

Multiple Ornstein-Uhlenbeck Processes for Maritime Traffic Graph Representation

Pasquale Coscia, Paolo Braca, *Senior Member, IEEE*, Leonardo M. Millefiori, *Member, IEEE*,
Francesco A. N. Palmieri, *Member, IEEE* and Peter Willett, *Fellow, IEEE*

Abstract—We propose an unsupervised procedure to automatically extract a graph-based model of commercial maritime traffic routes from historical Automatic Identification System (AIS) data. In the proposed representation, the main elements of maritime traffic patterns, such as maneuvering regions and sea-lanes, are represented, respectively, with graph vertices and edges. Vessel motion dynamics are defined by multiple Ornstein-Uhlenbeck (OU) processes with different long-run mean parameters, which in our approach can be estimated with a change detection procedure based on Page’s test, aimed to reveal the spatial points representative of velocity changes. A density-based clustering algorithm (DBSCAN) is then applied to aggregate the detected changes into groups of similar elements and reject outliers. To validate the proposed graph-based representation of the maritime traffic, two performance criteria are tested against a real-world trajectory data set collected off the Iberian Coast and the English Channel. Results show the effectiveness of the proposed approach, which is suitable to be integrated at any level of a JDL system.

Index Terms—DBSCAN, Graph-based Maritime Traffic Representation, Maritime Surveillance Awareness (MSA), Page’s Test, OU Process.

I. INTRODUCTION

The increasing demands of the global market are leading to a significant growth of the worldwide maritime traffic. A well-known fact about maritime traffic is that the majority of it is very regular: ships, especially those involved in goods transportation often, if not always, seek to optimize fuel consumption, and therefore will naturally follow the most convenient path allowed by international regulations and traffic separation schemata. In other words, it is apparent that the number of routes among specific points of interest (POIs), such as ports or stationary areas, is therefore limited.

In order to fully support human users and decision makers, it is crucial to understand the underlying structure of maritime

traffic. As the traffic has increased, so have the risks of ship collisions, piracy attacks and terrorism threats. Thanks to technological advances and international regulations, a global surveillance capability for ships at sea is closer day by day. Some issues remain, especially in terms of persistence and coverage.

Synthetic aperture radar (SAR) and coastal radar systems have for many years been the primary surveillance workhorse. However, recent maritime traffic monitoring networks have emerged based on the AIS that ships of certain categories required to have.

The AIS is a collaborative, self-reporting system that ships over 300 tonnes and all passenger ships have to be installed on board, as mandated by the Safety of Life at Sea (SOLAS) convention issued by the International Maritime Organization (IMO). AIS messages are regularly broadcast by ships and received by other ships, or terrestrial or satellite receivers. The ITU 1371-4 defines 64 different types of AIS messages that ships can broadcast; in this work we only focus on those that contain the position of the ship (latitude and longitude) and other information related to the ship’s motion (e.g. speed and course). In this work, we will use AIS data broadcast by cargo ships and collected by a global network of coastal AIS receivers from the 1st of June to the 31st of July, 2016, in an area spanning from 31.5° to 51.80° in latitude and from −15.5° to 3.8° in longitude.

The approach herein set forth is inspired by the works presented in [1]–[3], where the validity of the Ornstein-Uhlenbeck (OU) process to model vessel dynamics in open sea is demonstrated in terms of performance improvement compared to the more commonly adopted Nearly-Constant Velocity (NCV) model.

The main contribution of this work is a compact description of the maritime traffic by means of a graph-based representation. The nodes represent common waypoints, i.e. areas where ships are more likely of changing course, while edges are the recurrent navigational legs (sea lanes). Also, two performance criteria are proposed to adjust the representation according to the desired density.

Several Maritime Situational Awareness (MSA) applications could benefit from a network representation of maritime traffic: long-term prediction, anomaly detection, prevention from piracy at sea and decision support systems [4], [5], just to name a few. Finally, a graph-based representation of maritime traffic could also improve the low-level tasks of a data fusion system, e.g. tracking and prediction, as well as the high-level context for domain knowledge [6].

P. Coscia was with the NATO Science and Technology Organization (STO) Centre for Maritime Research and Experimentation (CMRE), Viale San Bartolomeo 400, 19126, La Spezia, Italy. He is now with the Dipartimento di Ingegneria Industriale e dell’Informazione, Università degli Studi della Campania “Luigi Vanvitelli”, Via Roma 29, 81031, Aversa, Italy. Email: {pasquale.coscia, francesco.palmieri}@unicampania.it.

