other sections. In addition, the average speed of tankers is slightly higher than the average speed of other ship types. In addition, according to the speed distributions divided by the density of ships in each section, it is found that the density of ships does not significantly affect the ship speed.

Tankers pose higher environmental risk and dominate the route passing Southbound Dangerous traffic lane of TSS off Cape Roca to the Port of Sines. Therefore, the characteristics of the maritime traffic in this route is further analysed. Two turning sections haven been identified, dividing this route in two legs. Speed and COG distributions as well as lateral distributions of the ship trajectories have been derived and analysed from the available AIS data. The investigation of the lateral distribution in the two legs indicates that the ship trajectories in Leg 1 are more concentrated on the route centreline compared with Leg 2. The lateral distributions along the ship route show that the route width

changes with the route location. Furthermore, the uncertainty in the route width also varies along the route and a high variability on the ship trajectories is observed at the entrance of the Port of Sines.

The lateral distributions of the maritime traffic along the ship route characterize in a probabilistic manner the route centreline and the route spatial extension followed by most of the ship trajectories. Ship off-route behaviour detection is then tested on an abnormal ship trajectory. The ship is first assigned to a route (ship route passing the southbound dangerous traffic lane to the Sines Port), and off-route behaviour is detected according to the lateral position probability defined by a Gaussian distribution. In this study, the route width is considered as three standard deviations from the mean route and the off-route detection threshold is set as the corresponding probability value. However, the choice of thresholds is a critical aspect of the detection strategy, which should be further investigated.

Author contributions section

Hao Rong: Methodology, Software, Formal analysis, Writing - Original Draft, Visualization.

Angelo Teixeira: Conceptualization, Methodology, Writing - Original Draft, Supervision.

C Guedes Soares: Conceptualization, Methodology, Writing - Review & Editing, Supervision.

Declaration of competing interest

There are no Conflicts of Interest.

Acknowledgements

The paper has been conducted through the project "Integrated System for Traffic Monitoring and Maritime Risk Assessment (MoniRisk)", which has been co-funded by the European Regional Development Fund (Fundo Europeu de Desenvolvimento Regional (FEDER) and by the Portuguese Foundation for Science and Technology (Fundação para a Ciência e a Tecnologia – FCT) under contract no. 028746. This work contributes to the Strategic Research Plan of the Centre for Marine Technology and Ocean Engineering (CENTEC), which is financed by the Portuguese Foundation for Science and Technology (Fundação para a Ciência e Tecnologia - FCT) under contract UID/Multi/00134/2013 - LISBOA-01-0145-FEDER-007629.

References

Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. ACM SIGMOD Record 28 (2), 49–60. https://doi.org/10.1145/304181.304187. Association for Computing Machinery (ACM).

Aarsæther, G.K., Moan, T., 2009. Estimating navigation patterns from AIS. J. Navig. 62 (4), 587–607.

Allianz Global Corporate & Specialty, 2018. Safety and shipping review 2018: an annual review of trends and developments in shipping losses and safety. Allianz.

Chen, P., Huang, Y., Mou, J., van Gelder, P.H.A.J.M., 2019. Probabilistic risk analysis for ship-ship collision: state-of-the-art. Saf. Sci. 117 (September 2018), 108–122.

Chen, Z., Xue, J., Wu, C., Qin, L.Q., Liu, L., Cheng, X., 2018. Classification of vessel motion pattern in inland waterways based on Automatic Identification System. Ocean. Eng. 161, 69–76.

De Vries, G.K.D., Van Someren, M., 2012. Machine learning for vessel trajectories using compression, alignments and domain knowledge. Expert Syst. Appl. 39 (18), 13426–13439.

Douglas, D.H., Peucker, T.K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartographica: The International *Journal for Geographic Information and Geovisualization* 10, 112–122.

Ester, M., Kriegel, H., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231.

Etienne, L., Devogele, T., Bouju, A., 2012. Spatio-temporal trajectory analysis of mobile objects following the same itinerary. Advances in Geo-Spatial Information Science 10, 47–57.

Goerlandt, F., Kujala, P., 2011. Traffic simulation based ship collision probability modeling. Reliab. Eng. Syst. Saf. 96 (1), 91–107.

Gouveia, J.V., Guedes Soares, C., 2010. Oil spill incidents in Portuguese waters. In: Guedes Soares, C., Parunov, J. (Eds.), Advanced Ship Design for Pollution Prevention. Taylor & Francis Group, London, U.K., pp. 217–223

Kang, L., Meng, Q., Liu, Q., 2018. Fundamental diagram of ship traffic in the Singapore Strait. Ocean. Eng. 147, 340–354.

Mou, J.M., Tak, C.V.D., Ligteringen, H., 2010. Study on collision avoidance in busy waterways by using AIS data. Ocean. Eng. 37 (5–6), 483–490.

Norris, A., 2007. AIS implementation – success or failure. J. Navig. 60, 373–389.

Pallotta, G., Vespe, M., Bryan, K., 2013. Vessel pattern knowledge discovery from AIS data: a framework for anomaly detection and route prediction. Entropy 15 (6), 2218–2245.

Pedersen, P.T., 1995. Collision and Grounding Mechanics. The Danish Society of Naval Architects and Marine Engineers, pp. 125–157.

Qu, X., Meng, Q., Li, S., 2011. Ship collision risk assessment for the Singapore Strait. Accid. Anal. Prev. 43 (6), 2030–2036.

