18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

Data-driven Detection and Context-based Classification of Maritime Anomalies

Giuliana Pallotta and Anne-Laure Jousselme
NATO-STO Centre for Maritime Research and Experimentation (CMRE), La Spezia, Italy
Emails:{Giuliana.Pallotta, Anne-Laure.Jousselme}@cmre.nato.int

***Abstract*—Discovering anomalies at sea is one of the critical tasks of Maritime Situational Awareness (MSA) activities and an important enabler for maritime security operations. This paper proposes a data-driven approach to anomaly detection, highlighting challenges specific to the maritime domain. This work builds on unsupervised learning techniques which provide models for normal traffic behaviour. A methodology to associate tracks to the derived traffic model is then presented. This is done by the pre-extraction of contextual information as the baseline patterns of life (*i.e.*, routes) in the area under investigation. In addition to a brief description of the approach to derive the routes, their characterization and representation is presented in support of exploitable knowledge to classify anomalies. A hierarchical reasoning is proposed where new tracks are first associated to existing routes based on their positional information only and "off-route" vessels" are detected. Then, for on-route vessels further anomalies are detected such as "speed anomaly" or "heading anomaly". The algorithm is illustrated and assessed on a real-world dataset supplemented with synthetic abnormal tracks.**

I. Introduction

The maritime domain is the most utilized environment for transportation, making maritime safety and security an important concern. Millions of vessels sail every day moving passengers, containers, and consumer goods. Some vessels are engaged in military operations, others carry out specific activities such as fishing. Some may be involved in illicit, illegal or dangerous activities. As discussed in [1], anomaly detection algorithms have many potential applications (e.g., countering piracy and organized crime, prevent terrorism). A key aspect of maritime safety and security is MSA, which is supported by surveillance and tracking systems. Sensors networks are ever increasing in order to complement and refine the picture of vessel motions at sea. More specifically, the Automatic Identification System (AIS) (Automated Identification System) technology is a cooperative vessel self-reporting system able to provide a vast amount of near real time information about vessel static and kinematic features [2]. These traffic spatio-temporal data streams can be collected from both coastal and satellite AIS receivers, requiring an ever increasing degree of automation and efficiency in order to detect and characterize ambiguities, redundancies, inconsistencies and ultimately transform data in actionable knowledge.

Previous work for the automated learning of vessel traffic routes from AIS data has shown that valuable knowledge about the behaviour of maritime traffic can be extracted through the analysis of historical data. The present paper starts from and expands the work on route classification and anomaly detection discussed in [3], [4]. starting from these papers, we here provide an example of exploitation of this contextual information for anomaly detection in case of short/incomplete tracks. A heuristic track to route association method is presented, whose parameters and decision thresholds are estimated from historical AIS data, previously mined into normal patterns of life categorized into routes and stationary areas. The parameters are essential characteristics of recurrent routes (e.g., physical extent, speed profiles, vessel type). It is demonstrated that the use of routes can provide a short-list of anomalous tracks (either incomplete or entire trajectories) via a hierarchical approach which incrementally triggers different *anomaly detectors* on specific features.

The advantage of the technique presented here is the relative ease of the modelling and parameterization of individual traffic routes which is key to the potential operationalization of track classification on real time and on a larger scale aiming at monitoring numerous different targets simultaneously.

The remainder of the paper is organized as follows. Section II reviews related and previous work in the field of maritime traffic analysis and classification while Section III gives an overview of the proposed approach. The context extraction and route representation is reviewed in Section IV and the route classifier and anomaly detector are presented in Sections V and VI. Section VII applies the methodology to a real world data set and discusses results. Conclusions are reported in Section VIII together with directions for future works.

II. Related work and state-of-the-art

The typical approach to both vessel behaviour analysis and anomaly detection involves the extraction of a normalcy model from a set of features in vessel track data [5], [6]. The general aim of this approach is to cluster the data in a multi-dimensional feature space, where the features of the track are attributes such as longitude and latitude, speed and course.

Generally, the process follows the three steps [7]:

- *Labelling*: vessel tracks are clustered into a number of routes, by compressing the input data into a compact object structure.
- *Prediction*: the information can be used to predict forward vessel tracks, based on the current location and direction.
- *Atypicality*: accumulating tracks over a long time period establishes a pattern of typical movements (e.g., patterns of life) and this can support the recognition of atypical or unusual (e.g., anomalous) movements.

The key assumption is: *tracks that are found to sit in or close to one of these clusters may be considered normal tracks, while those that sit at a larger distance from all the clusters may indicate an anomaly.*

As discussed in [8], when following this approach two key decisions need to be made:

- The learning technique which will lead to the model based on the training data;
- A suitable model representation to translate the learned knowledge. Models of normal vessel behaviour can be represented in many different forms, with varying degrees of complexity, including Support Vector Machines, Gaussian Mixture Models, Kernel Density Estimators, neural networks and Bayesian Networks (see, e.g., [8]).

The problem of track classification in the maritime domain is strictly linked to the anomaly detection. The anomaly detection task can be actually performed in two different ways:

- estimating the degree of deviation of the test trajectories from the learned normal trajectories;
- modelling directly the anomalies and develop specific detectors to automatically recognize the corresponding anomalous behavior.

