Unsupervised Maritime Traffic Graph Learning with Mean-Reverting Stochastic Processes

Pasquale Coscia*, Paolo Braca†, Leonardo M. Millefiori†, Francesco A. N. Palmieri* and Peter Willett‡
*Università della Campania "Luigi Vanvitelli", Dipartimento di Ingegneria, Aversa, Italy
Email:{pasquale.coscia, francesco.palmieri}@unicampania.it
†NATO-STO Centre for Maritime Research and Experimentation (CMRE), La Spezia, Italy
Email: {paolo.braca, leonardo.millefiori}@cmre.nato.int
‡University of Connecticut, ECE Department, Storrs CT, USA
Email: peter.willett@uconn.edu

***Abstract*—Inspired by the fair regularity of the motion of ships, we present a method to derive a representation of the commercial maritime traffic in the form of a graph, whose nodes represent way-point areas, or regions of likely direction changes, and whose edges represent navigational legs with constant cruise velocity. The proposed method is based on the representation of a ship's velocity with an Ornstein-Uhlenbeck process and on the detection of changes of its long-run mean to identify navigational way-points. In order to assess the graph representativeness of the traffic, two performance metrics are introduced, leading to distinct graph construction criteria. Finally, the proposed method is validated against real-world Automatic Identification System data collected in a large area.**

***Index Terms*—Ornstein-Uhlenbeck process, change detection, maritime traffic graph, graph learning, clustering, DBSCAN, real-world data, AIS, maritime situational awareness**

I. Introduction

Synthetic Aperture Radar (SAR) and Vehicle Tracking System (VTS) devices have been fundamental appliances to support maritime surveillance during the last decades. However, the Automated Identification System (AIS), firstly conceived only for collision avoidance, has been playing an increasingly important role in support to maritime surveillance and other activities too, such as search and rescue operations, fisheries and environment monitoring, and anomaly detection. AIS data is asynchronously broadcast by ships in the form of messages that contain their identity, position, velocity, and other vessel- and voyage-related information.

Being mandatory by international regulations for ships over 300 tonnes and passengers ships, AIS devices produce considerable amounts of data on maritime traffic, which can be used to highlight recurrent or regular patterns of ships at sea. Maritime traffic, and especially commercial traffic, is in fact fairly regular, because ships, and particularly the large ones, have reduced maneuverability and always seek to optimize fuel consumption in accordance with traffic regulations.

In order to eventually support operators and decision makers, this regularity of the traffic can be exploited to *synthesize* the information in clear and effective forms, i.e. traffic patterns, that highlight important features and reject noise.

To corroborate our findings, we use real-world AIS data of cargo vessels (also shown in Fig. 1) to learn, in an unsupervised fashion, a graphical representation of the traffic, where way-point areas correspond to the nodes of the graph whereas the edges are the navigational legs. The resulting graph can be considered a synthetic representation of the traffic under normal circumstances, that can be further exploited by tracking, data fusion or anomaly detection algorithms.

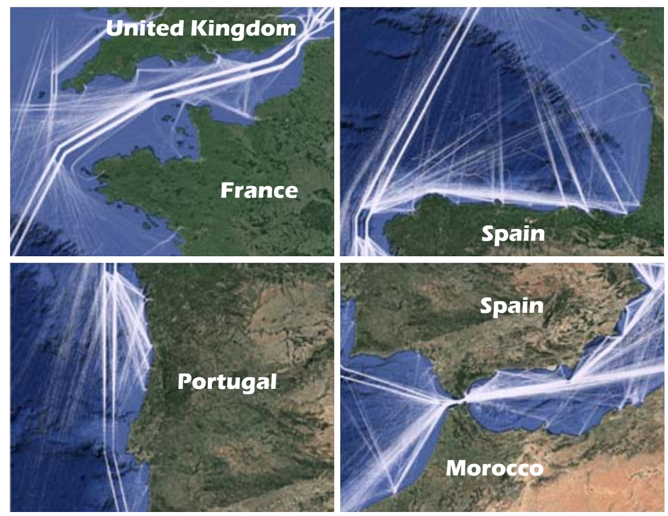

Fig. 1. Density map of the AIS messages broadcast by cargo ships in several areas during two months in 2016. The regularity of the traffic is clearly visible.

For the motion model, we rely on the Ornstein-Uhlenbeck (OU) stochastic process, which has been shown [1]–[3] to be a realistic model for the dynamics of the velocity of ships in open sea. The proposed approach, shown in Fig. 2, encompasses several steps: from the detection of way-points, to the graph construction, to the assessment of its performance.

This paper provides an overview of the work documented in [4] and is organized as follows. In Sec. II, we report a short overview of work related to this field. Sec. III is devoted to a description of the proposed approach. In Sec. IV some experimental results are reported and the related findings discussed. Finally, conclusions are drawn and future research ideas anticipated in Sec. V.

II. Related Work

Recognizing maritime traffic patterns is crucial to ultimately classify and predict activities of ships at sea. Unsurprisingly, several works are available in literature that chase this problem, mostly using historical AIS data as a starting point.

In [5], [6], the maritime traffic is modeled using decaying artificial potential fields, with the aim to detect both abnormal behaviors and potential drifts in unobserved dynamics.

