IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 7, 2020, accepted March 1, 2020, date of publication March 9, 2020, date of current version March 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2979612

A Distributed Spatial Method for Modeling Maritime Routes

DIMITRIS ZISSIS[1,2], (Senior Member, IEEE), KONSTANTINOS CHATZIKOKOLAKIS[2], GIANNIS SPILIOPOULOS[2], AND MARIOS VODAS[2], (Member, IEEE)

[1]Department of Product and Systems Design Engineering, University of the Aegean, 84100 Syros, Greece
[2]MarineTraffic, 115 25 Athens, Greece

Corresponding author: Dimitris Zissis (dzissis@aegean.gr)

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant 825070 and Grant 732310.

ABSTRACT In this work we propose a novel spatial knowledge discovery pipeline capable of automatically unravelling the ''roads of the sea'' and maritime traffic patterns by analysing voluminous vessel tracking data, as collected through the Automatic Identification System (AIS). We present a computationally efficient and highly accurate solution, based on a MapReduce approach and unsupervised learning methods, capable of identifying the spatiotemporal dynamics of ship routes and most crucially their characteristics, thus deriving maritime ''patterns of life'' at a global scale, without the reliance on any additional information sources or a priori expert knowledge. Experimental results confirm high accuracy of results and superior performance in comparison to other methods, with the entire processing duration completing in less than 3 hours for more than a terabyte of non-uniform spatial data. Finally, to clearly demonstrate the applicability and impact of our proposed method, we evaluate its ability to detect real world ''anomalies'', such as maritime incidents reported in the European Marine Casualty Information Platform. Numerical results show the advantages of our scheme in terms of accuracy, with an achieved anomaly detection accuracy of higher than 93%, by detecting 313 out of 335 relevant maritime incidents.

INDEX TERMS AIS, anomaly detection, data driven maritime traffic, patterns of life, routes.

I. INTRODUCTION

With more than 80% of the global trade today being carried by sea, shipping routes or ''sea roads'' are vital to the global economy [1]. Sea going vessels follow specific paths when travelling across the vast blue ocean; these roads connecting major ports are some of the busiest places on earth, often only a few kilometres wide, scattered with many physical constraints (e.g., reefs), where enormous vessels perform risky manoeuvres under constantly changing environmental conditions (e.g., wind, sea currents). These waterways form a global maritime exchange network. More than often these connections are not direct lines (e.g., the shortest distance from the point of departure to destination), but ''climatological routes'' along which higher speeds can be achieved due to the existence of currents or the prevalence of wind, sea or swell.

The associate editor coordinating the review of this manuscript and approving it for publication was Xinyu Du.

Sea roads though are not paved in concrete, as the location of the connector, its width and its content, can vary significantly over space and time, under the influence of various trade and carrier patterns, but also due to large infrastructure investments (e.g., canal expansions), climate changes (e.g., global warming), traffic restrictions (e.g., Emission Control Areas), political events and other international incidents (e.g., increase of piracy in specific regions). Accurately modelling this network and the spatiotemporal characteristics of ship traffic patterns is vital for improving maritime decision making and for applications such as navigation, traffic optimisation, policy making, environmental impact assessment and many more. Especially with respect to safety and security, a well-defined network of connections makes it possible to detect vessels deviating from normalcy. As such, anomaly detection can be understood as a method that supports situational assessment by indicating objects and situations that, in some sense, deviate from the expected, known or ''normal'' behaviour. But in the last century, less than a dozen

maps of worldwide maritime flows based on actual shipping data were published [2].

While in the past this could have been attributed to the lack of accurate surveillance data, today this is not the case. Nowadays, a multitude of tracking systems produce massive amounts of maritime data on a daily basis. The most commonly used is the Automatic Identification System (AIS), a collaborative, self-reporting system that allows vessels to broadcast their identification information, characteristics and destination, along with other information originating from on-board devices and sensors, such as location, speed and heading [3]. AIS messages are broadcast periodically and can be received by other vessels equipped with AIS transceivers, as well as by on the ground or satellite-based sensors. Since becoming obligatory by the International Maritime Organisation (IMO) for vessels above 300 gross tonnage to carry AIS transponders, large datasets are gradually becoming available and are now being considered as a valid method for maritime intelligence [4].

Towards this end, there is a growing body of literature on methods of exploiting AIS data for safety and optimisation of seafaring, namely traffic analysis, anomaly detection, route extraction and prediction, collision detection, path planning, weather routing, etc., [5]. Route definition and motion pattern extraction is used as a precursor for trajectory forecasting and anomaly detection. Perceiving and comprehending elements and their contextual meaning in the environment within a given volume of time and space, while projecting their status into a future timeframe, is a critical element of Maritime Domain Awareness (MDA). MDA is the effective understanding of activities, events and threats in the maritime environment that could impact global safety, security, economic activity or the environment.

As the amount of available AIS data grows to massive scales, researchers are realising that computational techniques must contend with difficulties faced when acquiring, storing, and processing the data. Applying traditional data processing techniques can lead to processing times of several days, if applied to global data sets of considerable size. In addition to this, conventional algorithms are proving incapable of dealing with the uncertainty and partial truth present in such datasets [6], [7]. As such, AIS datasets present some unique characteristics and difficulties. For example, the update interval for AIS is not constant (as it would be expected in trajectories generated with traditional positional trackers), but dependent on ships' behaviour. More specifically, it is common for a vessel to broadcast its data every three minutes when moving with less than 3 knots, while transmission period decreases to two seconds when travelling with more than 14 knots or when changing course. Therefore, in such occasions, the collected positions are spatially and temporally closer. Additionally, vessels may often travel in and out of the coastal-based receivers' network coverage, leading to surveillance gaps that range from several minutes to many hours. Especially in open seas, where coastal reception is limited, satellite AIS coverage can be used for tracking; in this case though, delays of several hours are not uncommon. Most existing studies only focus on analysing traffic patterns in ports or coastal areas, where these issues are not present, lacking to address the data uncertainty issue [8].

A number of empirical studies, attempt to support the understanding of maritime behaviour by generating heat maps or density maps of the data for operator engagement and supporting stakeholder decision making [9]–[11]. However, huge amounts of data overwhelm the end user and challenge the established analysis workflows, thus requiring more complex approaches to deal with the volume and veracity of implicated data [12]. Although heatmaps can be a useful visualisation tool, the resulting visualisations can be heavily affected by gaps inherent in the dataset; either due to limited network coverage in specific geographical regions or simply by rarely travelled areas. Consequently, important connections (i.e., commercial routes) may not be detected. A number of publications rely on statistics to generate simple analytics of ship traffic and frequencies [13], [14]. More complex solutions can be categorised to (i) grid based approaches or (ii) methods of using vectorial representations of traffic.

In grid based approaches the area of coverage is split into cells, which are characterised by the motion properties of the crossing vessels to create a spatial grid (e.g., [15]–[21]). For example in [15] a data-driven methodology is proposed to estimate the vessel arrival times in port areas. Grid based anomaly detection algorithms include Fuzzy ARTMAP [16], Holst Model [17], [18], Support Vector Machine [22], GeoHash encoding [23] and others [19]. In [18], Ristic *et al.*, use kernel density estimation (KDE, also known as Parzen Window estimation) to learn the distribution of kinematic variables (position and velocity) in each cell. Grid based methods have been considered effective only for small area surveillance where the majority of literature was focused and the computational burden was regarded as its limitation when increasing the scale [20], as well as the need for a priori selection of the optimal cell size. In areas characterised by complex traffic, like intersecting sea lanes, the resulting multi-modal behavioural description would lead to complex algorithms to perform anomaly detection. In [19], Wu *et al.*, demonstrate the ability of a grid based method for computing shipping density, fast enough to be performed at a global scale. Specifically, 33 months of data (i.e., from August 2012 to April 2015) were used to produce global monthly ship density, traffic density and AIS receiving frequency maps in 56 hours of processing.