The usability of the first approach seems easier to be applied on a larger scale to gain efficient classification performances and detect different type of anomalies. Most of the available literature can be reconnected to this first approach, especially in the maritime domain. However recently some examples of direct definition and recognition methods of specific behaviours at sea have been proposed such as:

- anomalous time sequences (as in the case of vessel reports received from different asynchronous sensors).
- specific trajectory evolution (e.g., fishing footprints [9], loitering vessels [10], U-turns or multiple loops such as to avoid an obstacle).

These latter approaches look promising for real time applications since they do not explicitly rely on historical information, although the behavioural modelling is generally helped by the analysis of the anomalous behaviours seen in the past.

Maritime anomaly detection methods using the historical patterns of life as the reference can be distinguished into two main classes, based on the format of input trajectories:

- *point-based* anomaly detection methods, which highlight individual anomalous points (either AIS reports or single anomalous points along a trajectory);
- *trajectory-based* anomaly detection methods, which receive as input entire trajectories and automatically classify them as anomalous or not.

Generally these anomaly detectors are informed by the way the normalcy is constructed, being the former ones deriving from point-based clustering methods (as K-means and DBSCAN), the latter ones related to the clustering of entire trajectories based on similarity measures, which include the shape of the trajectory into the classification problem.

Off-line vs on-line anomaly detection

Most of the algorithms available in literature are designed for off-line anomaly detection in trajectories. Hence, they are based on the assumption that the anomaly classification is performed after the entire trajectory has been observed.

This is an evident drawback for surveillance applications since it delays the alert on anomalous events, thus delaying the necessary actions to be taken. In contrast, algorithms for sequential anomaly detection allow detection in incomplete or partially observed trajectories (e.g., tracks, tracklets or trajectories) [11], e.g., real-time detection of anomalous trajectories as they evolve.

III. Methodology overview and proposed approach

Previous work for the automated learning of traffic routes from AIS data [12] [13] has shown that through the analysis of historical data, valuable knowledge about the behaviour of maritime traffic can be extracted. Specifically, some previous work has discussed the development of the Centre for Maritime Research and Experimentation (CMRE) tool called Traffic Route Extraction for Anomaly Detection (TREAD) [3], [4] which provides an unsupervised learning approach to derive a dictionary of the maritime traffic routes using spatio-temporal data streams from terrestrial and/or satellite AIS receivers. This dictionary will be the contextual information which we exploit in the present work to perform track classification.

The proposed scheme addresses the following Maritime Situational Indicators of interest for operators:

- "The vessel is off-route";
- "The vessel is in reverse traffic on the route";
- "The speed is not compatible with the route followed".

The flowchart of the proposed methodology is summarized in Fig. 1. As part of contextual information, the picture of the maritime traffic is first derived from this historical AIS data of the area of interest. The underlying assumption is that the feature values of the AIS data points come from a stationary distribution of normal traffic, estimated using training data. The route objects are extracted as labelled motion prototypes which can then be compared with newly observed tracks of vessels at sea in the same area. The new tracks are first associated to the derived routes based on the positional information only and labelled with the route label. "Off-route vessels are detected at this step as vessels not following an existing route. Then, further compatibility tests between the track and its associated route are performed on the kinematic information such as speed and heading. Anomalies about speed such as "speed incompatible with the route (too high or too slow), or "vessel in opposite direction are then possibly identified.

The originality of this work is the concrete association of a track to a route to be used as a basis for further reasoning. Most of the works of the literature [5], [14], [12] directly detect anomalies skipping this step of route association. The operational advantage is twofold:

- Once *assigned* to a specific route[1] , future vessel position and speed can be predicted, missing information type of vessel can be estimated, etc. Also, additional predictions

[1]In the presented approach, a different association method is used in the route extraction process (clustering) compared to the classification phase.

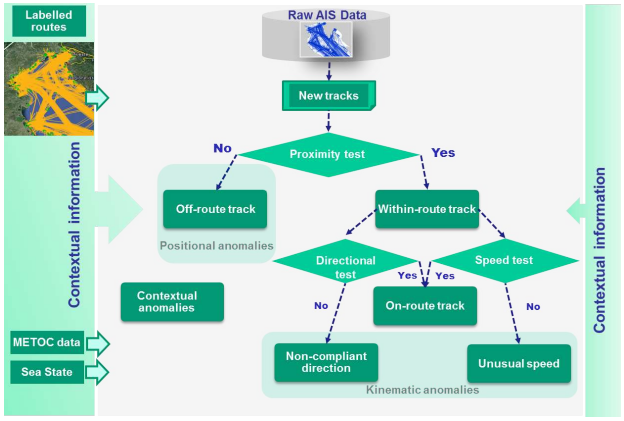

Fig. 1. The general flowchart of the Track-to-Route classification.

and compatibility tests can be performed. For instance, the vessel destination can be predicted or validated with the vessel declared Next Port of Call (NPOC) (Next Port Of Call) in AIS messages. To this aim, the use of the learned routes can provide more accurate predictions on a longer time scale, compared to conventional models (*e.g.* particle filtering). As an example, the beneficial inclusion of route average speed as prior contextual information into vessel prediction models has been discussed in [15].

- If the vessel is *not assigned* to any existing route, it enhances the detection of potentially anomalous behaviours, commonly defined as behaviours *deviating from* or *not compliant* to the learned traffic normality, and further actions at different levels (*e.g.*, sensor managements, asset allocation) can be decided.