In [7], the Gaussian mixture model (GMM) and adaptive kernel density estimation (KDE) are applied to the tasks of normalcy behavior identification and anomaly detection, with some limitations related to a discretization of the space in a cell grid and the simulation of abnormal behaviors. In [8], a maritime traffic monitoring system is proposed where the target detection relies on a competitive neural network (CNN); the target dynamics is modelled with a non-linear function; and the extended Kalman filter (EKF) is used for the prediction. A complete framework for knowledge discovery and anomaly detection can be found in [9], [10], where an unsupervised and incremental technique is presented to extract movement patterns of the maritime traffic from historical AIS data.

Related to the detection of anomalous events, intended as *deviations from common patterns of activity* [11], in [12] the maritime traffic in the Baltic Sea is modeled with a two-level representation. The inner layer represents each route individually, while the external one provides a high-level traffic representation, where way-points and stationary areas are obtained from AIS data. In [13], [14], the authors use discrete locations and velocities to determine abnormal behaviors and predict future vessels positions with an associative learning procedure based on biological principles. In [15], a restricted region of sea is considered to discover normal as well as abnormal ship behaviors. In [16], a spline-based clustering approach is presented that is able to detect abnormal vessels behaviors in coastal surveillance scenarios.

More recently, mean-reverting stochastic processes has been proposed to model the velocity of ships in open sea and the approach has been validated against a large real-world data set [1]. This modelling, which is also relevant to types of data other than AIS [3], finds application for anomaly detection [17] too, while its use for traffic pattern discovery is documented in this work and more thoroughly in [4].

III. Proposed Framework

The proposed framework is depicted in Fig. 2. It comprises several stages, which will be detailed in the next sections.

A. Dynamic Vessel Modelling

Building upon [1], we model vessel dynamics with OU stochastic processes. More precisely, a ship's velocity is modelled with an OU process, so that it has a tendency to drift towards a central value, with a greater attraction when the process is further away from it. This central velocity value, or long-run mean velocity, represents the *desired* velocity of the ship. The main implication of this modelling is that the velocity process has bounded variance. The target state is represented by the four-dimensional vector $\boldsymbol{s}(t) = [\boldsymbol{p}(t), \dot{\boldsymbol{p}}(t)]$, where $\boldsymbol{p}(t)$ and $\dot{\boldsymbol{p}}(t)$ are the position and velocity of the target, respectively, in a two-dimensional Cartesian coordinate system. The dynamics of the target is ruled by the stochastic differential equation (SDE)

$$\dot{\boldsymbol{s}}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{B}\boldsymbol{v} + \boldsymbol{W}\dot{\boldsymbol{n}}(t), \tag{1}$$

where $\boldsymbol{v} = [v_x, v_y]^T$ is the long-run mean velocity and $\boldsymbol{n}(t)$ is a standard 2-D Wiener process. The matrices $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{W}$ are defined as follows:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{0}_2 & \boldsymbol{I}_2 \\ \boldsymbol{0}_2 & -\boldsymbol{\Gamma} \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{0}_2 \\ \boldsymbol{\Gamma} \end{bmatrix}, \quad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{0}_2 \\ \boldsymbol{H} \end{bmatrix}, \tag{2}$$

where $\boldsymbol{0}_2$ and $\boldsymbol{I}_2$ are the 2-by-2 null and identity matrices, respectively, $\boldsymbol{H}$ is a generic bi-dimensional matrix and $\boldsymbol{\Gamma} \in \mathbb{R}^{2\times 2}$ quantifies the mean-reversion effect. If $\boldsymbol{\Gamma}$ has linearly independent eigenvalues then can be factorized as $\boldsymbol{T}\boldsymbol{\Theta}\boldsymbol{T}^{-1}$, where $\boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{\gamma})$ and $\boldsymbol{T}$ contains the eigenvectors of $\boldsymbol{\Gamma}$. The diag operator applied to the vector $\boldsymbol{\gamma} = [\gamma_1, \gamma_2]$ stores its elements on the diagonal components of a diagonal matrix. Introducing now a discretization of the time, the first moment of the solution of (1) represents the target evolution, and can be arranged in matrix form as follows:

$$\begin{aligned} \boldsymbol{s}_k = {} & \widetilde{\boldsymbol{T}}\boldsymbol{\Phi}(t_k - t_{k-1}, \boldsymbol{\gamma})\widetilde{\boldsymbol{T}}^{-1}\boldsymbol{s}(t_{k-1}) \\ & + \widetilde{\boldsymbol{T}}\boldsymbol{\Psi}(t_k - t_{k-1}, \boldsymbol{\gamma})\boldsymbol{T}^{-1}\boldsymbol{v} + \boldsymbol{n}_k, \end{aligned} \tag{3}$$

where $\widetilde{\boldsymbol{T}} = \boldsymbol{I}_2 \otimes \boldsymbol{T}$, being $\otimes$ the Kronecker product. The full expression of $\boldsymbol{\Phi}(t, \boldsymbol{\gamma})$ and $\boldsymbol{\Psi}(t, \boldsymbol{\gamma})$ can be found in [1], [4], along a more detailed discussion about the target state and its time evolution.