P. Braca and L. M. Millefiori are with the NATO Science and Technology Organization (STO) Centre for Maritime Research and Experimentation (CMRE), Viale San Bartolomeo 400, 19126, La Spezia, Italy. Email: {paolo.braca, leonardo.millefiori}@cmre.nato.int.

F. A. N. Palmieri is with the Dipartimento di Ingegneria Industriale e dell’Informazione, Università degli Studi della Campania “Luigi Vanvitelli”, Via Roma 29, 81031, Aversa, Italy. Email: {pasquale.coscia, francesco.palmieri}@unicampania.it.

P. Willett is with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs CT, 06269-2157 USA. Email: peter.willett@uconn.edu.

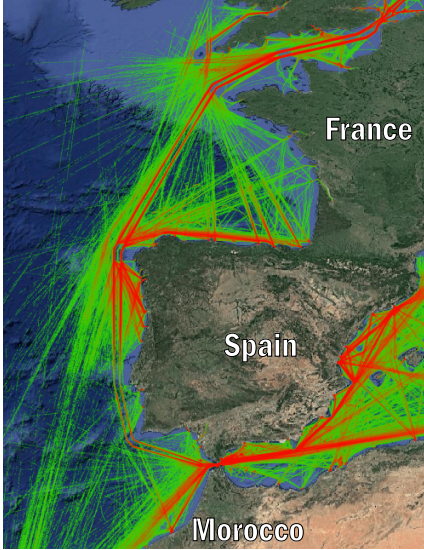


Fig. 1. Density map of AIS messages collected from cargo ships during two months in the area of interest. Several waypoints, i.e. localized spatial regions where vessels show velocity or orientation changes are observable. Linear piecewise lanes that make up standardized routes followed by vessels are clearly visible. The color is proportional to the number of ships whose reported positions fall within the ground footprint, from green (low traffic density) to red (high traffic density).

The paper is organized as follows. Section II presents a brief overview of the approaches previously proposed to create a network representation of the maritime traffic. Section III describes the vessel dynamics model. Section IV introduces a graph-based representation of maritime traffic. In Section V, the performance criteria for the network are introduced. Finally, Sections VI and VII are devoted to the presentation of the experimental results, conclusions and suggestions for future developments.

II. RELATED WORK

Extracting recurrent motion patterns is very important to achieve a useful understanding of the maritime traffic and it is also a very active research area [7]–[11]. The applications of this research are various; for example, in [9] the uncertainty is integrated to a reasoning system to improve the quality of the estimation using the paradigm of the Markov Logic Networks (MLNs) defining weighted first-order formulas, and in [12] a decision support system is proposed to monitor high risk areas, treating it as an optimization problem that accounts for fuel consumption, environmental risks, traffic and the specifics of the covered area. There is little work that addresses the graphical structure that might describe spatio-temporal moving objects data. Some that does is applied to video surveillance (e.g., [13]), but unfortunately its relevance to the maritime domain is unclear; and what does specifically focus on maritime traffic (e.g., [7]) still leaves many research issues open.

Among the trajectory pattern mining techniques available in the literature, the great majority rely on AIS data to provide support to decision makers in critical situations [14], [15]. In [16], AIS data is used to estimate the seaport location and

the operational areas combining an adaptive Kernel Density Estimation (KDE) with the MapReduce processing framework. Ship trajectory patterns are extracted in [17] with a density-based approach that uses the IMO rules during the clustering of trajectories in close-to-port operations. The clustering therein employed is based on speed and direction, as well as spatial positioning. The major drawback of this approach is the lack of scalability: in any relevant real-world application, such a system must be able to capture patterns which cover large areas, dealing with very large volumes of data. Similarly to the approach presented in this work, in [18] a framework is proposed to build a probabilistic normalcy model and demonstrated using S-AIS data for behavior analysis and path prediction. However, the method appears unable to capture patterns in high density areas increasing the related uncertainty. Finally, in [19], [20] the authors propose a Dynamic Bayesian Network (DBN) based on a topological map for port areas to detect abnormal behaviors. The approach considers a cumulative sum updated via triggered events when a vessel moves from zone to zone. The normalcy models are built using simulated data. Even though it appears well-suited for limited areas such as harbors, the proposed approach needs a large number of reliable sensors, e.g. cameras and trackers, to attain a high percentage of success.