A distinctive feature of the methodology is the possibility to classify tracks which are *incomplete* (e.g., they have gaps in the data sequences or their reports are incomplete in some of their fields as in the case of Course over Ground (COG) and/or Speed over Ground (SOG) is missing or inconsistent). Compared to other trajectory-based classification methods where input trajectories need to be uniformly sampled over time (either directly or after re-sampling techniques), our approach takes as input a partial track from that vessel with a variable time rate. Moreover, being the proposed approach hierarchical in the feature domain, the classification does not assume complete trajectories, since it is based on an incremental handling of the track, updating the classification as soon as a new contact from the same vessel is received. Consequently, the longer the track duration and/or the larger the number of feature measurements available, the better the track classification performance is expected. Also, the explicit association of a vessel (or track) to an existing route is original to this work.

IV. Context Extraction via TREAD

As discussed in Section II, the way in which the classification is performed is very much influenced by the representation of the patterns of life extracted in the labelling phase. This means that, depending on the specific final aim of the classification, different levels of complexity can be reached in the way normal patterns are represented.

A. Routes construction by clustering

The increasing quantity of historic AIS reports poses new challenges in the related fields of data mining and machine learning techniques when applied within the context of big data and MSA. The large amount of vessel movement data collected by terrestrial networks and satellite constellations of AIS receivers requires the aid of automatic processing techniques to fully exploit this data, since the initial amount of raw information can overwhelm human operators. A comprehensive knowledge of recurrent vessel patterns in an area under investigation is valuable contextual information to support accurate vessel prediction and track classification. The motion patterns are here learned using TREAD tool. More details about the learning methodology behind TREAD can be found in [3], [4], where the way the contextual information is derived is discussed with examples. We here exploit this contextual information to perform track association for anomaly detection. In the learning-labelling phase TREAD uses an incremental Density-Based Spatial Clustering of Applications with Noise (DBSCAN) procedure (see [16]) adapted to the maritime domain by a spatio-temporal clustering which detects the following events in vessel motion AIS data streams:

- "the vessel enters the area"
- "the vessel exits the area"
- "the vessel is steady"
- "the vessel is sailing"

These discrete events are then incrementally clustered to create, update and merge the labelled clusters, being of three different types: "route R_k", "entry/exit gate","stationary area".

The clustering parameters are dependent on the specific scale and traffic intensity/density of the selected area and are affected by other contextual factors such as weather conditions, seasonality, etc. Traffic knowledge for human operators, providing an up-to-date situation assessment information (e.g., level 2 processing in the Joint Directors of Laboratories (JDL) model [17]). The way routes and traffic change with time and season can help operators in enhancing the knowledge of vessel pattern of life activity and analysts in predicting vessel movement and the impact on traffic when the routing systems are modified. Information from the derived routes, such as the vessel type, mean velocity or the series of route points provided by previous transits, represent a set of contextual constraints that can be used to characterize the behaviour of specific classes of vessels along the route. The functional architecture of the proposed methodology is summarized in Fig. 2. In Section V, we will illustrate how these contextual feeds can be used into the proposed context-based classification model, following a hierarchical *data-driven* approach.

B. Representation of maritime routes

The design of an anomaly detector is strictly linked to the representation of the data in which we search for anomalies (see, e.g., [6]). Some anomaly detection methods assume that

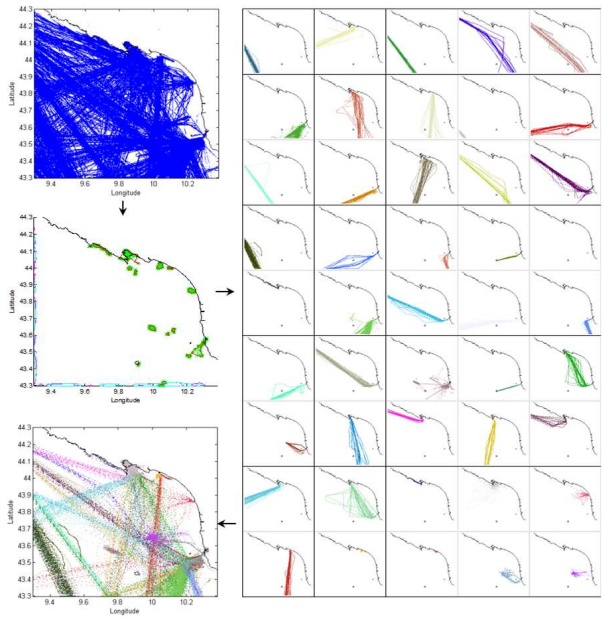

Fig. 2. The general overview of TREAD Knowledge Discovery Process: the raw AIS data stream is processed. Vessel movements are clustered. The discovered traffic route system is organized as a dictionary of motion models.

data, in our case trajectories, are represented as points in a fixed-dimensional feature space. This implies that a fixed number of features, e.g., the position at a fixed number of points in time, has to be extracted from each trajectory or trajectory segment. Other methods only assume that a similarity or dissimilarity measure is defined for pairs of trajectories or trajectory segments. The choice of features or similarity or dissimilarity measure essentially determines the type of anomalies that can be detected and is therefore very important. We propose a spatial model for representing routes on the sea, similarly to [7].