B. Velocity changes detection

In order to reveal way-point areas, i.e. spatial regions where vessels typically change their direction, we assume that at an unknown time instant t^* the long-run mean velocity of the process might abruptly change from a value $\boldsymbol{v}_{\mathrm{in}} \neq \boldsymbol{0}$ to another value $\boldsymbol{v}_{\mathrm{out}} \neq \boldsymbol{0}$. We will formalize the problem after the change detection theory [18], in order to identify the time instants of possible changes. To this aim, let us consider the difference between two consecutive target velocity samples defined conveniently as follows:

$$\boldsymbol{z}_k = \dot{\boldsymbol{p}}_k - \boldsymbol{J}\dot{\boldsymbol{p}}_{k-1}\boldsymbol{z}_k \sim \mathcal{N}(\boldsymbol{z}; (\boldsymbol{I} - \boldsymbol{J}_k)\boldsymbol{v}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{J}_k = \boldsymbol{T}e^{-\boldsymbol{\Theta}(t_k - t_{k-1})}\boldsymbol{T}^{-1}$, $\boldsymbol{v} \in \{\boldsymbol{v}_{\mathrm{in}}, \boldsymbol{v}_{\mathrm{out}}\}$ and being $\boldsymbol{I}$ the identity matrix. Since such difference is a Gaussian random variable with definite mean and variance, a way-point can be detected through a change in the mean of $\boldsymbol{z}_k$. More specifically, the detection procedure is based on Page's test, which makes use of the clipped CUSUM statistic, defined as follows

$$S_k = \max\left\{0, S_{k-1} + \log\frac{f_{\mathrm{out}}(\boldsymbol{z}_k)}{f_{\mathrm{in}}(\boldsymbol{z}_k)}\right\}, \quad S_0 = 0, \tag{4}$$

where $f_{\mathrm{in,out}}(\boldsymbol{z}_k) = \mathcal{N}((\boldsymbol{I} - \boldsymbol{J}_k)\boldsymbol{v}_{\mathrm{in,out}}, \boldsymbol{\Sigma})$. A change is declared if $S_k > s$, where s is a suitable fixed threshold.

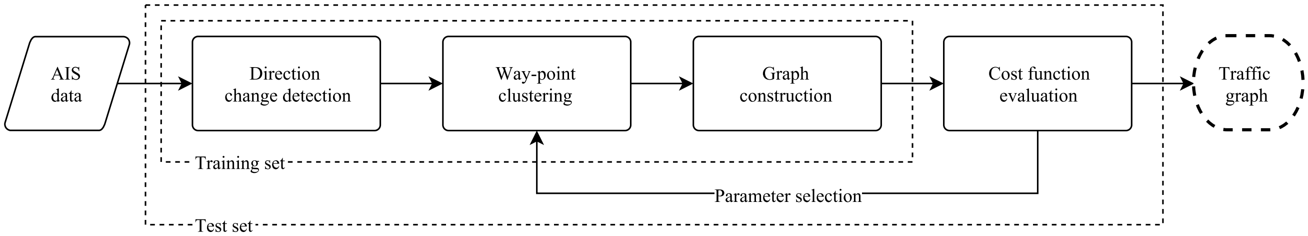

Fig. 2. Flowchart of the proposed method, consisting of four main steps: detection of way-points; clustering of way-points; learning of the adjacency matrix of the graph; and evaluation of the cost functions to optimize the choice of the parameters of the clustering algorithm. The output is the adjacency matrix of the traffic graph.

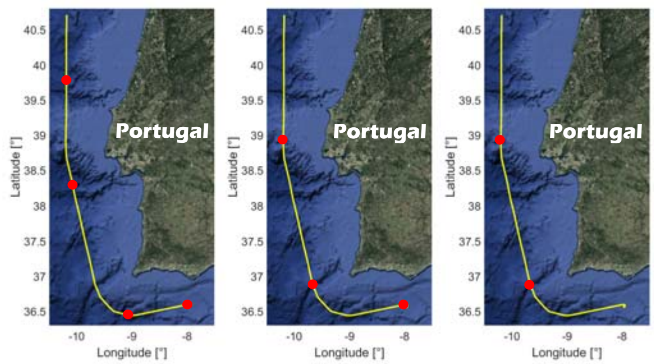

Fig. 3. Change detection procedure applied on a vessel's track along the Iberian Coast. We use three different values for the threshold s, namely $0.1, 4, 10$, while N is set to 15. The images show how the value of the threshold controls the detection capability. Too low values of the threshold would lead to a high numbers of detections, while too high values would either prevent the detection or worsen its performance in terms of average run length.

Since the values of the long-run mean velocity under the two hypotheses, that a change in the parameter happened or not, are not known in advance, Page's test is not directly applicable, but several steps are required in order to detect arbitrary changes in the parameter:

1) the value $\boldsymbol{v}_{\text{in}}$ is estimated using a sample mean estimator (SME) [2] on N velocity samples;
2) two CUSUM statistics are initialized in parallel to detect possible positive and negative changes of the direction. Specifically, assuming a constant speed between two consecutive positions, i.e. $||\hat{\boldsymbol{p}}_{n+1}|| = ||\hat{\boldsymbol{p}}_n||$, we consider the two following statistics:
$$\angle\hat{\boldsymbol{p}}_{n+1} = \begin{cases} \angle\hat{\boldsymbol{p}}_n + \delta & \textit{(for positive changes)}, \\ \angle\hat{\boldsymbol{p}}_n - \delta & \textit{(for negative changes)}; \end{cases} \tag{5}$$
3) if one of the statistics exceeds the threshold s, time and position at the time of change are estimated and the new hypothesis of no change updated accordingly.

Fig. 3 shows a graphical representation of the results of this procedure applied to real data.