III. DYNAMIC MODEL

To model the vessel dynamics we use the following linear Stochastic Differential Equation (SDE) [1], [2]:

$$\dot{s}(t) = \mathbf{A}s(t) + \mathbf{G}v + \mathbf{B}\dot{w}(t) \quad (1)$$

where $s(t) = [p(t), \dot{p}(t)]^T$ is the four-dimensional target state with the vessel position and velocity in a 2-D coordinate system and $w(t)$ is a standard bi-dimensional Wiener process. The main difference between this model and the NCV one that is more widely adopted in the target tracking literature, is that here a feedback loop is included that regulates the velocity of the target, so that there is a tendency of the walk to move towards a central velocity value, with a greater attraction when the process is further away from it. The matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{G}$ are defined as:

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{0} & -\mathbf{\Theta} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{0} \\ \mathbf{C} \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} \mathbf{0} \\ \mathbf{\Theta} \end{bmatrix}, \quad (2)$$

where $\mathbf{0}$ and $\mathbf{\Theta}$ are 2×2 generic matrices and the matrix $\mathbf{C}$ represents the process noise. A key parameter of the equation above is the term v , which represents the control input and specifies the *desired* vessel's velocity. Without the feedback loop, the velocity of the target would tend to drift over time; with the loop, the velocity of the target instead tends to revert to the value v as it gets farther away from it. This mean-reversion effect is quantified by the matrix $\mathbf{\Theta}$, whose components rule how quickly the process velocity will tend back to the control input v once that the process value has moved away from it. Its diagonal elements quantify this effect along the x and y components, respectively, while the off-diagonal elements quantify the coupling effects between them.

More specifically, to model a target's velocity, we apply the OU mean-reverting stochastic process, which is the solution

of (1) when the long-run process mean $\mathbf{v} = [v_x, v_y]^T$ is constant. Recent works [1], [2] have shown that mean-reverting stochastic processes, such as OU, are better suited to represent the velocity of non-maneuvering vessels than other stochastic processes commonly adopted in the target tracking literature, as they do not suffer of the divergence with time [21], [22].

Assuming positive and distinct eigenvalues, the matrix Θ can be decomposed as $\Theta = \mathbf{R}\mathbf{\Gamma}\mathbf{R}^{-1}$ where the columns of the matrix $\mathbf{R}$ are the eigenvectors of Θ and $\mathbf{\Gamma}$ is a diagonal matrix whose elements are the corresponding eigenvalues.

As already mentioned, the OU process is the solution of (1), and its first moment can be computed analytically [23] and rearranged in a matrix form as follows (see also [2] for more details):

$$\mathbf{s}(t|t_0) = \tilde{\mathbf{R}}\Phi(t-t_0, \gamma)\tilde{\mathbf{R}}^{-1}\mathbf{s}(t_0) + \tilde{\mathbf{R}}\Psi(t-t_0, \gamma)\mathbf{R}^{-1}\mathbf{v}, \quad (3)$$

where

$$\tilde{\mathbf{R}} = \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}, \quad \Phi(t, \gamma) = \begin{bmatrix} \mathbf{I} & (\mathbf{I} - e^{-\mathbf{\Gamma}t})\mathbf{\Gamma}^{-1} \\ \mathbf{0} & e^{-\mathbf{\Gamma}t} \end{bmatrix} \quad (4)$$

and

$$\Psi(t, \gamma) = \begin{bmatrix} t\mathbf{I} - (\mathbf{I} - e^{-\mathbf{\Gamma}t})\mathbf{\Gamma}^{-1} \\ \mathbf{I} - e^{-\mathbf{\Gamma}t} \end{bmatrix}. \quad (5)$$