Let us denote by $\mathcal{R} = \{R_1, \ldots, R_K\}$ the finite set of extracted routes for a given region. As described above,a route R_k is built from a cluster of vessel AIS reports following the same itinerary. The concept of itinerary can be defined as an ordered sequence of positions (i.e.,, directed path). A route describes the entire trajectory of an object from the time it enters one waypoint to when it exits another waypoint and can be described as a curve with specific start and end points.

Each route is associated with a *route prototype* representing an "average route" or be interpreted as a *representative trajectory*. A route prototype is represented by a set of intermediate estimated waypoints that form a poly-line bounded by two extrema on the variation in trajectories sampled for the route (see Figure 3). Vessels move in an open area, some of them can move away from the "average route" as noted in [18], but these deviations can still be normal when they fall within a certain interval. For each route we define a *route spatial extent* (e.g., area of influence) which can be probabilistically represented using a Kernel Density Estimation (KDE) method of the random variable "vessel position" adopting a Gaussian kernel

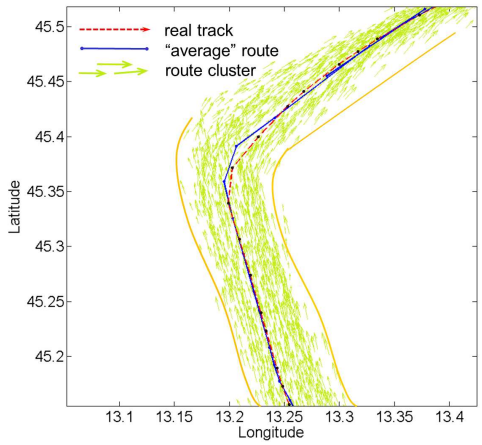

Fig. 3. A route cluster with the "average route" and boundaries superimposed.

with an optimized bandwidth. KDE has shown a superior ability to accurately model traffic routes, even in the case of skewed distribution of vessel positions across the main route longitudinal axis as shown in [3].

More specifically, the route R_k can be represented by:

- an *origin* point O_k, which is the centroid of the entry gate;
- a *destination* point D_k, which is the centroid of the exit gate; this is a relevant information mostly when the exit gate area corresponds to a port;
- an *"average route"* $\bar{R}_k$, e.g., a sequence of waypoints WP_j represented by a 2-D position vector. A median COG and SOG is associated to each waypoint; each waypoint can be expressed as follows:

$$\mathbf{WP}_j = [x_{WP_j}, y_{WP_j}, \dot{x}_{WP_j}, \dot{y}_{WP_j}] \tag{1}$$

  where x_{WP_j} and y_{WP_j}, are related to the coordinates of an ideal vessel moving along the "average route" and passing by that specific WP_j; the velocity components, $\dot{x}_{WP_j}$ and $\dot{y}_{WP_j}$ are derived by combining median SOG, $S\bar{O}G_{WP_j}$ and COG, $C\bar{O}G_{WP_j}$ information, based on the conditions: $S\bar{O}G^*_j = \sqrt{(\dot{x}_{WP_j})^2 + (\dot{y}_{WP_j})^2}$ and $C\bar{O}G_{WP_j} = \tan^{-1}\left(\frac{\dot{y}_{WP_j}}{\dot{x}_{WP_j}}\right)$.
- a route *spatial extent* which is computed using the aforementioned KDE;
- a route *width* w_k, which is the maximum of the distances of each route cluster point (vessel positions associated to the route in the learning phase) to the *closest* waypoint on the "average route";
- a route *global kinematic variability* measure expressed in terms of the overall standard deviations σ_{COG_k} and σ_{SOG_k} of the COG and SOG values of the vessels associated to the route;
- a route *"life span"* which is computed as the average duration of the transits of the vessels associated to the route;
- a *ship-type frequency* distribution of the vessels which transited along that route.

It is advantageous to collect data for a long time period to include variations in motion patterns over the day, the week and the year in order to account for the temporal variability of routes, depending on the specific time scale of the analysis. The set of routes $\mathcal{R} = \{R_k\}$ with $k = 1,..K$ will be used hereafter as the *data-driven* contextual information for classification.

V. Track-to-Route (T2R) association

We consider a Vessel of Interest (VOI) under observation about which we receive a partial track with a variable time rate. A vessel track, $\mathbf{V}$, is a temporal sequence of T observed state vectors, $\mathbf{v}_t$:

$$\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T\} \tag{2}$$

where we assume that the state vector observation, $\mathbf{v}_t$, is extracted from AIS information. However, it could be provided by any other source such as coastal radar with associated tracker or a combination of other sources. In this study, it includes both position and velocity information as extracted by the vessel track properties.

$$\mathbf{v}_t = [x_t, y_t, \dot{x}_t, \dot{y}_t]^\intercal \tag{3}$$

where x_t and y_t are related to the vessel coordinates and the velocity components, $\dot{x}_t$ and $\dot{y}_t$, are derived by combining SOG and COG information, as defined above. $\mathbf{v}_t$ can be further expanded to include additional features such as the length, the type, etc.