C. Way-point Clustering

The change detection procedure leaves us with a set of directional change positions and long-run mean ship velocities before and after the time instants of change. Since commercial maritime traffic appears fairly regular for the examined areas, we can assemble such elements to identify regions of sea where there is a tendency of ships to change their navigational course. These regions will represent the nodes of the maritime traffic graph. To this aim, we use a popular density-based clustering algorithm, DBSCAN [19]. Compared to other clustering algorithm (k-means or GMM), DBSCAN can recognize regions of arbitrary shape and reject points whose feature vectors are too distant from those of the other clusters. To reveal dense regions (clusters), the algorithm requires two user-defined parameters: *minPoints*, which defines the minimum number of points to create a connected region, and ϵ, which defines the maximum radius of a point's neighborhood in the feature space. Intuitively, DBSCAN looks for regions where points are very close in the feature space and identifies outliers as points lying in low-density regions. When such regions are *density-connected* one another, they are recursively joined in the same cluster, until only non-reachable points remain unclustered, which are marked as outliers.

As distance function, we use the Mahalanobis distance in the four-dimensional feature space. The main reason of this choice for the distance function is the inhomogeneity of the components of the feature vectors, made up by positions and angles.

In fact, the feature vector is defined as $\boldsymbol{f}_i = [\boldsymbol{d}_i, \alpha_i^{(\text{in})}, \alpha_i^{(\text{out})}]$ where $\boldsymbol{d}$ is the position of the change estimated with Page's test, in a geographic coordinate system, and $\alpha_i^{(\text{in})}$ and $\alpha_i^{(\text{out})}$ are the directions before and after the change, respectively. Note that, thanks to the assumption of unchanged speed before and after the change, it is not necessary to include the velocity vectors $\boldsymbol{v}_{\text{in}}$ and $\boldsymbol{v}_{\text{out}}$ in the feature vectors. The distance between two feature vectors, $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$, is therefore $D_M(\boldsymbol{f}_i, \boldsymbol{f}_j) = \sqrt{(\boldsymbol{f}_i - \boldsymbol{f}_j)^T \boldsymbol{Q} (\boldsymbol{f}_i - \boldsymbol{f}_j)}$ where the weighting matrix $\boldsymbol{Q}$ is set as $\boldsymbol{Q} = \text{diag}(1, 1, \pi/18, \pi/18)$ in order to consider $1°$ and $10°$ significant changes in geographic positions and directions, respectively.

Then, for each established cluster, we estimate the distributions of the input and output velocity directions. A Gaussian approximation is encompassed in this step, which, however, proves to be realistic when compared with real data, as it is shown in Fig. 4. Therefore, a cluster (or node of the graph) N_i is defined by the 4-D vector $N_i = [\boldsymbol{c}_i, \boldsymbol{\mu}_i^{(\text{in,out})}, \boldsymbol{\sigma}_i^{(\text{in,out})}]$

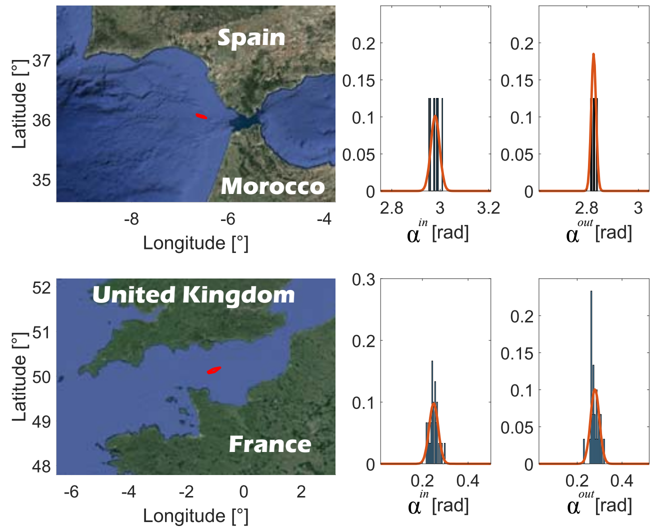

Fig. 4. Left: two clusters along the Strait of Gibraltar and the English Channel, represented by their convex hulls. Right: histograms of the velocity directions for the input and output changes along with the estimated Gaussian probability distributions.

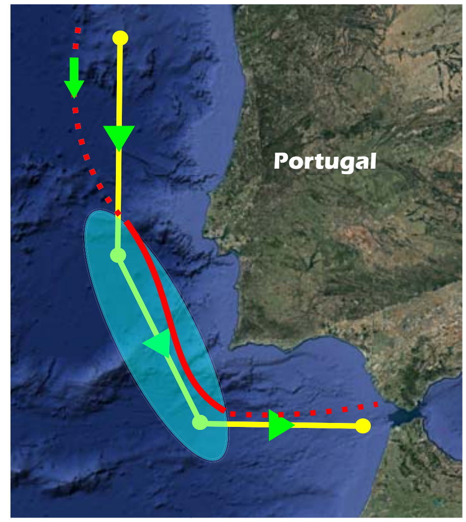

Fig. 5. Graphical representation of the association procedure of the geometric criterion with the graph edges. The vessel's track (in red) is associated to a sea-lane (in yellow) if it belongs to its elliptic area, with a cost penalty that accounts for a possible misalignment between the direction of the ship and the orientation of the edge.

which encloses the cluster's centroid, mean and variance of the estimated probability distributions, respectively. The input and output parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are obtained as the sample mean and standard deviation of the empirical distributions.

D. Graph Construction

The compact representation of the maritime traffic that we propose is based on a directed graph, whose nodes and edges represent regions of likely changes of directions, i.e. navigational way-points, and navigational legs, respectively.