If we consider only the velocity term in (3), we have:

$$\dot{\mathbf{p}}_k = \mathbf{v} + \mathbf{J}_k(\dot{\mathbf{p}}_{k-1} - \mathbf{v}) + \mathbf{w}_k \quad (6)$$

where $\mathbf{J}_k = \mathbf{R}e^{-\mathbf{\Gamma}\Delta_k}\mathbf{R}^{-1}$ and $\mathbf{w}_k$ are independent zero-mean Gaussian random variables with bi-dimensional covariance Ξ_k . We used subscripted indices to denote the time dependency, i.e. $\mathbf{p}_k = \mathbf{p}(t_k)$ and $\Delta_k = t_k - t_{k-1}$. The parameters of the process can be estimated from the data in a batch fashion following the procedure described in [2], where are also provided the empirical distributions of the estimated values of $\mathbf{\Gamma}$ and $\mathbf{C}$. In this work, we only estimate the value of the long-run mean velocity $\mathbf{v}$, leaving fixed the values of the other parameters and set as follows: $\mathbf{\Gamma} = \text{diag}(\gamma_1, \gamma_2)$ with $\gamma_1 = \gamma_2 = 3 \cdot 10^{-3}$ and $\mathbf{C} = \text{diag}(c_1, c_2)$ with $c_1 = c_2 = 14.1 \cdot 10^{-3}$.

IV. GRAPH-BASED REPRESENTATION

A. Change Detection

A well-known fact about maritime traffic is that the majority of it is —unsurprisingly— very regular, as also shown in Fig. 1, where a ship traffic density map is reported. Ships, especially those involved in goods transportation, often if not always seek to optimize fuel consumption, and therefore will naturally follow the most convenient path allowed by international regulations and traffic separation schemata. Thanks to this overall regular behavior, detections of abrupt changes in the long-run velocity parameter $\mathbf{v}$ are more likely expected in proximity of waypoint areas.

The proposed model assumes that at an unknown time instant t^* there is a transition of the long-run mean velocity from a value α (hypothesis $\mathcal{H}_0$) to a value β (hypothesis $\mathcal{H}_1$) around the point $\mathbf{p}(t^*)$. Formalized via change detection theory [24], the procedure that will be described in this section

aims to detect these changes and subsequently to estimate the time instants of change.

Let us consider now the difference between two consecutive velocities samples, i.e. $\mathbf{z}_k = \dot{\mathbf{p}}_k - \mathbf{J}_k\dot{\mathbf{p}}_{k-1}$. In this way, $\mathbf{z}_k$ is a Gaussian random variable, i.e. $\mathbf{z}_k \sim \mathcal{N}(\mathbf{z}; (\mathbf{I} - \mathbf{J}_k)\mathbf{v}, \Xi_k)$ with $\mathbf{v} \in \{\mathbf{v}_0, \mathbf{v}_1\}$. Therefore, a change of the long-run mean velocity parameter can be revealed by detecting a change in the mean of $\mathbf{z}_k$. Under the assumption that all the parameters are known, Page's test [25] is the most suitable procedure to detect changes of the value $\mathbf{v}$, based on the following clipped CUSUM statistic:

$$S_k = \max \left\{ 0, S_{k-1} + \log \frac{f_1(\mathbf{z}_k)}{f_0(\mathbf{z}_k)} \right\}, \quad S_0 = 0, \quad (7)$$

where

$$f_0(\mathbf{z}_k) = \mathcal{N}(\mathbf{z}; (\mathbf{I} - \mathbf{J}_k)\alpha, \Xi_k), \\ f_1(\mathbf{z}_k) = \mathcal{N}(\mathbf{z}; (\mathbf{I} - \mathbf{J}_k)\beta, \Xi_k).$$

A change is detected when S_k exceeds a given threshold h , and the stopping time is defined as $\mathcal{K} = \min\{k : S_k > h\}$. Finally, a Maximum Likelihood procedure [24] estimates the time of change as the time of the last CUSUM reset before the stopping time. The performance of this procedure can be defined in terms of the Average Run Length (ARL) metric [24], under both $\mathcal{H}_0$ and $\mathcal{H}_1$, defined as the expected number of samples before correctly ($\mathcal{H}_1$) or erroneously ($\mathcal{H}_0$) detecting a change. The choice of the threshold h consequently affects the detection performance, as lower values of h imply a smaller ARL and therefore a higher rate of false alarms; vice versa, higher values of h will correspond to a higher ARL and a higher rate of missed detections. In turn, this will eventually impact on the overall graph structure, leading to an increased or decreased number of turning points (and, consequently, connections).