Rong, H., Teixeira, A.P., Guedes Soares, C., 2015a. In: Soares, Guedes, Santos (Eds.), Simulation and Analysis of Maritime Traffic in the Tagus River Estuary Using AIS Data. Maritime Technology and Engineering. Taylor & Francis Group, London, pp. 185–193.

Rong, H., Teixeira, A.P., Guedes Soares, C., 2015b. Evaluation of near-collisions in the Tagus River Estuary using a marine traffic simulation model. Sci Journals Marit Univ Szczecin 43, 68–78.

Rong, H., Teixeira, A.P., Guedes Soares, C., 2016. In: Guedes Soares, C., Santos, T.A. (Eds.), Assessment and Characterization of Near Ship Collision Scenarios off the Coast of Portugal. *Maritime Technology And Engineering 3*. Taylor & Francis Group, London, pp. 871–878.

Rong, H., Teixeira, A.P., Guedes Soares, C., 2018. In: Soares, Guedes, Teixeira (Eds.), A Model for Predicting Ship Destination Routes Based on AIS Data. Maritime Transportation and Harvesting of Sea Resources. Taylor & Francis Group, London, pp. 257–264.

Rong, H., Teixeira, A.P., Guedes Soares, C., 2019a. Risk of ship near collision scenarios off the coast of Portugal. In: Beer, M., Zio, E. (Eds.), 29th European Safety and Reliability Conference (ESREL 2019) (Hannover, Germany).

Rong, H., Teixeira, A.P., Guedes Soares, C., 2019b. Ship trajectory uncertainty prediction based on a Gaussian Process model. Ocean. Eng. 182, 499–511.

Rong, H., Teixeira, A.P., Guedes Soares, C., 2020. Collision probability assessment based on uncertainty prediction of ship trajectories. In: Soares, Guedes (Ed.), Developments in the Collision and Grounding of Ships and Offshore Structures. Taylor & Francis Group, London, pp. 283–290.

Sebastião, P., Guedes Soares, C., 2006. Uncertainty in predictions of oil spill trajectories in a coastal zone. J. Mar. Syst. 63 (3–4), 257–269.

Silveira, P.A.M., Teixeira, A.P., Guedes Soares, C., 2013. Use of AIS data to characterise marine traffic patterns and ship collision risk off the coast of Portugal. J. Navig. 66, 879–898.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Routledge.

Sun, F., Deng, Y., Deng, F., Zhu, Q., Chu, H., 2015. Unsupervised maritime traffic pattern extraction from spatial temporal data. In: *11th International Conference On Natural Computation*, pp. 1218–1223.

Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., Huang, G., 2018. Exploiting AIS data for intelligent maritime navigation: a comprehensive survey from data to methodology. IEEE Trans. Intell. Transp. Syst. 19 (5), 1559–1582.

van Dorp, J.R., Merrick, J.R.W., 2011. On a risk management analysis of oil spill risk using maritime transportation system simulation. Ann. Oper. Res. 187 (1), 249–277.

Wu, B., Yan, X., Wang, Y., Zhang, D., Guedes Soares, C., 2017. Three-stage decision-making model under restricted conditions for emergency response to ships not under control. Risk Anal. 37 (12), 2455–2474.

Wu, B., Zong, L., Yan, X., Guedes Soares, C., 2018a. Incorporating evidential reasoning and TOPSIS into group decision-making under uncertainty for handling ships without command. Ocean. Eng. 164, 590–603.

Wu, X., Mehta, A.L., Zaloom, V.A., Craig, B.N., 2016. Analysis of waterway transportation in Southeast Texas waterway based on AIS data. Ocean. Eng. 121, 196–209.

Wu, X., Rahman, A., Zaloom, V.A., 2018b. Study of travel behaviour of vessels in narrow waterways using AIS data–A case study in Sabine-Neches Waterways. Ocean. Eng. 147, 399–413.

Xiao, Z., Ponnambalam, L., Fu, X., Zhang, W., 2017. Maritime traffic probabilistic forecasting based on vessels' waterway patterns and motion behaviors. IEEE Trans. Intell. Transp. Syst. 18 (11), 3122–3134.

Xin, X., Liu, K., Yang, X., Yuan, Z., Zhang, J., 2019. A simulation model for ship navigation in the "Xiazhimen" waterway based on statistical analysis of AIS data. Ocean. Eng. 180, 279–289.

Xu, H., Rong, H., Guedes Soares, C., 2019. Use of AIS data for guidance and control of path-following autonomous vessels. Ocean. Eng. 194, 106635. https://doi.org/10.1016/j.oceaneng.2019.106635.

Yoo, S.L., 2018. Near-miss density map for safe navigation of ships. Ocean. Eng. 163, 15–21.

Zhang, S., Shi, G., Liu, Z., Zhao, Z., Wu, Z., 2018. Data-driven based automatic maritime routing from massive AIS trajectories in the face of disparity. Ocean. Eng. 155 (1), 240–250.

Zhang, W., Goerlandt, F., Montewka, J., Kujala, P., 2015. A method for detecting possible near miss ship collisions from AIS data. Ocean. Eng. 107, 60–69.

Zhang, W., Goerlandt, F., Kujala, P., Wang, Y., 2016. An advanced method for detecting possible near miss ship collisions from AIS data. Ocean. Eng. 124, 141–156.

Zhou, Y., Daamen, W., Vellinga, T., Hoogendoorn, S.P., 2019. Ship classification based on ship behaviour clustering from AIS data. Ocean. Eng. 175, 176–187.