The graph is defined by its adjacency matrix $\mathcal{G} \in \mathbb{N}^{N_n \times N_n}$, being N_n the number of nodes. The adjacency matrix is constructed from the data itself by simply recording the transitions of ships from a generic pair of nodes i and j, in the specific order, as to also record the direction of movement. Each edge of the graph has a corresponding navigational leg, which is represented with a 5-D vector, namely $L_i = [\boldsymbol{p}_{1_i}, \boldsymbol{p}_{2_i}, m_i]$, where $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are the centroid positions of the way-point clusters of the origin and arrival nodes and m_i is a weight proportional to the number of vessels which transitioned from the origin to the arrival nodes. Finally, it is worth mentioning that we estimate a finite simple graph, so diagonal elements of $\mathcal{G}$ are set to zero to avoid meaningless loops and that one node could be linked to more nodes simultaneously.

E. Graph Validation Criteria

A graph constructed as above should ideally provide a synthetic representation of the maritime traffic, or at least the regular part of it. In order to understand to which extent the resulting graph is representative of maritime traffic, two performance criteria are defined in this section to assess the representativeness of the graph when compared to real data. The first criterion is more focused on the association of maritime traffic with the navigational legs, i.e. the edges of the graph, while the second one gives more importance to way-points and direction changes, i.e. the nodes of the graph.

1) Geometric criterion: With the first performance criterion, we consider the surrounding area of a navigational leg, i.e. a graph edge. The larger the area is, the more vessels can be associated to it. A simple and reasonable assumption for the shape of the surrounding area is an ellipse whose foci are the nodes connected by the edge and whose major axis is the edge itself. The eccentricity parameter rules the total deviation across the leg (see Fig. 5). To verify if a vessel can be considered as in navigation on a leg, the elliptic region is defined as

$$E(\boldsymbol{p}) := \boldsymbol{p}^T \boldsymbol{Q}^T \boldsymbol{M} \boldsymbol{Q} \boldsymbol{p},$$

where

$$\boldsymbol{Q} = \begin{bmatrix} \cos\alpha & -\sin\alpha & x_0\cos\alpha - y_0\sin\alpha \\ \sin\alpha & \cos\alpha & x_0\sin\alpha + y_0\cos\alpha \\ 0 & 0 & 1 \end{bmatrix} \tag{6}$$

and

$$\boldsymbol{M} = \begin{bmatrix} 1/a^2 & 0 & 0 \\ 0 & 1/b^2 & 0 \\ 0 & 0 & -1 \end{bmatrix}. \tag{7}$$

Eventually, we verify the following condition: $E(\boldsymbol{p}) \leq 0$. However, vessels could be in the elliptical region that surrounds the navigational leg but not aligned with its direction. To account for this effect, we define an association cost c as follows:

$$c = w\left(\frac{|E(\boldsymbol{p})|}{||\boldsymbol{F}_{12}||_2} + d(\dot{\boldsymbol{p}}, \boldsymbol{F}_{12})\right) \tag{8}$$

where w is the weight of the edge and $d(\dot{\boldsymbol{p}}, \boldsymbol{F}_{12})$ returns the angle between the ship's velocity $\dot{\boldsymbol{p}}$ and the major axis of the ellipse, $\boldsymbol{F}_{12}$; both addends are suitably normalized. The

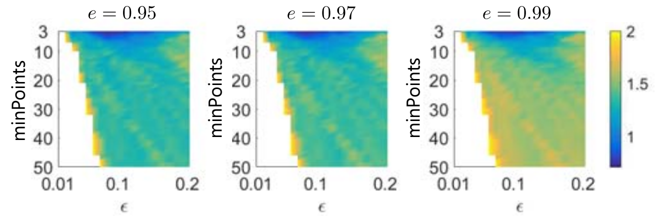

Fig. 6. Cost J_{gc} (geometric criterion) for several eccentricity values. The white color refers to combinations of $minPoints$ and ϵ for which neither graph nodes nor edges are identified. Based on the geometric criterion, the optimal choice of $(minPoints, \epsilon)$ is that for which J_{gc} is minimum.

TABLE I

OPTIMAL VALUES OF THE DBSCAN PARAMETERS FOR SEVERAL ECCENTRICITY VALUES WITH CORRESPONDING GLOBAL COST J_{gc} COMPUTED WITH THE GEOMETRIC CRITERION

	$minPoints$	ϵ	J_{gc}
$e = 0.95$	3	0.11	0.5849
$e = 0.97$	3	0.11	0.6116
$e = 0.99$	3	0.11	0.7016

problem of multiple associations due to overlapping ellipses is avoided selecting the leg at minimum cost c. The global cost J_{gc} is the sample mean of each cost c_i, i.e. $J_{gc} = N^{-1} \sum_i c_i$, which is then averaged w.r.t. the total number of tracks.

2) Cluster-based criterion: With the second approach, the change detection procedure is performed on the test set to identify points of direction change. These points are then compared with the way-point clusters, i.e. the nodes of the graph. The more points belong to one of the clusters, the more the clusters and, consequently, the graph, should be representative of the ship traffic.

The first step is to associate change detections to the closest cluster in terms of Euclidean distance; then the compatibility of the directions before and after the change detection is tested against the input and output direction distributions of the cluster. This amounts to check if

$$\alpha_{\text{in,out}} \in \mathcal{I} = [\mu_{\text{in,out}} - h\sigma_{\text{in,out}},\, \mu_{\text{in,out}} + h\sigma_{\text{in,out}}],$$

where h rules the desired confidence level.