In the second category, vessel trajectories are modelled as a set of connected waypoints. Thus, vessel motions in large areas (e.g., at a global scale) can be managed thanks to the high compactness of the waypoint representation [24], [25]. In [26], Mazzarella *et al.*, apply a Bayesian vessel prediction algorithm based on a Particle Filter (PF) on AIS data. Pallotta *et al.*, present the TREAD (Traffic Route Extraction and Anomaly Detection) methodology, which relies on the DBSCAN algorithm for automatically detecting anomalies and projecting current trajectories and patterns into the

future [20]. In [24], Li *et al.*, present a two-step method to achieve a balance between computational time and performance; first performing data simplification by applying the Douglas-Peucker (DP) algorithm before processing the simplified trajectories with Kernel Density Estimation. Similarly in [27] a DBSCAN is used for clustering purposes, before Ant Colony algorithm is used to find the optimal path from the starting turning node to the ending turning node. In [13], a big data analytical approach that analyses ship traffic demand and the spatiotemporal dynamics of ship traffic in Singapore's port waters using big AIS data is described. Recently [28], [29] presented an approach to learn automatically and represent compactly commercial maritime traffic in form of a graph, whose nodes represent clusters of waypoints, which are connected together by a network of navigational legs (graph edges). This work also focuses on a limited geographical region, specifically the Iberian Coast and the English Channel where network coverage is potentially good and there are few gaps in the dataset. In [30], Breithaupt *et al.*, attempt to compute vessel routes between ports and to delineate route boundaries for a dataset covering the Atlantic coast of the US over a ten year time span. This approach resulted in continuous but not smooth boundaries that in some cases had boundary lines changing abruptly at transects. A combined trajectory classification and long short-term memory (LSTM) networks framework is proposed in [23] where the longest common subsequence algorithm is used to measure similarity when performing trajectory clustering with DBSCAN. Clustered trajectories indicate vessels' mobility patterns that are further modelled via LSTM networks for long-term prediction. However, the algorithm is applied in small geographical area and the prediction is accurate only for few minutes ahead.

Table 1 below summarises the shortcomings that recent studies have faced. In summary, our approach attempts to overcome those shortcomings, by i) increasing accuracy while avoiding information loss and dealing with data uncertainty present at global scales, ii) adopting big data processing techniques so as to minimise processing time and iii) making use of methods which do not require manual tuning or a priori expert knowledge as this would impact the generalisation capacity of the proposed solution.

A. RESEARCH CONTRIBUTIONS

In this paper we present a distributed data driven approach for uncovering the ''roads of the sea'' and maritime traffic patterns. We propose a maritime data modelling methodology, named **ROTA** (Maritime **RO**ute Ex**T**r**A**ctor), capable of accurately extracting ''origin to destination'' connections and their spatial characteristics at a global scale automatically, without the reliance on any additional information sources (e.g., nautical maps) or a priori knowledge. The main contribution of this work is to address the challenge of automatically generating an accurate data driven representation of maritime traffic. Our novel methodological contribution shows how to overcome big data challenges (i.e., processing huge volumes of uncertain data that arrive at unprecedented speed) and transform surveillance data into a representative model of vessel traffic patterns. The proposed methodology addresses a number of important requirements and literature shortcomings. More specifically, our key contributions are the following:

1) The majority of related literature is focused on specific geographic areas or short time periods as presented in Table 1, where the volume and veracity of the data is manageable with traditional architectural and algorithmic approaches. ROTA is capable of extracting traffic patterns at a global scale from non-uniform spatial and irregular temporal data distributions without requiring manual tuning or expert knowledge. ROTA proposes a spatial big data processing pipeline capable of automatically processing huge amounts of global data in a short time period. The entire maritime traffic network can be produced on request in a few hours by using big data technologies (Spark and MapReduce) on a cluster of distributed computing nodes so as to depict any changes in the network.
2) Most of the related literature define ''sea routes'' as thin lines connecting ports across the globe, causing substantial information loss. In the real world though, ships do not travel on thin lines, as traffic corridors have a variable width, volume, and distribution. ROTA is capable of identifying the specific characteristics of each corridor, and fundamentally the spatial variations of ship traffic dynamics. For our work, we adopt the guidelines published by the UK Department of Trade & Industry, the Department for Transport and the Maritime & Coastguard Agency, according to which a route's width should be such that the route accommodates 95% of all traffic transiting it [31]. Additionally, ROTA is capable of defining contextual routes capturing variations in traffic patterns e.g., routes followed by different categories of vessels under specific conditions. Thus, ROTA supports understanding the organic behaviour of maritime traffic and ''patterns of life'', a feature currently lacking from the majority of navigation charts [12].
3) ROTA is tested against a series of maritime security incidents officially published by the European Maritime Security Agency achieving highly accurate results. To the best of our knowledge this contribution is the only approach that provides evaluation against ground truth.

B. ORGANISATION

The structure of this paper is as follows. Section II presents the proposed knowledge discovery approach and the distributed method for maritime data modelling, capable of automatically extracting ship routes and their spatial characteristics at a global scale. Following this, we report on the experimental and evaluation results in Section III and analyse the results from five real world serious maritime incidents.

TABLE 1. Previous research methods for route extraction and anomaly detection.

Ref.	R. / G.	B. D.	S.	P.T.	R. R.	E. I.	I. L. H.	G. T. E.
[9]	Regional	No	No	N/A	Line interpolation leading to heatmap	N/A	No	No
[10]	Regional	No	No	N/A	Heatmap	N/A	No	No
[11]	Regional	No	No	60h for 1 year of data	Heatmap	N/A	No	No
[13]	Regional	Yes	Time	N/A	Line Interpolation	Yes	Partially with line interpolation	No
[14]	Regional	No	No	N/A	No	N/A	No	No
[15]	Regional	No	No	N/A	Grid based interpolation	Partially. Automated but dependent on a set of parameters optimised for the area under investigation	No	No
[16]	Regional	No	No	N/A	Grid based solution for incremental learning of vessel behaviour.	Partially. The learning procedure can be annotated from an operator	No	No
[17]	Regional	No	No	N/A	Grid based solution	Yes. Anomaly threshold is manually tuned by the operator.	N/A	No
[18]	Regional	No	No	N/A	No	No. KDE is used to determine thresholds that partition space to two regions (i.e., normal and anomalous behaviour)	No	No
[19]	Global	No	Yes	56 hours	No. Heatmap produced through a grid based solution	N/A	No, but MMSI spoofing is applied	No
[20]	Regional	No	Not tested	N/A	Sequence of successive waypoints	No. Unsupervised and incremental learning	Yes, through waypoints.	No
[21]	Regional	Yes	Yes	N/A	GeoHash encoding filtering and Delaunay Triangulation	N/A	No	No
[22]	Regional	No	No	N/A	Line interpolation	Yes. Categorisation of normal tracks based on visual analysis process	No	No
[23]	Regional	No	No	N/A	Line interpolation and clustering of common subsequences of trajectories	No	No	No
[24]	Regional	No	No	3-4 min. but tested against small dataset.	Cubic spline interpolation, heatmap of vessel trajectories	Yes	No	No
[25]	Global	No	No	N/A	Route density map	Yes	No	No
[26]	Regional	No	No	N/A	Linear interpolation	No. Unsupervised construction of traffic patterns.	Yes	No
[27]	Regional	No	No	N/A	Interpolation of characteristic points after Douglas Peucker algorithm	Yes. Manual parameter selection.	No	No
[28]	Regional	No	No	N/A	Directed graph connecting density-based clusters of positions	Partially. Clustering parameters are applicable for the region but not generalised to global dataset.	No	No
[29]	Regional	No	No	N/A	Not applicable. Prediction of long-term target state based on a target motion model.	No	No	No
[30]	Regional	No	No	N/A	Line Interpolation leading to density maps	N/A	Partially. Gaps longer than one day cause voyage split.	No

R. / G.= Regional or Global geographical coverage, B.D.= Uses Big Data technologies, S.= Scalable in time, area, P.T.= Processing Time, R.R.= Route Reconstruction method, E.I.= Needs Expert Input, I.L.H. = Information Loss Handling, G.T.E. = Ground Truth evaluation

Then, we discuss the merits and the limitations of our work in Section IV, before concluding this research in Section V.