Moreover, a vessel generally maneuvers several times along its route and the parameter $\mathbf{v}$ needs to be treated as a piecewise-constant function of time, whose values are $\mathbf{v}_j, j = 0, \dots, M-1$. At the beginning of each segment ($k=0$), the first N samples are used to estimate the long-run mean velocity under the null hypothesis, with a sample mean operation: $\alpha = N^{-1} \sum_{k=0}^{N-1} \dot{\mathbf{p}}_k$. In order to detect a change using Page's test we need also the alternative hypothesis β , which is built from the null hypothesis as a deviation from it by a parameter δ . Actually, two hypotheses are compiled, to account for positive and negative deviations from the parameter under $\mathcal{H}_0$, in order to detect both possible increasing and decreasing changes. In this work, we assume that ships maintain their speed constant, and only the long-run velocity vector angle changes over a waypoint. Under these two hypotheses, the long-run velocity vector is indicated with β_+ and β_- , where $|\beta_+| = |\beta_-| = |\alpha|$ and

$$\angle \beta_+ = \angle \alpha + \delta, \\ \angle \beta_- = \angle \alpha - \delta.$$

The CUSUM test is computed in parallel for the alternative hypotheses and, if one detects a change, a waypoint is revealed and a new segment is initiated. The two alternative hypotheses,

β_+ and β_- , allow us to detect both positive and negative changes in the velocity direction with respect to α . This assumption is not entirely unrealistic, as ships seek to optimize fuel consumption and will naturally avoid any unnecessary speed reduction or increase. The parameter δ has to be determined experimentally as a trade-off between the detection performance, in terms of detection delay, and the detection of small deviations from the null hypothesis. In fact, the smaller δ , the larger will be the detection delay for a given false alarm rate.

Fig. 2 shows the result of the change detection procedure on a single track. Each detected waypoint is defined by its position and two velocity vectors, v_0 and v_1 , which represent the long-run mean velocity value before and after the change.

Remark: The CUSUM test can be applied directly on the target state, as long as it lays in a generic Cartesian coordinate system, whose specific choice will not affect the CUSUM performance. In the case of an abrupt change in the parameter, the CUSUM test will be able to detect it, independently on the specific orientation of the coordinate system. However, the choice of a specific coordinate system is still important, just from another perspective. The OU process itself is defined in a Cartesian coordinate system; in this form, it has been shown to be suitable to represent, in a statistical sense, the dynamics of vessels at sea [2]. On the other hand, ship positions are in a geographic coordinate system and therefore a suitable projection has to be applied to transform latitude and longitude coordinates into locations on the Cartesian plane. The choice of a specific projection depends on the specific application; we use the Universal Transverse Mercator (UTM) projection, but, in principle, other projections can be used in place of UTM.

B. Clustering

In order to discover significant waypoint areas, i.e. regions where vessels are more likely to show long-run mean velocity changes, the change detections need to be first aggregated in clusters. In this work we use the Density-Based Spatial Clustering Of Applications With Noise (DBSCAN)¹ algorithm [26]. As the name suggests, it is a density-based clustering algorithm and classifies a set of points into *core points*, *density-reachable points* and *outliers*. More specifically, given a set of points $\mathcal{P}$, the algorithm iteratively looks for *core points* $p_i \in \mathcal{P}$, s.t. $N_\epsilon(p_i) \geq \text{minPts}$, where minPts is the minimum number of points to create a dense region and $N_\epsilon(p_i)$ is the ϵ -neighborhood of p_i , defined as $N_\epsilon(p_i) = \{q \in \mathcal{P} | d(p_i, q) \leq \epsilon\}$, where $d(\cdot, \cdot)$ is a generic distance function. If the condition is satisfied, a cluster is initialized. Then, for each point in its ϵ -neighborhood, the core condition is evaluated and the point is possibly merged to the cluster. This evaluation is repeated until no more points can be added to the cluster, i.e. a density-connected region is completely found. Then, the process is repeated for the

unvisited points. Finally, the non-reachable points are labeled as *outliers*.