Finally, we compute the global cost $J_{cc} = N^{-1} \sum_j [\mathbb{1}_{\mathcal{I}}(\alpha_{\text{in}_j}) \cdot \mathbb{1}_{\mathcal{I}}(\alpha_{\text{out}_j})]$, where $\mathbb{1}_{\mathcal{I}}(\alpha) = 1$ if $\alpha \in \mathcal{I}$ and otherwise is zero. Finally, the global cost is averaged w.r.t. the total number of tracks.

IV. EXPERIMENTAL RESULTS

The proposed approach has been applied to a real-world dataset of about 11 000 AIS messages broadcast by commercial cargo ships in 2016 from June, 1st to July, 31st in a fairly large area that includes all the Iberian Coast and up to the English channel. Of the about 11 000 messages, 8000 have been used as training set, the remaining being the test set.

The dataset is shown in Fig. 1, where we report a density map of the spatial distribution of the AIS messages in the area of interest; the overall regularity of the traffic is apparent. The parameters required by the proposed approach have been set as follows: $\mathbf{\Theta} = \theta \boldsymbol{I}_2$, $\boldsymbol{H} = h\boldsymbol{I}_2$, with $\theta = 3 \times 10^{-3}$ and $h = 14.1 \times 10^{-3}$, $s = 4$, $N = 15$, $\delta = 5°$.

Since different values of the parameters of the clustering algorithm, i.e., $minPoints$ and ϵ, lead to significantly different resulting graphs, one first analysis is devoted to characterize the effects of the clustering parameters on the resulting graph. In our analysis, we use the validation criteria defined in Sec. III-E to identify the $(minPoints, \epsilon)$ pairs that minimize or maximize, respectively, the global costs J_{gc} and J_{cc}.

Specifically, we selected the intervals $minPoints \in [3, 50]$ and $\epsilon \in [0.01, 0.2]$. The choice of the clustering parameters is usually the result of a trade-off between graph resolution and fragmentation. In this analysis, the trade-off is found by minimizing or maximizing the cost functions J_{gc} and J_{cc}, respectively.

In Fig. 6 it is reported a graphical representation of the cost J_{gc}, defined in Sec. III-E, as function of the parameters of the clustering algorithm and for three distinct values of eccentricity, i.e., $0.95, 0.97$ and 0.99. As the figure shows, the cost is directly proportional to the eccentricity value. In other words, small clusters seem to be favored by this criterion, as this leads to the highest number of associations with navigational legs. In particular, Table I reports the optimal values, along with the corresponding minimum costs for the three eccentricity values. A qualitative representation of the resulting graph is shown in Fig. 8, where the main commercial routes, along with the connections to the most important ports in the area have been correctly identified. Note that some edges of the graph (e.g. in proximity of the Strait of Gibraltar) are unrealistically drawn on land. This effect can be credited to the graph construction stage, which, at the moment, only checks for ship transitions from one cluster to the other, not including any land avoidance logic.

The second performance criteria is validated using three different values for the interval $\mathcal{I} = [\mu - h\sigma, \mu + h\sigma]$. Unlike the previous one, the performance of the graph is now inversely proportional to the cost function, due to the definition of J_{cc} provided in Sec. III-E2, meaning that a value of $J_{cc} = 1$ corresponds to the *best* possible graph representation of the traffic. As shown in Fig. 7 and reported in Table II, this criterion identifies three optimal networks varying the width of the confidence interval $\mathcal{I}$. The resulting graphs are shown in Fig. 9, where a Traffic Separation Scheme (TSS) is clearly visible and captured by the graph, even though no sea-lanes are detected along the Iberian Coast, due to the limited number of detected clusters in that area.

From the qualitative analysis done, it is clear how the first performance criterion tends to favor a more detailed traffic description at the cost of higher numbers of graph edges, which hinder to capture the main navigational routes. Conversely, such routes are correctly captured when using the second criterion to build the graph, even though several important sea lanes are not identified.

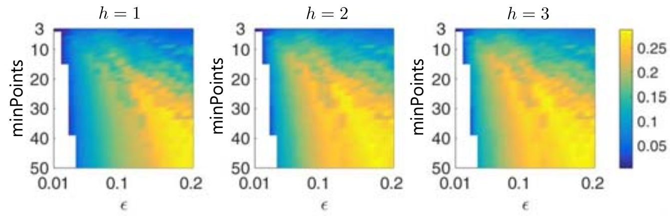

Fig. 7. Cost J_{cc} (cluster-based criterion) for several confidence levels. Better graph representations correspond to higher values of J_{cc} and are obtained with $minPoints \sim 40$. The white color refers to combinations of $minPoints$ and ϵ for which neither graph nodes nor edges are identified. Based on the cluster-based criterion, the optimal choice of $(minPoints, \epsilon)$ is that for which J_{cc} is maximum.

TABLE II
OPTIMAL VALUES OF THE DBSCAN PARAMETERS FOR SEVERAL CONFIDENCE LEVELS WITH CORRESPONDING GLOBAL COST J_{cc} COMPUTED WITH THE CLUSTER CRITERION

	$minPoints$	ϵ	J_{cc}
$h = 1$	34	0.20	0.1883
$h = 2$	42	0.17	0.2561
$h = 3$	35	0.20	0.2873

V. CONCLUSION

We presented an unsupervised procedure to learn a synthetic graph representation of the maritime traffic from ship mobility data. The velocity of ships that navigate by way-points is statistically modeled with multiple Ornstein-Uhlenbeck processes with different long-run mean velocities. Based upon this modelling, a statistical change detection procedure is performed to detect changes in a ship's direction through phase changes of the long-run mean velocity. Eventually, change detections can be clustered in groups of similar elements to become the nodes of the graph, and ship transitions from node to node to instantiate the edges of the graph.