The classification problem of a set of vessel positioning observations (forming a track) with respect to a labelled set of routes can be formulated as building the mapping Ψ from the set of tracks of interest $\mathcal{T}$ to $\mathcal{R}^*$, where $\mathcal{R}^* = \mathcal{R} \cup R^*$ is the extended set containing a rejection label corresponding to off-route vessels:

$$\Psi(\mathbf{v}_t) = A_R \tag{4}$$

where A_R is the set of routes compatible with $\mathbf{v}_t$. In the general case, we consider a multi-label classification since vessels can be compatible with several routes. The way routes are defined and extracted, they may partially overlap and share some physical regions with common directions of motion.

The proposed Track-to-Route (T2R) classification computes the probability of a vessel track matching each route compatible to the vessel features extracted from the partially observed vessel track.

A. Definition of meaningful distances

The first task of the T2R classification algorithm is to infer whether the vessel track, $\mathbf{V}$ deviates significantly from the route model R_k, given the intrinsic variability of the vessels within each route. An effective technique should be capable of distinguish data point deviations that occur due to anomalous events from outliers associated with the tails of the reference route points. Then the distance of the track point from the closest point on the synthetic route can be computed and compared to it for each labelled route R_k.

In the following section some distances are defined, useful for the classification process. Some similar definitions for ground surveillance have been proposed in [7].

Distance of a point from a route waypoint: Let $v_i = (x_i, y_i)$ be a measured vessels position. The distance from v_i to a waypoint is defined as:

$$d(v_i, WP_j) = \|[x_i, y_i] - [x_{WP_j}, y_{WP_j}]\|; \tag{5}$$

In order to take into account the approximated spherical geometry of the Earth's surface we calculate this distance by using the Haversine formula. When the scale of the area of interest is relatively limited, the curvature effect can be neglected and the Euclidean distance formula can be used.

Distance of a track point from a route: The distance (5) can be extended to compute the distance to a route as:

$$\Delta_p^k(v_i) = d(v_i, R_k) = \min_{j=1}^{J} d(v_i, WP_j) \tag{6}$$

for all the $WP_j \in \bar{R}_k$. An approximation of this can be the minimum distance of the track point from all the waypoints of the route. This is the distance of the point v_i from the closest waypoint WP_j on the route $\bar{R}_k$.

Distance of a track from a route: The distance $\Delta_p^k(v_i)$ can be extended to define the distance of the vessel track $\mathbf{V}$ (of length T) from the route R_k:

$$\Delta_p^k = \Delta_p(\mathbf{V}, k) = \max_{i=1}^{T} d(v_i, R_k) \tag{7}$$

It is the maximum distance over all the distances of the track points from the route. Itcan be considered a modified directed Hausdorff distance for moving objects at sea. The Hausdorff distance measures how far two subsets are from each other [6]. It has been used as an objective measure of each trajectory similarity and resulted to be useful for clustering ship trajectories. This distance is not symmetrical, and can be defined as a *directed Hausdorff distance*.

VI. Maritime anomaly detection

The observation sequence is used as a base for the T2R classification. We assume there is no error in the measurement system. The feature distances are defined for each route R_k as follows and included in the observation sequence:

$$\mathbf{O}_\mathbf{V}^k = [\Delta_p^k, \Delta^k heading, \Delta^k speed, \Delta^k Tseq, \Delta^k type] \tag{8}$$

with the following contributions:

- Distance Δ_p^k between the vessel track $\mathbf{V}$ and the route R_k (here indicated as Δ_p, for simplicity of notation);
- Distance in vessel heading ($\Delta^k heading$). This could indicate deviation from normal route vessel flow;
- Distance in vessel speed ($\Delta^k speed$). This could help in identifying vessels showing a high-speed or low-speed compared to the route average speed;
- probability of the temporal sequence $\Delta^k Tseq$ of the track compared to the route which is practically null when the vessel track is not associated to a route;
- Mismatch in vessel type ($\Delta^k type$). This could indicate a misplaced vessel by ship type (e.g., a fishing vessel transiting on a tanker route, and, to some extent, can

be considered an *off-route* anomaly in a broader sense). Distances in vessel type are not considered in the present study.

A. Track classification and off-route vessel detection

The distance $\Delta_p^k = \Delta_p(\mathbf{V}, k)$ is used to test the spatial proximity of the vessel track $\mathbf{V}$ from to all the routes R_k, together with the route width w_k. We consider sequentially the received data points of the vessel track, starting from the first received one.

Thresholding the distance of a track from a route: The route width w_k is defined as the maximum of the distances of each route cluster point (vessel positions associated to the route in the learning phase) to the closest waypoint on the "average route" $\bar{R}_k$:

$$w_k = \max_{z=1}^{P} d(v_z, R_k) \tag{9}$$

where

$$d(v_z, R_k) = \min_{j=1}^{J} d(v_z, WP_j) \tag{10}$$

This distance can be again considered an approximated *directed Hausdorff distance*. In our tests we also comparatively investigated the use of an average function, instead of the maximum value, but no significant differences in the classification were observed.

The proximity test assigns a vessel track to a route R_k if the track distance Δ_p^k is below the route width w_k:

$$\Psi(\mathbf{v}_t) = \{R_k \in \mathcal{R} | \Delta_p^k \leq w_k\} \tag{11}$$

This means that more than one route can be compatible with the observed track.