II. BIG DATA KNOWLEDGE DISCOVERY APPROACH

Knowledge discovery from data (KDD) and data mining are not new topics; The first knowledge discovery in databases workshop was held back in 1989 in Detroit, during IJCAI-89. The basic problem addressed by the KDD process is one of mapping low-level data into other forms that might be more compact (e.g., a short report), more abstract (e.g., a descriptive approximation or model of the process that generated the data), or more useful (e.g., a predictive model for estimating the value of future cases) [32]. At the core of this process is data mining, an essential step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data. Hence, extracting patterns also means fitting a model to the

data, finding structures, or in general any high-level description of a set of data. The fitted models play the role of inferred knowledge [32]. As previously discussed, nowadays KDD is hindered by the volume, velocity, variety, and veracity of data produced by numerous distributed real-world sensors and systems.

Within this context, ROTA is a maritime big data modelling approach, capable of accurately identifying the spatiotemporal dynamics of ship routes and most crucially their characteristics (such as route variable width, types of vessels, direction of travel, etc.), thus deriving the maritime ''patterns of life'' at a global scale, without the reliance on any additional information sources or a priori knowledge. ''Patterns of Life'' are understood as observable human activities that can be described as patterns in the maritime domain related to a specific activity (e.g., fishing) taking place at a specific time and place [33]. Essentially, vessel based maritime activity can be described in space and time, while classified to a number of known activities at sea (fishing etc). The spatial element describes recognised areas where maritime activity takes place; thus, including ports, fishing grounds, offshore energy infrastructure, dredging areas, etc. The transit paths to and from these areas also describe the spatial element (e.g., commercial shipping, ferry routes, etc.), while the temporal element often holds additional information for categorising these activities (e.g., fishing period, time of year, etc.). The development of an accurate network for modelling these activities, enables the deployment of relevant descriptive and predictive analytics, which are critical for improved situational awareness and anomaly detection.

The data used in this study includes three sources: a) AIS data covering the entire globe for an extended period of 2 years (January 2016 till December 2017), b) port geometries as provided by the World Port Index from the National Geospatial Intelligence Agency [34] and c) information regarding nautical accidents and incidents which took place in European waters. This dataset is provided by the European Maritime Safety Agency and used for validation purposes [35]. Regarding the AIS data, there is an upper limit of 64 possible types of messages that AIS transceivers can exchange [3]. These message types may be related to vessel position tracking, vessel's identification or voyage information. In our study we focus on types 1-3, 18, 19 that are linked to tracking vessel positions and type 5 which comprises vessel identification and voyage information. Table 2 provides an overview of the all the AIS information fields taken into account in our analysis. Our AIS dataset contains approximately 9 billion positions, broadcast from more than 200,000 ships of all ship types. Although the collected data includes all the required information to identify spatiotemporally the operation of each ship, significant processing is needed to extract additional value. In the following subsections we provide a detailed analysis of the approach followed and the impediments ROTA had to overcome for effectively constructing an accurate representative model of vessel traffic patterns. A high level overview of the spatial knowledge discovery pipeline is provided in Fig. 1.

TABLE 2. AIS data fields description [3].

Name	Description	Range
Maritime Mobile Service Identity	Identification number for the vessel	-
Rate of turn	(Right or left) Turn angle of vessel	0 to 720 degrees with minute resolution
Speed over ground	Ship's speed measured in knots	0 to 102 knots with 0.1-knot resolution
Position Co-ordinates	Vessel's latitude and longitude	Latitude ranges from -90 to 90 and longitude from -180 to 180. Both are provided with up to 0.0001 minutes accuracy
Course over ground	Vessel's motion direction relative to the magnetic north pole	0 to 359 degrees with 0.1 minute resolution
Heading	Vessel's heading direction relative to the magnetic north pole	0 to 359 degrees
International Maritime Organisation number	9-digit number that is assigned by IHS Maritime (Information Handling Services) when a commercial vessel is constructed	-
Destination	The vessel's destination that is manually inserted by the crew members	Free text, up to 20 characters
Type	The vessel's type id	0-255 code that is mapped to its type (e.g., tanker, passenger, etc.)
Dimensions	Dimensions of ship in meters	Four integers indicating dimension to bow, dimension to stern, dimension to port (i.e., left side of the vessel when facing the bow), and dimension to starboard (i.e., right side of the vessel when facing the bow)
Name	The vessel's name which is manually inserted by the crew members	Free text, up to 20 characters

A. DATA PREPROCESSING: ORIGIN-DESTINATION ASSIGNMENT

A safe ship journey begins and ends at a sea port (or an anchorage within or close to the port's operational area). An essential preprocessing task is assigning to all positional data collected through AIS, origin-destination information. Although AIS messages often include a destination port, this field is ignored in our study, as it is manually entered by each vessel's crew, without following a specific standard, making it thus prone to errors. For this purpose we recalculate destination and departure ports by making use of the World Port Index dataset, which contains the location and physical characteristics of major ports and terminals worldwide [34]. We execute a spatial query to assess intersections of port geometries (or operational areas) with vessel positions.

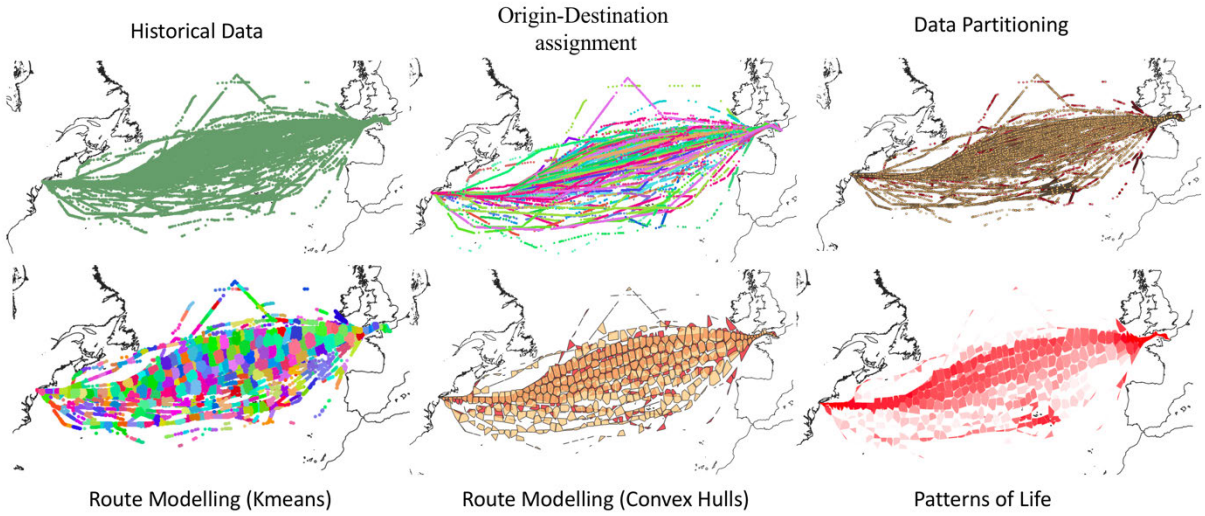

FIGURE 1. ROTA a spatial knowledge discovery pipeline. The above figure depicts the several steps the methodology proposes and are analysed in detail in the following sections.