In order to assess the representativeness of the graph when compared to real traffic data, two different cost functions are introduced, which can also be maximized or minimized to determine the parameters of the clustering algorithm. Findings were corroborated using real ship traffic data collected during two months in a fairly large area.

Since traffic patterns are clearly the results of several factors, including weather conditions and maritime navigation rules, our future work will also focus on the inclusion of these elements in the synthesis process to provide a more comprehensive sight into the dynamics of commercial ship traffic.

ACKNOWLEDGMENT

This work was supported by the NATO Allied Command Transformation (ACT) via the project Data Knowledge Operational Effectiveness (DKOE) at the NATO Science and technology Organization (STO) Centre for Maritime Research and Experimentation (CMRE). P. Willett was supported by NPS via ONR contract N00244-16-1-0017.

REFERENCES

[1] 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.

[2] L. M. Millefiori, P. Braca, and P. Willett, "Consistent estimation of randomly sampled Ornstein-Uhlenbeck process long-run mean for long-term target state prediction," *IEEE Signal Process. Lett.*, vol. 23, no. 11, pp. 1562–1566, Nov 2016.

[3] G. Vivone, L. M. Millefiori, P. Braca, and P. Willett, "Performance assessment of vessel dynamic models for long-term prediction using heterogeneous data," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6533–6546, nov 2017.

[4] 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.*, pp. 1–1, to be published.

[5] E. Osekowska, S. Axelsson, and B. Carlsson, *Potential Fields in Modeling Transport over Water*. Cham: Springer International Publishing, 2015, pp. 259–280.

[6] E. Osekowska, H. Johnson, and B. Carlsson, "Maritime vessel traffic modeling in the context of concept drift," *Transportation Research Procedia*, vol. 25, pp. 1457–1476, 2017, world Conference on Transport Research - WCTR 2016 Shanghai. 10-15 July 2016.

[7] R. Laxhammar, G. Falkman, and E. Sviestins, "Anomaly detection in sea traffic – A comparison of the Gaussian Mixture Model and the Kernel Density Estimator," in *2009 12th International Conference on Information Fusion (FUSION)*, July 2009, pp. 756–763.

[8] L. P. Perera, P. Oliveira, and C. G. Soares, "Maritime traffic monitoring based on vessel detection, tracking, state estimation, and trajectory prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1188–1200, Sept 2012.

[9] M. Vespe, I. Visentini, K. Bryan, and P. Braca, "Unsupervised learning of maritime traffic patterns for anomaly detection," in *9th IET Data Fusion & Target Tracking Conference (DF&TT 2012): Algorithms & Applications*. IET, 2012.

[10] 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, jun 2013.

[11] L. Snidaro, C. Piciarelli, and G. L. Foresti, "Fusion of trajectory clusters for situation assessment," in *2006 9th International Conference on Information Fusion (FUSION)*, July 2006, pp. 1–7.

[12] V. F. Arguedas, G. Pallotta, and M. Vespe, "Maritime traffic networks: From historical positioning data to unsupervised maritime traffic monitoring," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[13] N. A. Bomberger, B. J. Rhodes, M. Seibert, and A. M. Waxman, "Associative learning of vessel motion patterns for maritime situation awareness," in *2006 9th International Conference on Information Fusion (FUSION)*, July 2006, pp. 1–8.

[14] B. J. Rhodes, N. A. Bomberger, and M. Zandipour, "Probabilistic associative learning of vessel motion patterns at multiple spatial scales for maritime situation awareness," in *2007 10th International Conference on Information Fusion (FUSION)*, July 2007, pp. 1–8.

[15] M. Fiorini and S. J. Chang, "Vessel route analysis by in situ measurements. a pre-requisite for maritime spatial planning," in *2012 12th International Conference on ITS Telecommunications*, Nov 2012, pp. 380–384.

[16] A. Dahlbom and L. Niklasson, "Trajectory clustering for coastal surveillance," in *2007 10th International Conference on Information Fusion (FUSION)*, July 2007, pp. 1–8.

[17] E. d'Afflisio, P. Braca, L. M. Millefiori, and P. Willett, "Detecting stealth deviations from standard routes using the Ornstein-Uhlenbeck process," in progress.