Once the vessel track $\mathbf{V}$ has been assigned to at least one route, the related VOI is declared *within-route* (as shown in Figure 1) and other features can be incrementally investigated (i.e., heading and speed), being these features conditionally independent (e.g., on a locally-based evaluation).

If the vessel track $\mathbf{V}$ cannot be assigned to any route, i.e., $\Psi(\mathbf{v}_t) = \varnothing$, the related VOI is declared *off-route* (as shown in Figure 1) and *external* contextual information (e.g., weather conditions, restricted areas, fishing areas, etc.) can help in refining the classification of its behaviour.

B. Kinematic anomaly detection

If vessel track $\mathbf{V}$ is assigned to a route or a series of routes (e.g., it is declared within-route), we can start investigating its kinematic features in terms of the COG, SOG and time intervals Δ_t between subsequent observations.

Directional distance: The distance $\Delta heading$ is used to test the angular alignment of the vessel track $\mathbf{V}$ from to the compatible routes $\Psi(\mathbf{v}_t)$. The track distance $\Delta heading$ is computed as the mean of the observed $\Delta heading_t$ for each point of the track v_t. $\Delta heading_t$ is computed as the circular distance of the track point COG_t from $\bar{COG}_j^*$ associated to the closest $\mathbf{WP}_j^*$ on the *"average route"* $\bar{R}_k$. $\Delta heading$ is normalized with the maximum observed $\Delta heading_t$ along the track.

Speed difference: In the same way, to test the speed compatibility; the distance $\Delta speed$ for the entire track is computed as the mean of the observed $\Delta speed_t$ for each point of the track v_t. $\Delta speed_t$ is computed as the absolute value of the difference between the track speed SOG_t from $\bar{SOG}_j^*$ associated to the closest $\mathbf{WP}_j^*$ on the *"average route"* $\bar{R}_k$. $\Delta speed$ is normalized with the maximum observed $\Delta speed_t$ along the track.

Track kinematic anomaly score: The final track kinematic anomaly score AD_V is computed as the average of the two kinematic contributions. It has an upper bound at 1 and is directly proportional to the likelihood $P(\mathbf{V}|R_k)$ of observing a track, given the compatible route R_k.

If a vessel is in the proximity of at least one route and the kinematic anomaly score AD_V is greater than a given threshold Th, the track $\mathbf{V}$ is declared *anomalous*, due to its kinematic features and further investigations are needed. This score will not be used in the results of section VII.

C. Track sequence anomaly score

Another contribution to the anomalous track detection is provided by accounting for the likelihood of the temporal sequence $\Delta Tseq$ of the track. This anomalous track sequence detector assumes both the SOG and COG information are available at each point of the track and can be triggered once the track is associated to a route. The aim is to detect *changes* in the track sequence, once it is associated to a route. If so, we can compute the probability of transition from one observed point of a track (i.e., state, following a Markov-like approach) to another while the track is evolving over time indicated by $\Delta Tseq$. This probability can be computed only if the track has been associated to at least one route and aims at detecting on-line deviations from the assigned route. The probability, $P(\Delta Tseq|R_k) = P(\mathbf{V}|R_k)$, of the observation sequence, $\mathbf{V}$, for the state sequence, $\bar{\mathbf{S}}$, given the route, R_k, can be then expressed as follows:

$$P(\Delta Tseq|R_k) = P(\mathbf{V}|R_k) = \prod_{t=1}^{T} P(\mathbf{v}_t|R_k). \tag{12}$$

where:

$$P(\mathbf{v}_t|R_k) = \exp\left[-\left(\frac{\Delta_p^*}{\alpha_k}\right)^{\beta_k}\right]. \tag{13}$$

The α_k and β_k estimates are obtained of each route R_k using the sampled distances Δ_p^* between the predicted points $[\hat{x}_t, \hat{y}_t]$, and the actual observed points, $[x_t, y_t]$, in the specified route, R_k, for each given time lag, Δ_t, using Maximum Likelihood methods. These results have been fully detailed and discussed in [3]. The concept is here framed into the more general scenario of track classification. This allows to compute the corresponding Δ_p^* as a function of the specific time lag Δ_t along the track, without requiring re-sampling techniques to be able to process the track sequence. The distance, Δ_p^*, can be used to estimate the likelihood of observing the track state at time t, given the previous state at time $t-1$, along the route, R_k. In this way, a consistent transition probability for the considered likelihood estimation problem is obtained, as

discussed in [3] and this can be used to set up an on-line anomaly test of the time sequence.

Anomaly test of the time sequence: The detection of an anomaly, H_1, at time, *t*, can be thought of as deviation from the normality, H_0, learned using historical data and can be approached by setting a minimum threshold in Equation (14), according to the detection and false alarm rates required by the specific surveillance application:

$$\arg\max_{k} P(\mathbf{V}|R_k) \underset{H_0}{\overset{H_1}{\gtrless}} Th \tag{14}$$

where **V** is the observed track for the Vessel Of Interest (VOI).

In order to perform the anomaly detection on-line a sliding time window, which captures only the most recent points of the partially observed track, can be used.

VII. Illustrations and Results on Real Data

The proposed T2R classification method was tested on a reference data set of AIS data recently published by CMRE [19].