All the positions that intersect with a port geometry are assigned the corresponding geometry unique identifier (i.e., port id). Then, ROTA sorts data per ship id and timestamp and for each consecutive pair of messages with the same ship id, it detects changes in port id, to determine port call events (i.e., departures / arrivals). As depicted also in Table 3, four different cases may occur for each pair of consecutive messages received:

- Only the 1st message has a port id assigned: Since the messages are sorted chronologically, this case indicates that the vessel moved out of the port's geometry, which is marked as a port departure event.
- Only the 2nd message has a port id assigned: Reversely to the previous case, this one indicates that a vessel arrived at a port.
- Both messages have the same port id: This means that the vessel was moving inside a port.
- Neither of the messages has a port id assigned: This indicates that the vessel was travelling in open sea.

Following this, all vessel positions that are between departure and arrival time are considered as part of the same voyage. As a second curation step, voyages are automatically confirmed and any erroneous entries are removed. For instance, a vessel that is travelling at open sea, may intersect with a port geometry without necessarily calling at that port (e.g., unintended crossing). This may occur due to a complex port geometry in relation to the trajectories vessels follow in the given area. Such a case would lead to a port arrival event followed by a port departure event, that would split incorrectly the vessel's voyage into two different voyages. However, when a vessel arrives at its destination port, it is obliged to reduce speed as it moves inside the port, and typically will spend several hours (or even days) at the port. Thus, based on the state changes described afore we generate sessions (in-port stay) by binding positions and calculating aggregates of each port-call session such as minimum, maximum and average speed of the vessel and in-port stay duration. The latter will let us distinguish the actual port call events and merge parts of voyages that were split due to unintended crossings. Another case of unintended crossing occurs when anchored vessels drift and cross multiple times a geometry boundary due to weather conditions. This behaviour results in falsely identifying multiple port visits, though for some vessels (e.g., pilot vessels, fishing boats, etc.) this can be considered as typical behaviour. Consequently, additional preprocessing and data cleaning steps are performed in our work (both on the static and dynamic data), such as:

- Evaluating for each data field whether it is complete and determining its integrity.
- Identifying for each vessel if dynamic positional reports received are possible based on kinematic equations.
- Removing entries with empty, erroneous, inconsistent (e.g., ship characteristics such as ship dimensions, changing during a voyage) or conflicting fields (e.g., ship name changing during voyage).

The majority of these processes are beyond of the scope of the current paper, but are adequately documented in related papers such as [36].

TABLE 3. Port call events.

Case	Message (t-1)	Message (t)	Travelling Status	Port Move
#1	In Port(A)	Not in Port	Departure from Port(A) at t (time)	True
#2	Not in Port	In Port(A)	Arrival at Port(A) at t (time)	True
#3	In Port(A)	In Port(A)	In Port(A)	False
#4	Not in Port	Not in Port	Travelling at open sea	False

B. DATA PARTITIONING AND COMMON ROUTE EXTRACTION

After the initial curation process all the positional data are linked with a specific journey (voyage id). However, vessel

voyages should be grouped according to their departure and their arrival port to perform route analysis. Furthermore, vessel type limitations should be applied when extracting ''Origin to Destination'' routes. For instance, tankers and cargo ships are not capable of sailing in areas with shallow waters and thus different routes will be followed in each case. Thus, all the positions of each voyage are assigned an identifier that uniquely characterises the ''Origin to Destination'' connection for the specific type of vessel. Due to the volume of related data, a distributed MapReduce programming approach is adopted for this step, executed on a spark cluster [37]. During the map phase each position is transformed into a key-value pair where the key is the identifier and the value is the message itself. Then, taking advantage of the inherent parallelisation of Spark, a reduce-by-key method is applied upon the data to group them into lists, each one containing a ''common route''. A common route comprises all the messages of the voyages that have the same identifier, meaning that they have the same vessel type, departure port and arrival port. The data is processed and stored in a distributed fashion across the worker nodes in the form of those key-value pairs. ROTA leverages the distributed architecture of Spark to process multiple routes simultaneously in different nodes. After this step, we filter out all messages that correspond to in-port movements and thus are not useful for our analysis. The resulting data set comprises of more than 1 million unique connections distinct by ''Origin to Destination'' and vessel type.

C. ROUTE MODELLING

During the previous step, the original dataset was segmented into a set of partitions where keys corresponded to a unique ''Origin to Destination'' connection per vessel type (route) and values contained all AIS messages. The next step was to build a model from the positional data for every ''Origin to Destination'' connection. For this, we apply a clustering algorithm with the aim of reducing the size of the initial dataset and generating a model of normalcy for the trajectories. Most models and approaches reviewed for AIS trajectory reconstruction rely on density-based clustering algorithms (such as DBSCAN). Due to the lack of temporal and spatial uniformity existent in global AIS datasets, density-based clustering algorithms underperform or require case by case parameter selection. For example, in coastal areas the spatial and temporal distance between the collected positions is much smaller as opposed to open sea journeys where the lack of coverage can create much sparsely defined trajectories (e.g., a single position received in several hours). Thus, the high variance of density poses a serious obstacle to using a density-based clustering method for identifying the regions, mainly because it becomes difficult to find a suitable threshold to provide as a parameter to those methods. Moreover, a density-based method would identify very few or no regions, based on the threshold, in sparse areas that are commonly found in the parts of the route that lack terrestrial coverage. This kind of output is not suitable for defining the traffic patterns of vessels, since there would be large gaps in the related routes.

Partitioning clustering methods, on the other hand, do not require setting any threshold for the density. Though, they do require a user defined parameter, (k) that indicates the number of clusters to be identified. However it is possible to provide an automatic estimate of the number of clusters by making it proportional to the number of points of each ''Origin to Destination'' route, so their importance is proportional to their density, using the following formula:

$$k = max\{min\{\lfloor\sqrt{N}\rfloor, k_{max}\}, 1\}; \quad (1)$$

where N is the number of points and k_{max} defines the upper limit of clusters and equals 300.

The partitioning clustering algorithm we use in our methodology is an adaptation of K-Means, using the previous formulation to automatically estimate the k parameter to cluster AIS positions based on longitude and latitude [38]. To minimise the maximum diameter of clusters detected, increase stability and computational efficiency of our approach, regardless of the spatial distribution of the positions on each route, we provide the k farthest positions of each route as initial centroids to the k-means algorithm, instead of using a random selection [39]. Then, the rest of the positions (i.e., pairs of latitude and logitude values) of each route are clustered using the route's centroids. Moreover, we take full advantage of the parallel processing of routes, defining the upper bound k_{max} of clusters to a fixed number, which in turn bounds the maximum processing time required per partition. For the specific dataset used in this study, $k_{max} = 300$ is a non reachable upper limit for more than 99% of connections. Bounding the number of clusters to an upper limit allows us to control the expected size of the output and minimise processing time, without any significant loss in accuracy. For most of the routes the number of clusters will be equal to the square-root of the number of points included in the route. Square-root is a simple empirical method of finding number of clusters that allows to have an adaptive threshold per route.

D. EXTRACTING THE GLOBAL NETWORK AND PATTERNS OF LIFE

The result of the previous step is a collection of k clusters calculated from the positions of vessels departing from one port to another for each ship type. A ''sea road'' can be represented as a collection of high level geometries e.g., convex hulls or polygons that consist of the positions that belong to each cluster (Fig. 2). This kind of representation is particularly suitable for anomaly detection, e.g., identifying positions that fall outside of convex hulls can be considered as unusual, raising an alert (Fig. 3).

According to the report published by the UK Department of Trade & Industry (DTI), in co-operation with the Department for Transport (DfT) and the Maritime & Coastguard Agency (MCA) [31], a route width should accommodate 95% of all traffic transiting each route, while it is noted that this

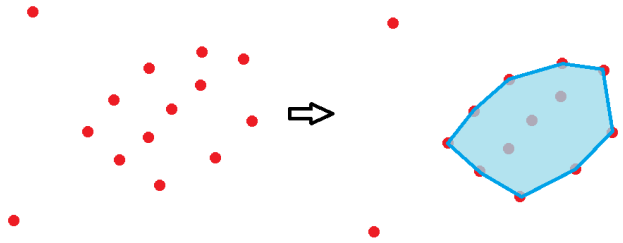

FIGURE 2. A depiction of convex hulls or polygons that consist of the positions that belong to each cluster. These are then used to model sea routes (belonging to the same departure and destination port) across the globe.