[18] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

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

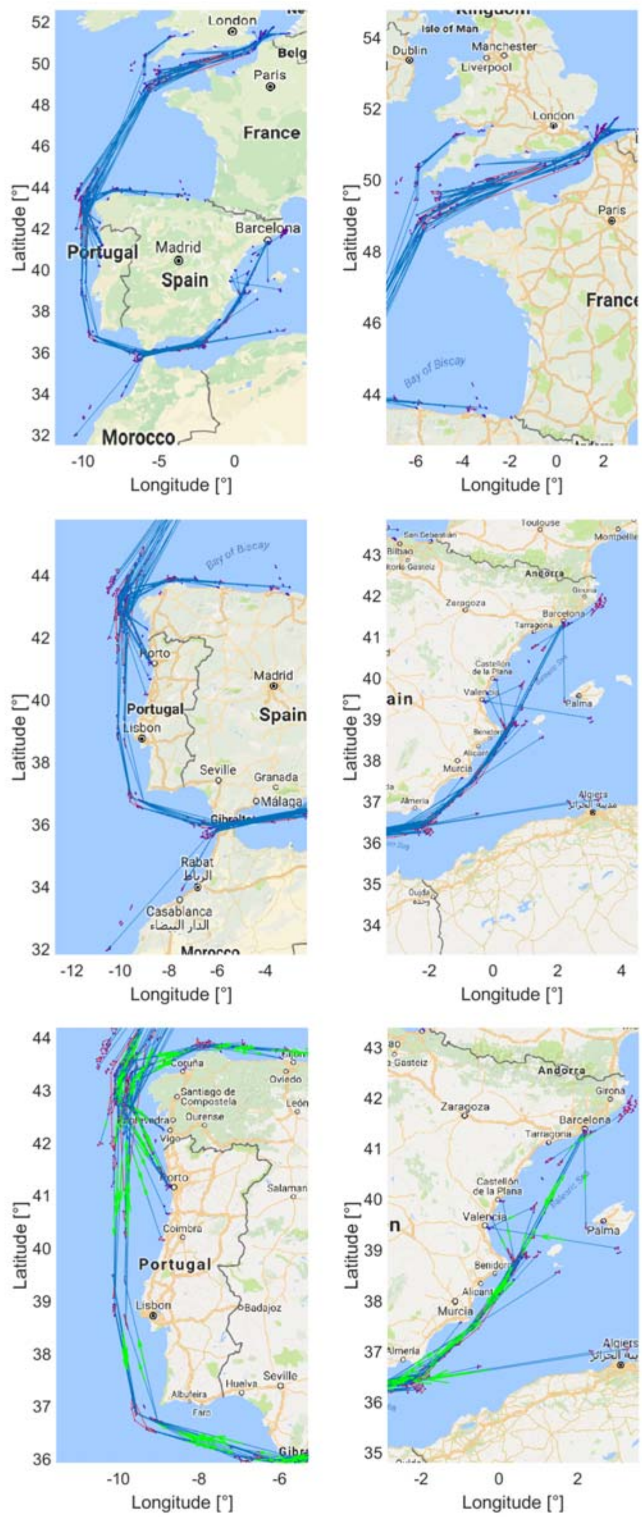

Fig. 8. Graphs constructed with the geometric criterion. There are approximatively 650 nodes and edges. Panels in the bottom row show also the orientation of the edges.

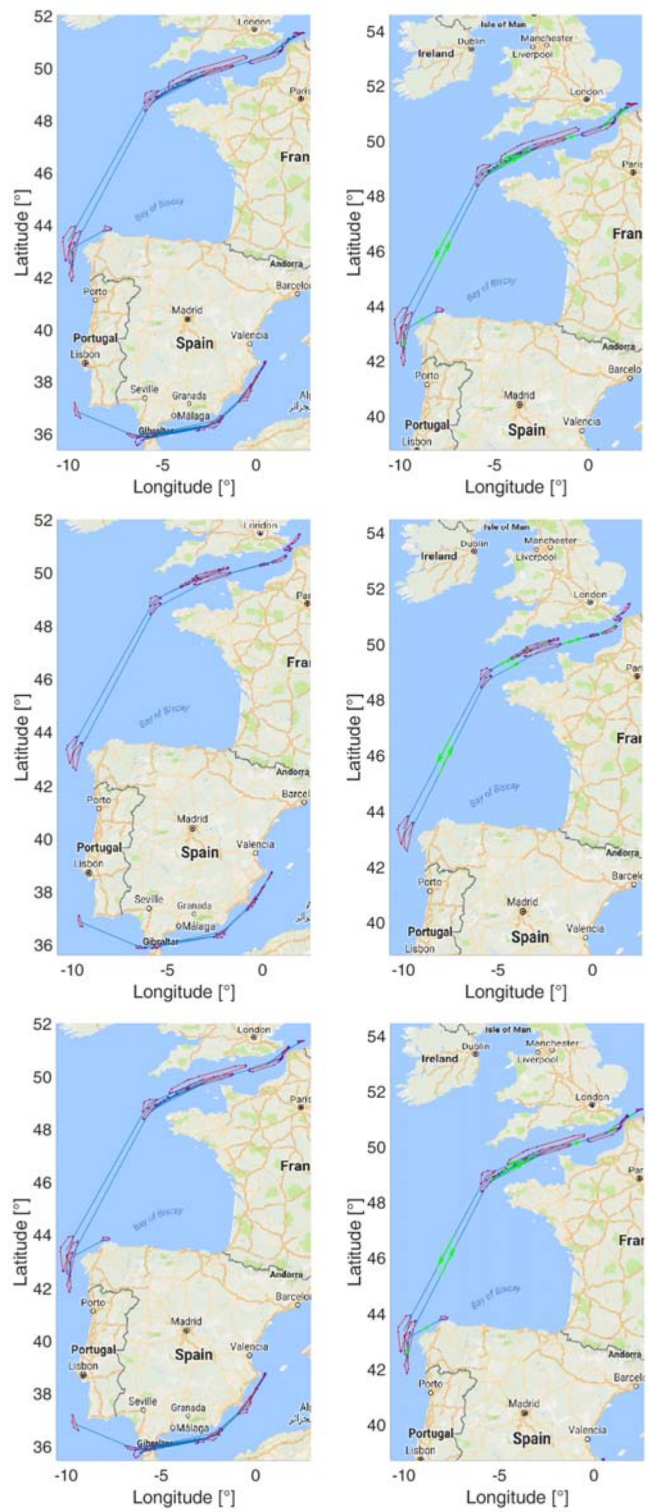

Fig. 9. Graphs constructed with the cluster-based criterion. There are approximatively 30 nodes and edges. Panels on the right column show also the orientation of the edges.