A. Case study: Castellana reference data set

The Area of Interest (AOI) covers an area of 46 x 60 nautical miles in the Ligurian Sea in front of La Spezia Harbour. 904 vessels transited in total in the considered scenario from January 1st to February 20th 2013, over the considered AOI. The area includes La Spezia Harbour, where different typologies of ship traffic are observed. From each AIS report the vessel's Maritime Mobile Service Identity (MMSI), the timestamp, the latitude and longitude of the vessel (in decimal degrees), the vessel speed over ground (in knots), the course over ground (in degrees, North referenced) and the ship type (coded according the AIS standards, e.g., 30 stands for fishing vessel) were processed. The route extraction via TREAD provides the atlas of the main routes prototypes. We selected 18 route prototypes to set up our validation method, as shown in Fig. 4.

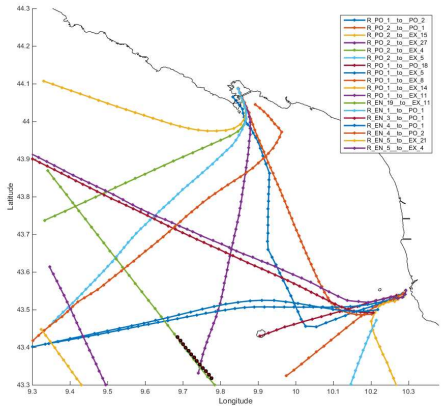

Fig. 4. Historical route prototypes extracted via the TREAD model in the area from AIS data (Jan 1- Feb20 2013)

B. Some results

For each extracted route R_k with $k = 1, ...18$, a route model has been computed, as discussed in Section IV-B. Each route was then decomposed into a set of the elementary trajectories of the vessels which transited along that route in the given time window. These trajectory data were then further split into tracks (i.e., partial segments of trajectories) and they were ingested into the T2R algorithm in order to test the T2R association algorithm on incomplete tracks. Each extracted track contains 10 data points (corresponding to an average track duration of about 20 minutes, given the average AIS time rate in the area).

We provide here some results about the proposed methods. The T2R association percentages are reported in Table I.

First, we analyzed the association performance in terms of the percentage of tracks associated to the extracted system of routes and the total number of processed tracks. In particular, 94 is the percentage of the tracks, originated from the extracted system of routes, which were associated to the extracted system of routes (and classified as "on-route").

Secondly, we estimated the accuracy of the anomaly detectors, which is equivalent to the overall rate of correctly classified anomalous tracks, as discussed in [6]. We generated synthetic tracks, in order to counterbalance for the lack of ground truth which is one of the main issues when assessing the performance of anomaly detectors. To this aim, we started from the "on-route" tracks and altered their features (e.g., position, heading and speed) independently, in such a way they can reflect the Maritime Situational Indicators discussed in Section III. An aggregated percentage of 87.3 tracks were correctly detected as anomalous. One result from the analysis is that the level of traffic variability of the originating routes affects the association performance of both normal and abnormal tracks and is a direct consequence of the data-driven approach, even in dense traffic areas.

TABLE I
Track-to-Route Association Rates

		Detector output	
Ground truth		***Normal***	***Anomalous***
Normal	on-route	***94***	*6*
Anomalous	off-route	11	89
	on-route but high speed	12	88
	on-route but opposite heading	15	85
	Total	*12.6*	***87.3***

Figure 5 reports an example of a measurement of a track sequence, using the method presented in Section VI-C. This method suites the on-line classification of a track, once it is associated to a route, by computing the transition probability using jointly the COG and SOG information. This can help highlighting unexpected evolutions of the vessel motion along the route.

VIII. Conclusions and Future Work

A method has been presented for associating and classifying tracks (i.e., partial trajectories) enhanced by the traffic routes derived via TREAD model for route extraction. The method is data-driven: routes are learned from the historical data for

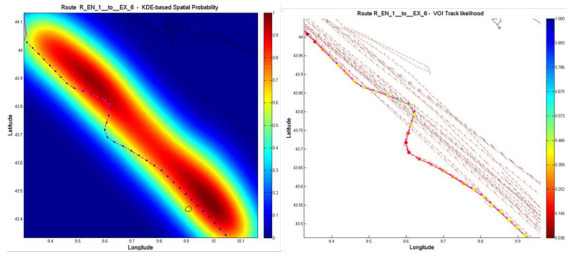

Fig. 5. Route represented via KDE-computed extent (left); Example of a low-likelihood track classified using the sequential scoring method discussed in Section VI-C. (right)

the area of interest to get performance results suitable to the users needs and proper thresholds must be selected.

The proposed approach has been tested on a real data set supplemented with simulations of abnormal tracks. In the light of these preliminary results, we observed that the association accuracy is affected by the level of traffic variability in the area. In [4] the route cluster quality has been discussed. As a consequence, the detection of anomalies can be linked to the traffic entropy: the capability to successfully recognize low-likelihood behaviors is enhanced in areas where the traffic patterns are highly regular and therefore, the associated level of disorder is low. The effect of entropy on association performance will be further investigated as part of future research.

The spatial proximity is the primary factor affecting the track association. The algorithm assigns to multiple routes a track when this falls in the proximity of overlapping motion patterns with similar flow directions (i.e., kinematic features).