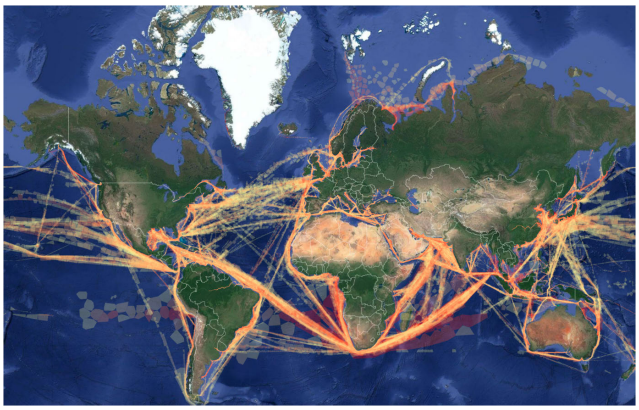

FIGURE 3. A random selection of roads of sea. Road of sea as represented as set of convex hulls. Colours depict quantiles of global distribution of the calculated density (yellow is low density and red is high density).

process will result in variable route widths (dependent upon the traffic activity). It also suggested that, where appropriate, route widths should encompass the lateral deviation associated with +/−2 standard deviations of the displacement of the traffic associated with movement between two locations. In [30], the difficulty of calculating smooth vessel routes and their accurate boundaries was identified by the authors, while adopting an approach which could deal with uncertainties and errors in the data. A convex hull is the smallest (w.r.t. area) polygon created from a set of points that encloses all of them. Convex hulls are the most accurate geometric shapes we can use in order to represent the area clusters of positions occupy.

The adapted K-Means clustering algorithm, used to produce the clusters-convex hulls, assigns every position to a cluster (no outliers); therefore, the initially formed clusters will cover every part of the regions for which training data exists. It is safe to assume, and indeed our experiments validate this, that not all convex hulls need to be retained to accurately represent a ''sea road''. To discard the unnecessary convex hulls, we create an indicator of the density of a convex hull, calculated using the following formula:

$$Density = \frac{N_{points}(convex\ hull)}{Area(convex\ hull)} \times \frac{N_{trips}(convex\ hull)}{N_{trips}(route)} \quad (2)$$

Convex hulls with high densities indicate areas of high concentration of vessel traffic. Using this density metric and a threshold, provided by the user, our method selects which convex hulls with low density to discard. The threshold regulates the extent to which the coverage will be affected. As an additional step we calculate for each convex hull, specific voyage characteristics; such as median speed, number of turns, entry exit points etc.

III. EVALUATION AND VALIDATION

In this section we present numerical results on accuracy categorised per vessel type. As a first validation step we divide our dataset into a training and a test set separately for each connection. The hypothesis here is that, the majority of vessel positions future to the cut-off date used for the route definitions should fall within the convex hulls (or polygons) of that sea road.

We maintain a 70/30 ratio for each split for each port-connection and vessel type, while we maintain the original order of appearance of transmitted messages (i.e., the last 30% of messages that appeared in a specific connection are used as the test set). We calculate the accuracy of our methodology by computing the percentage of the positions in the test set that fall within the convex hulls produced only from the training set:

$$Accuracy = \frac{N_{points}(test\ set\ \textbf{\textit{intersect}}\ convex\ hulls)}{N_{points}(test\ set)} \quad (3)$$

The aggregated training and test sets include almost 4 billion and 1,5 billion messages respectively. The output of our method is consisted of 35 million convex-hulls in total and approximately 30 convex-hulls on average per connection and vessel type. The resulting dataset corresponds to more than 1 million unique combinations out of which only the 18% correspond to periodically transited routes (i.e., there are at least 10 repetitions of the route within the training set). We consider sea routes with less than 10 trips within our dataset to be out of the scope of this work, as these are not common behavioural patterns and routes.

Fig. 4 which depicts the accuracy of our method is highly correlated to the number of trips for specific type of vessels. Moreover, the breakdown of routes per vessel type exposes some characteristics of different vessel types (i.e., AIS reported types) and their markets. We focus on three key shipping markets (i.e., ''PASSENGER SHIPS'',''CARGO'',''TANKER'') that follow repetitive patterns due to the nature of their operations. As expected, ship routes are governed from the characteristics of the market segment they serve; for example, commodity transportation frequency and travelling times are regulated by chartering conditions, while passenger transportation is affected by seasonality. It is evident that vessels in the first category converge to 95% of accuracy earlier than the later, while the later achieves higher accuracy on average due to highest average number of vessels on the same route. Other vessel types that do not follow repetitive patterns (such as tug boats, fishing vessels etc.) were excluded from our analysis. Further analysis and classification of the generic ship types and markets,

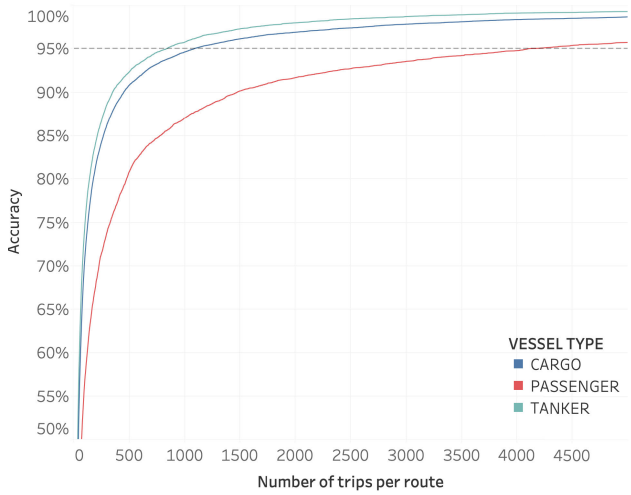

FIGURE 4. Cumulative distribution function of accuracy per number of trips per route.

TABLE 4. Results and statistics for routes with more than 10 trips.

Vessel Type	On Average	
	# Trips	Accuracy
PASSENGER (60-69)	2162,18	0,8354
TANKER (80-89)	553,04	0,7372
CARGO (70-79)	1563,28	0,7767
	4278,5	**0,7831**
TANKER (80-89) SPECIALISED MARKETS		
LNG CARRIERS	46,54	0,7246
LPG CARRIERS	71,58	0,7228
WET BULK	434,92	0,7643
		0,7372
CARGO (70-79) SPECIALISED MARKETS		
DRYBULK	87,25	0,7376
DRYBREAKBULK	1145,38	0,7799
RO/RO	109,51	0,7920
CONTAINERSHIPS	221,14	0,7973
		0,7767

into more specialised ones, results in a better understanding of the achieved accuracy and patterns followed. In Table 4, we split the generic vessel classification into a more specialised categorisation based on similar vessel characteristics.

The overall accuracy of our approach is depicted in Table 4 where accuracy is above 73% for each vessel type and on average 78% for all vessel types. So as to be able to compare ROTA with other state of the art approaches, we calculate ROTA accuracy in specific geographical areas such as those reported in [20]. Although an overall global accuracy is not provided in this work, according to the authors [20], accuracy in the Strait of Gibraltar (an area of high traffic density) was 95%, decreasing to 70% in the North Adriatic Sea, declining further to only 40% in the Indian Ocean due to a lack of traffic constraints over a large area combined with the low update rates of satellite-based AIS data. In comparison, ROTA, is capable of 95% accuracy in the Strait of Gibraltar, 80% in the North Adriatic and 70% in the Indian Sea. In terms of performance, overall execution time is less than 3 hours,

TABLE 5. Real world accuracy as measured against reported incidents.

Type of Incident	Reported	Detected	Accuracy
Collision	127	123	96,85
Grounding	80	71	88,75
Loss of Control	128	117	91,40
TOTAL	335	313	93,43

indicating that ROTA is at least one order of magnitude faster than related approaches [19].

To further demonstrate the relevance of ROTA for maritime anomaly detection and practically validate our hypothesis that defining normal vessel behaviour can assist in detecting deviations, we test our method against data from real-world maritime incidents. Real anomalies leading to accidents or being indicators of irregular behaviour are rare, as often tracking devices malfunction or tracking data is not available, thus most researchers usually rely on synthetic data for validation purposes [40].

The European Marine Casualty Information Platform (EMCIP) is a database and a data distribution system operated by European Maritime Safety Agency. Data includes information regarding incidents that took place in the EU waters; containing information regarding the location of the incident, the type, nature, ships involved, time, severity, casualties, and weather conditions. We rely on this data, to test our method's accuracy in detecting specific types of real accidents which took place in the last 7 years. Although this database contains data on a variety of incidents, we focus on severe incidents which would cause a vessel to alter its expected voyage, such as ship collisions, groundings, and loss of control due to engine failure. In sum, from a total of 335 maritime incidents reported, ROTA was capable of detecting 313; achieving an accuracy of 93,4%. In Table 5 below, further statistics are presented.

In the following sections of our work, we look into five serious maritime incidents as examples which include real AIS data as broadcast from the vessels "Norman Atlantic", "KEIT", "INDRA II", "YM Pluto" and the "Costa Concordia" prior and soon after related incidents.

A. INCIDENT 1: THE "COSTA CONCORDIA" GROUNDING (13th JANUARY 2012)

On the night of January the 13th 2012, "Costa Concordia", with 3206 passengers and 1023 crew members on board, was sailing off Isola del Giglio, during its planned seven-day cruise from Civitavecchia to Savona and five other ports. After deviating from course the ship struck its port side on a reef, known as the "Scole Rocks", about 800 metres south of the entrance to the harbour of port of Giglio, on the island's east coast [41]. Soon after running aground, water flushed in causing the vessel to tilt, while the engine rooms flooded and propulsion was lost. Thirty two lives were lost, while the "Costa Concordia" was officially declared a "constructive total loss" by the insurance company, and its salvage was "one of the biggest maritime salvage operations" [42].

(a) Overview (b) Zoom

FIGURE 5. AIS track history for the "Costa Concordia" during the time period of the incident while travelling from Civitavecchia to Savona. The green polygons depict the commonly travelled route between these ports, while the red star marks the position of the incident.

In Fig. 5 the normal vessel behaviour on route to Savona is depicted with green polygons (''convex hulls''), while the positions of the ''Costa Concordia'' are depicted with red markers and the incident location with a red star. It is evident that the ''Costa Concordia'' had clearly deviated from normal behavioural patterns as defined by ROTA; thus being classified as an anomalous incident.

B. INCIDENT 2: THE YM PLUTO INCIDENT (27th APRIL 2013)

YM Pluto departed from Ceuta, Spain on the 25th of April 2013 bound for Rotterdam. On the 27th of April 2013 the vessel was North of the Western coast of Portugal while the weather forecast reported northerly winds of Force 7 to 8 on the Beaufort scale with severe gusts and very rough seas. During the early morning of 27th April 2013, the master of YM Pluto was on the forecastle deck in adverse weather conditions, attempting to stop a water leakage. Unexpectedly, the ship slammed into a very large wave. The master was exposed to the violent impact of the breaking wave and was severely injured. While arrangements were made to dispatch a helicopter to airlift the master, the vessel altered its course and headed towards the port of Averio.

In Fig. 6, the depicted normality model defined by ROTA is presented with green polygons, while the deviation from the normal maritime traffic patterns by ''YM Pluto'' is clear.

FIGURE 6. AIS track history for the "YM Pluto" during the time period of the incident while travelling from Ceuta to Rotterdam. The green polygons depict the commonly travelled route between these ports, while the red star marks the position of the incident.

C. INCIDENT 3: THE NORMAN ATLANTIC FIRE AT SEA (28th DECEMBER 2014)

On 28th December 2014 the Norman Atlantic caught fire in the Strait of Otranto, while travelling on route from Patra (Greece) to Ancona (Italy). As a result of the fire, 11 persons died, 12 went missing and 31 were injured. Similarly to the previous incidents, the green polygons (''convex hulls'') in Fig. 7 depict the extracted maritime patterns for this route. It is clear that after the occurrence of the incident the vessels drifts out of the normal travelling patterns (red markers), providing indicators of irregular behaviour.

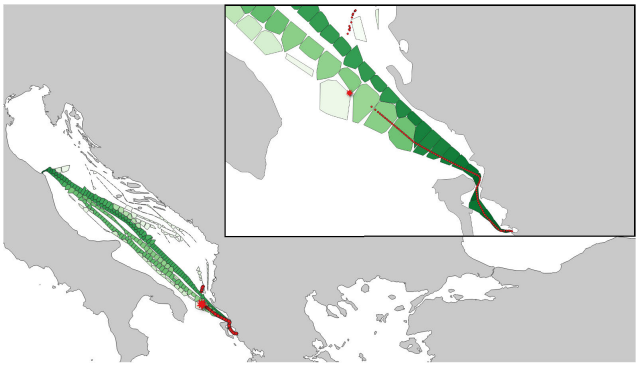

FIGURE 7. AIS track history for the "Norman Atlantic" during the time period of the incident while travelling from Patra to Ancona. The green polygons depict the commonly travelled route between these ports, while the red star marks the position of the incident.

D. INCIDENT 4: THE CARGO VESSEL KEIT AUTO PILOT FAILURE (27th DECEMBER 2017)

The cargo vessel KEIT lost control after an auto pilot failure while transiting the Kiel Canal on Dec 27th 2017, on route from Rotterdam to Klaipeda Lithuania [43]. The ship berthed

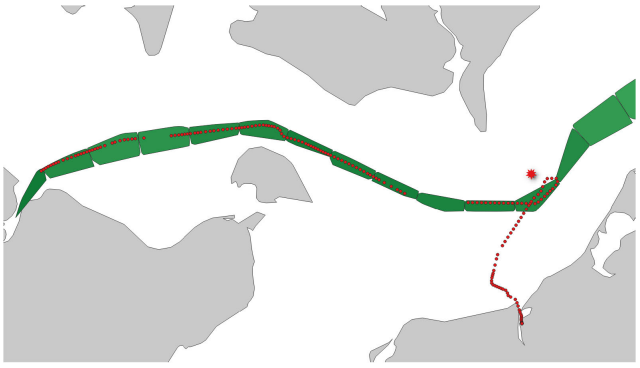

FIGURE 8. AIS track history for the "KEIT" during the time period of the incident while travelling from Kiel to Sodertajle. The green polygons depict the commonly travelled route between these ports, while the red star marks the position of the incident.

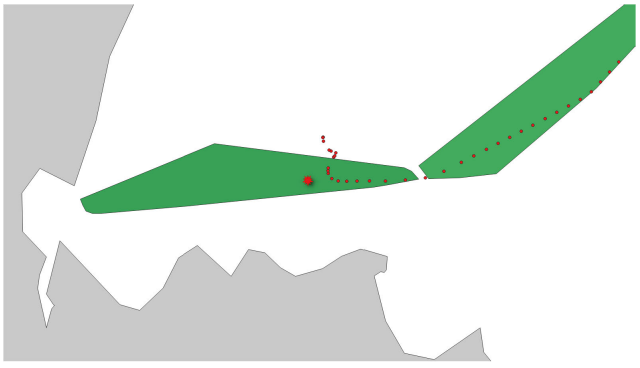

FIGURE 9. AIS track history for "INDRA II" during the time period of the incident while travelling from Odessa to Burgas. The green polygons depict the commonly travelled route between these ports.

near the accident site. Fig. 8 presents the normal route in green and the position of the incident with a red start, while vessel positional data as collected through the AIS are highlighted with red markers. It is evident that after the incident the KEIL positions fall outside of the "normal" behavioural patterns.