Due to the complex nature of the maritime surveillance and anomaly detection, the presented algorithm is meant be one component of a larger integrated system of maritime anomaly detection to improve automated processing techniques. Indeed, the risk of high false alarm rates when using automatic systems can be mitigated using multiple sensors and sources, in order to get a confirmation from complementary sources.

The choice of thresholds is a critical task in the design of an anomaly detection strategy. As part of the future work, a sensitivity study is planned in comparison with other available methodologies (such as [20]), in order to assess the impact of the thresholds on the anomaly detector performance. Operators can provide guidelines about the the deviations of interest in the kinematic features of the vessel tracks. To this aim, a support in the threshold selection can be provided by the context, represented via routes (as discussed in the present paper) and via "external" contextual constraints if available (e.g., direction/speed indications from the traffic separation schemes in proximity of ports or in highly dense areas). A planned extension of the present work will consider the evolution of route representation using semantic regions (common sub-paths shared by different routes) similar to the approach presented in [21]. Also, the differentiation of route layers filtered by ship-type will be considered, in order to enhance the detection of the anomaly "The type of the vessel is not compatible with the route followed" as discussed in [12].

ACKNOWLEDGEMENTS

The authors wish to thank the NATO Allied Command Transformation (NATO-ACT) for supporting the CMRE project on Maritime Security (MSEC).

REFERENCES

[1] R. Lane and K. Copsey, "Track anomaly detection with rhythm of life and bulk activity modeling," in *Information Fusion (FUSION), 2012 15th International Conference on*, July 2012, pp. 24–31.
[2] S. of Life at Sea (SOLAS) convention Chapter V, "Regulation 19."
[3] 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, pp. 2218–2245, 2013.
[4] ——, "Traffic knowledge discovery from ais data," in *16th International Conference on Information Fusion*, Istanbul, Turkey, 2013.
[5] B. Ristic, B. L. Scala, M. Morelande, and N. Gordon, "Statistical analysis of motion patterns in AIS Data: Anomaly detection and motion prediction," in *11th Conference on Information Fusion (FUSION)*, 2008.
[6] R. Laxhammar, "Conformal anomaly detection," Ph.D. dissertation, University of Skvde, Skvde, Sweden, 2014.
[7] D. Makris and T. Ellis, "Learning semantic scene models from observing activity in visual surveillance," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 35, pp. 397–408, 2005.
[8] S. Mascaro, A. E. Nicholso, and K. B. Korb, "Anomaly detection in vessel tracks using bayesian networks," *International Journal of Approximate Reasoning*, vol. 55, no. 1, Part 1, pp. 84 – 98, 2014, applications of Bayesian Networks.
[9] F. Mazzarella, M. Vespe, D. Damalas, and G. Osio, "Discovering vessel activities at sea using ais data: Mapping of fishing footprints," in *Information Fusion (FUSION), 2014 17th International Conference on*, 2014.
[10] L. Cazzanti and G. Pallotta, "Mining maritime vessel traffic: Promises, challenges, techniques," in *IEEE OCEANS2015*, 2015.
[11] C.-Y. Chong, S. Mori, W. Barker, and K.-C. Chang, "Architectures and algorithms for track association and fusion," *IEEE Aerospace and Electronic Systems Magazine*, vol. 15, no. 1, pp. 5–13, jan 2000.
[12] G. De Vries and M. van Somerem, "Machine learning for vessel trajectories using compression, alignments and domain knowledge," *Expert Systems with Applications*, no. 18, 2012.
[13] N. Le Guillarme and X. Lerouvreur, "Unsupervised extraction of knowledge from s-ais data for maritime situational awareness," in *Information Fusion (FUSION), 2013 16th International Conference on*, July 2013, pp. 2025–2032.
[14] R. Laxhammar and Falkman, "Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories," *Annals of Mathematics and Artificial Intelligence*, pp. 1–28, 2013. [Online]. Available: http://dx.doi.org/10.1007/s10472-013-9381-7
[15] G. Pallotta, S. Horn, P. Braca, and K. Bryan, "Context-enhanced vessel prediction based on ornstein-uhlenbeck processes using historical ais traffic patterns: Real-world experimental results," in *Information Fusion (FUSION), 2014 17th International Conference on*, 2014.
[16] H. S. J. W. M. Ester, M.; Kriegel and X. Xu, "Incremental clustering for mining in a data warehousing environment," in *24th International Conference on Very Large Data Bases*, New York, USA, 1998.
[17] D. L. Hall and A. K. Garga, "Pitfalls in Data Fusion (and How to Avoid Them)," in *2nd International Conference on Information Fusion*, vol. 1, 1999, pp. 429–436.
[18] D. T. Etienne L. and B. A., "Spatio-temporal trajectory analysis of mobile objects following the same itinerary," *Advances in Geo-Spatial Information Science*, 2012.
[19] G. Pallotta and M. Vespe, "Castellana reference dataset," NATO UNCLASS, Tech. Rep., 2014.
[20] N. Bomberger, B. Rhodes, D. Garagic, J. Dankert, M. Zandipour, L. Stolzar, G. Castanon, and M. Seibert, "Adaptive spatial scale for cognitively-inspired motion pattern learning analysis algorithms for higher-level fusion and automated scene understanding," in *MILCOM IEEE*, 2008.
[21] K. Wang, "Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models," in *Int Journal of Computer Vision*, 2011.