E. INCIDENT 5: INDRA II (9th NOVEMBER 2015)

At 21:05 on 9 November 2015, while the bulk carrier INDRA II was at about 6 nm off port of Burgas, the main engine was brought to an emergency stop, due to an "automatic protection of ME from low oil pressure" alarm appearing [42]. The ship went out of the Traffic Separation Scheme and drifted on the Burgas roadstead. The crew commenced cleaning of filters of Oil Lubricating System. At 23:10 an engineer started opening the main oil filter, after the oil pump had been stopped. When the cover was released, the filter unexpectedly blew and hit the engineer in the right part of his chest. Emergency First Medical Aid was given and Emergency Shore Medical team was called, but all attempts to revive the engineer were unsuccessful. In Fig. 9, the irregular behaviour of the "INDRA II" is evident when compared with the normality model calculated by ROTA.

IV. DISCUSSION AND LIMITATIONS

The experimental results show that applying distributed computation approaches can achieve high efficiency, thus reducing processing time (from days as reported in [11], [19]) to less than 3 hours. In sum, the accuracy of our approach is above 73% for most types of vessels that frequently travel on "sea roads", with an overall of 78%. Some interesting insights of our work include, i) approximately just over 10 percent of global trips have a moderate repetition number (10 trips in two years are insignificant), while ii) the accuracy of our method is highly correlated with number of trips we have in our training set for each route. As it would be expected, certain types of vessels such as supporting vessels (e.g., tugs, etc.) are not accurately identified, as they do not follow spatially and temporally repetitive patterns (e.g., do not operate in the same geographical area at repeating times), although, they do have many common characteristics between their routes (i.e., similar trajectory pattern).

ROTA can be considered as a fundamental step towards performing anomaly detection on historical and real time data. We evaluate ROTA's ability to detect true "anomalies", such as maritime incidents reported in the European Marine Casualty Information Platform. Our results show the advantages of ROTA in terms of accuracy, with an achieved anomaly detection accuracy of higher than 93%, by detecting 313 out of 335 relevant maritime incidents reported in EU waters. Furthermore we present a number of examples demonstrating the applicability and accuracy of the suggested approach.

ROTA performs clustering of historical data (i.e., vessels' positions) to create "sea roads". A fundamental prerequisite for the accurate representation of sea roads is the existence of sufficient amount of data for each route. In areas with limited network coverage (e.g., open sea at oceans) the data available are scarce and consequently, the clusters created are sparse, covering large areas and limiting the precision of the produced sea roads in those areas. Finally, we avoid using other data-driven methods that repeat the clustering process for various k-s (e.g., elbow method, gap statistic method, etc.), since the clustering method is needed for each route, and thus, the computational time would be increased dramatically.

V. CONCLUSION

In this paper we presented a novel methodology that uses historical AIS positional data and port geometries to extract maritime "patterns of life" at a global scale. The methodology has been applied on a per-port-connection basis, to real-world data covering the entire globe for an extended period of 2 years (January 2016 until December 2017). ROTA has been proven capable of extracting traffic patterns at a global scale from non-uniform spatial and temporal data distributions (such as AIS), without requiring manual tuning or a priori knowledge achieving accuracy higher than 73% for the produced "sea roads". Furthermore, ROTA would be of great

value for maritime security applications, as it has been able to detect security incidents with accuracy higher than 93%.

Our goal is to extend this work and calculate additional characteristics for the extracted ''routes'', on a ''convex-hull'' level, which could provide even more early indicators of irregular behaviour. Our results could become even more relevant in the years ahead with the advent of autonomous vessels. For future work, we also intend to examine trajectory similarity algorithms to capture mobility patterns of vessel types that do not conform to the presented ''Origin to Destination'' connections such as supply vessels and tug boats that generally perform port operations inside or near port. Finally, we will extend the clustering algorithm by testing alternative (to the empirical method) data-driven methods and optimise the number of clusters used in the k-means algorithm.

REFERENCES

[1] *Review of Maritime Transport 2014*, UNCTAD, Geneva, Switzerland, 2014.

[2] C. Ducruet, *Advances Shipping Data Anal. Modeling: Tracking Mapping Maritime Flows Age Big Data*. Evanston, IL, USA: Routledge, 2018.

[3] *M.1371: Technical Characteristics for an Automatic Identification System Using Time-Division Multiple Access in the VHF Maritime Mobile Band*. Accessed: Sep. 24, 2018. [Online]. Available: https://www.itu.int/rec/R-REC-M.1371/en

[4] *Solas Chapter V—Regulation 19—Carriage Requirements for Shipborne Navigational Systems and Equipment*. Accessed: Sep. 24, 2018. [Online]. Available: http://solasv.mcga.gov.uk/regulations/regulation19.htm

[5] E. Tu, G. Zhang, L. Rachmawati, E. Rajabally, and G.-B. Huang, ''Exploiting AIS data for intelligent maritime navigation: A comprehensive survey from data to methodology,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1559–1582, May 2018.

[6] J. Poļevskis, M. Krastiņš, G. Korāts, A. Skorodumovs, and J. Trokšs, ''Methods for processing and interpretation of AIS signals corrupted by noise and packet collisions,'' *Latvian J. Phys. Tech. Sci.*, vol. 49, no. 3, pp. 25–31, Jan. 2012. [Online]. Available: https://content.sciendo.com/view/journals/lpts/49/3/article-p25.xml

[7] M. Yang, Y. Zou, and L. Fang, ''Collision and detection performance with three overlap signal collisions in space-based AIS reception,'' in *Proc. IEEE 11th Int. Conf. Trust, Secur. Privacy Comput. Commun.*, Jun. 2012, pp. 1641–1648.

[8] D. Nguyen, R. Vadaine, G. Hajduch, R. Garello, and R. Fablet, ''A multi-task deep learning architecture for maritime surveillance using AIS data streams,'' in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2018, pp. 331–340.

[9] L. Goldsworthy and B. Goldsworthy, ''Modelling of ship engine exhaust emissions in ports and extensive coastal waters based on terrestrial AIS data–an australian case study,'' *Environ. Model. Softw.*, vol. 63, pp. 45–60, Jan. 2015.

[10] M.-C. Tsou, ''Discovering knowledge from AIS database for application in VTS,'' *J. Navigat.*, vol. 63, no. 3, pp. 449–469, Jul. 2010.

[11] N. Willems, H. van de Wetering, and J. J. van Wijk, ''Visualization of vessel movements,'' *Comput. Graph. Forum*, vol. 28, no. 3, pp. 959–966, Jun. 2009.

[12] L. Cazzanti, A. Davoli, and L. M. Millefiori, ''Automated port traffic statistics: From raw data to visualisation,'' in *Proc. IEEE Int. Conf. Big Data*, Dec. 2016, pp. 1569–1573.

[13] L. Zhang, Q. Meng, and T. Fang Fwa, ''Big AIS data based spatial-temporal analyses of ship traffic in singapore port waters,'' *Transp. Res. E, Logistics Transp. Rev.*, vol. 129, pp. 287–304, Sep. 2017.

[14] Q. Meng, J. Weng, and S. Li, ''Analysis with automatic identification system data of vessel traffic characteristics in the singapore strait,'' *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2426, no. 1, pp. 33–43, Jan. 2014.

[15] A. Alessandrini, F. Mazzarella, and M. Vespe, ''Estimated time of arrival using historical vessel tracking data,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 7–15, Jan. 2019.

[16] N. Bomberger, B. Rhodes, M. Seibert, and A. Waxman, ''Associative learning of vessel motion patterns for maritime situation awareness,'' in *Proc. 9th Int. Conf. Inf. Fusion*, Jul. 2006, pp. 1–8.

[17] R. Laxhammar, ''Anomaly detection for sea surveillance,'' in *Proc. 11th Int. Conf. Inf. Fusion*, Jun. 2008, pp. 1–8.

[18] B. Ristic, B. F. L. Scala, M. R. Morelande, and N. J. Gordon, ''Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction,'' in *Proc. 11th Int. Conf. Inf. Fusion*, Jun. 2008, pp. 1–7.

[19] L. Wu, Y. Xu, Q. Wang, F. Wang, and Z. Xu, ''Mapping global shipping density from AIS data,'' *J. Navigat.*, vol. 70, no. 1, pp. 67–81, Jan. 2017.

[20] G. Pallotta, M. Vespe, and K. Bryan, ''Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction,'' *Entropy*, vol. 15, no. 12, pp. 2218–2245, 2013.

[21] G. Wang, J. Meng, and Y. Han, ''Extraction of maritime road networks from large-scale AIS data,'' *IEEE Access*, vol. 7, pp. 123035–123048, 2019.

[22] D. O. D. Handayani, W. Sediono, and A. Shah, ''Anomaly detection in vessel tracking using support vector machines (SVMs),'' in *Proc. Int. Conf. Adv. Comput. Sci. Appl. Technol.*, Dec. 2013, pp. 213–217.

[23] W. Li, C. Zhang, J. Ma, and C. Jia, ''Long-term vessel motion predication by modeling trajectory patterns with AIS data,'' in *Proc. 5th Int. Conf. Transp. Inf. Saf. (ICTIS)*, Jul. 2019, pp. 1389–1394.

[24] Y. Li, R. W. Liu, J. Liu, Y. Huang, B. Hu, and K. Wang, ''Trajectory compression-guided visualization of spatio-temporal AIS vessel density,'' in *Proc. 8th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2016, pp. 1–5.

[25] M. Fiorini, A. Capata, and D. D. Bloisi, ''AIS data visualization for maritime spatial planning (MSP),'' *Int. J. E-Navigat. Maritime Economy*, vol. 5, pp. 45–60, Dec. 2016.

[26] F. Mazzarella, V. F. Arguedas, and M. Vespe, ''Knowledge-based vessel position prediction using historical AIS data,'' in *Proc. Sensor Data Fusion. Trends, Solutions, Appl. (SDF)*, Oct. 2015, pp. 1–6.

[27] S.-K. Zhang, G.-Y. Shi, Z.-J. Liu, Z.-W. Zhao, and Z.-L. Wu, ''Data-driven based automatic maritime routing from massive AIS trajectories in the face of disparity,'' *Ocean Eng.*, vol. 155, pp. 240–250, May 2018.

[28] P. Coscia, P. Braca, L. M. Millefiori, F. A. N. Palmieri, and P. Willett, ''Multiple Ornstein–Uhlenbeck processes for maritime traffic graph representation,'' *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 5, pp. 2158–2170, Oct. 2018.

[29] L. M. Millefiori, P. Braca, K. Bryan, and P. Willett, ''Modeling vessel kinematics using a stochastic mean-reverting process for long-term prediction,'' *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 5, pp. 2313–2330, Oct. 2016.

[30] S. A. Breithaupt, A. Copping, J. Tagestad, and J. Whiting, ''Maritime route delineation using AIS data from the atlantic coast of the US,'' *J. Navigat.*, vol. 70, no. 2, pp. 379–394, Mar. 2017.

[31] *Methodology for Assessing the Marine Navigational Safety & Emergency Response Risks of Offshore Renewable Energy Installations*, Dept. Trade Ind., New Delhi, India, 2013.

[32] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996.

[33] M. R. MacLeod and W. M. Wardrop, ''Operational analysis at combined maritime forces,'' in *Proc. 32nd Int. Symp. Mil. Oper. Res.*, 2015, pp. 1–14.

[34] *World Port Index*. Accessed: Sep. 24, 2018. [Online]. Available: https://msi.nga.mil/NGAPortal/MSI.portal?_nfpb=true&_pageLabel=msi_portal_page_6&pubCode=0015

[35] *Accident Investigation—EMSA—European Maritime Safety Agency*. Accessed: Oct. 12, 2018. [Online]. Available: http://www.emsa.europa.eu/implementation-tasks/accident-investigation.html

[36] P. Last, C. Bahlke, M. Hering-Bertram, and L. Linsen, ''Comprehensive analysis of automatic identification system (AIS) data in regard to vessel movement prediction,'' *J. Navigat.*, vol. 67, no. 5, pp. 791–809, Sep. 2014.

[37] J. Dean and S. Ghemawat, ''Mapreduce: Simplified data processing on large clusters,'' in *Proc. OSDI*, 2004, pp. 137–150.

[38] S. Lloyd, ''Least squares quantization in PCM,'' *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[39] T. F. Gonzalez, ''Clustering to minimize the maximum intercluster distance,'' *Theor. Comput. Sci.*, vol. 38, pp. 293–306, Jan. 1985.

[40] E. Osekowska, H. Johnson, and B. Carlsson, ''Grid size optimization for potential field based maritime anomaly detection,'' *Transp. Res. Procedia*, vol. 3, pp. 720–729, Jan. 2014.

[41] *Maritime History: Costa Concordia Disaster*. Accessed: Jan. 13, 2018. [Online]. Available: https://safety4sea.com/maritime-history-costa-concordia-disaster/
[42] *Published Maritime Casualty Investigation Reports*, Eur. Maritime Casualty Inf. Platform—Eur. Maritime Saf. Agency, Lisbon, Portugal, 2013.
[43] *Near Miss Accident in Kiel Canal, Caused by Navigational Rules Violation*. Accessed: Oct. 12, 2018. [Online]. Available: https://www.fleetmon.com/maritime-news/2017/20951/near-miss-kiel-canal-caused-navigational-rules-vio

DIMITRIS ZISSIS (Senior Member, IEEE) is currently an Associate Professor with the University of the Aegean. His published scientific work includes more than 70 publications, which have received more than 2200 citations to date. His research interests and areas of expertise include several aspects of architecting and developing complex Information Systems (IS), including distributed and cloud-based big data deployments. He is a member of the editorial boards of *Future Generation Computer Systems* (FGCS) (Elsevier) and the *International Journal of Internet of Things and Cyber-Assurance* (Inderscience) and a PC Member for numerous conferences, including CLOUDCOM, GECON, SerCO, and others. He is also a member of the IEEE Computer Society, the IEEE Oceanic Engineering, the IEEE Intelligent Transportation Systems Societies, and the Young Researchers Committee of the World Federation on Soft Computing. His professional experience includes senior consulting and researcher positions in a number of private and public institutions.

KONSTANTINOS CHATZIKOKOLAKIS received the B.Sc., M.Sc., and Ph.D. degrees from the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, in 2008, 2012, and 2016, respectively. He has participated in several European projects, including METIS 2020, BigDataOcean, and SELIS. His research interests include data mining, machine learning, big data, and distributed and cloud computing. His professional experience includes software engineer and senior researcher positions both in private and public sector.

GIANNIS SPILIOPOULOS received the Diploma and M.Sc. degrees from the Electronics and Computer Engineering Department, Technical University of Crete, in 2008 and 2010, respectively. He has been a member of Technical Chamber of Greece, since 2009. He has worked as a Software Engineer and an Analyst in the private sector, for six years. His research interests include statistical analysis of spatial data and data-driven extraction of spatiotemporal patterns using unsupervised machine learning methods.

MARIOS VODAS (Member, IEEE) received the B.Sc. and M.Sc. degrees in informatics from the University of Piraeus, Greece, in 2011 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Department of Informatics, University of Piraeus. He is also a Researcher wih MarineTraffic. His research interests include spatio-temporal indexing and mining on big data.