In order to perform a cluster analysis on the change detections revealed with the procedure described in the previous section, the waypoints' positions alone would not be sufficient to isolate groups of similar elements, i.e. clusters. In fact, in order to consider *similar* two different change detections, not only the positions have to be close one to the other, but also the long-run velocity mean values have to be comparable, both before and after the change. In other words, the feature vector would be six-dimensional, comprising the position of the detection (latitude and longitude coordinates), and the velocity components before and after the change. Assuming that ships maintain their speed across waypoints, the feature vector can be reduced to four dimension, i.e. the position of the detection and the velocity angles before and after the change². In the clustering analysis that will be presented later, we adopted the Mahalanobis distance to compare the change detection feature vectors. More specifically, as depicted in Fig. 4, each of the N_D detected waypoints w_{p_i} , $i = 1, \dots, N_D$, is represented by a four-dimensional vector whose elements are the position and the velocity directions before and after the change: $\mathbf{d}_i = (w_{p_i}, \theta_{0_i}, \theta_{1_i})^T$. The Mahalanobis distance between waypoint $\mathbf{d}_i$ and $\mathbf{d}_j$ is then $d_M(\mathbf{d}_i, \mathbf{d}_j) = \sqrt{(\mathbf{d}_i - \mathbf{d}_j)^T \Sigma^{-1} (\mathbf{d}_i - \mathbf{d}_j)}$, where Σ shapes the equivalent ellipsoids. We use $\Sigma = \text{diag}(1, 1, \pi/18, \pi/18)$, which scales by a standard deviation of 1° the position components and 10° the velocity directions. The values of Σ have been chosen empirically, such that clusters are continuous and well separated.

The DBSCAN stage leaves us with a set of clusters, for each of which, said $\mathcal{C}_i$, we compute an estimated probability density function (PDF) for the input and output directions. We have seen experimentally, as in the example shown in Fig. 5, that it is sufficient to assume normal³ distributions for such PDFs, i.e. $\Theta_0 \sim \mathcal{N}(\theta_0; \mu_{\theta_0}, \sigma_{\theta_0}^2)$ and $\Theta_1 \sim \mathcal{N}(\theta_1; \mu_{\theta_1}, \sigma_{\theta_1}^2)$, with means and variances estimated with sample averages.

Each cluster is therefore characterized by a set of elements: $\mathcal{C}_i = \{p_{C_i}, \mathbf{n}_0, \mathbf{n}_1\}$, where p_{C_i} is the position of the cluster's centroid, while $\mathbf{n}_0 = [\mu_0, \sigma_0]^T$ and $\mathbf{n}_1 = [\mu_1, \sigma_1]^T$ are the estimated parameters of the input and output distributions. Fig. 5 shows three clusters, together with their convex hulls and the corresponding PDFs of the input and output velocities.

The results of the clustering analysis performed with DBSCAN depend mainly on two parameters: minPts and ϵ , which define the minimum number of points to create a cluster and the searching area around a single point, respectively. Different values of these parameters may lead to quite different graphs (see Fig. 6). In Section VI the results obtained with two

¹Compared with other clustering algorithms, such as k -means or the Gaussian Mixture Model (GMM), DBSCAN has the advantage to be able to identify an arbitrary number of clusters of arbitrary shape and it does not necessarily associate a point to a cluster, i.e. it is robust to noise, making it well-suited for our application.

²Performing the cluster analysis with DBSCAN in a four-dimensional space instead of in a six-dimensional one is important not only from a computational cost perspective, but also because it helps to mitigate the *curse of dimensionality* issue, i.e. there is little difference in the distances between pair of samples as the number of coordinates increases when a measure like the Euclidean distance is used, making very difficult to find an appropriate value for ϵ .

³Circular distributions, such as the Von Mises distribution, can also be considered since the random variables are angles. However, the Gaussian approximations' tails would not be a problem and can be neglected due to the variance of a few degrees.

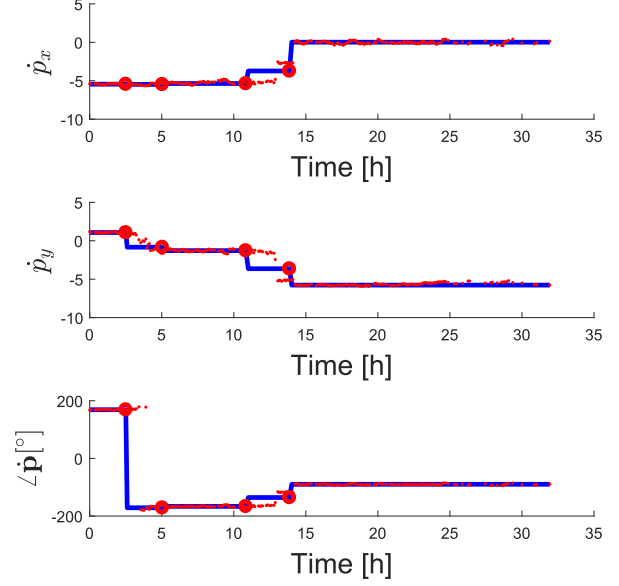
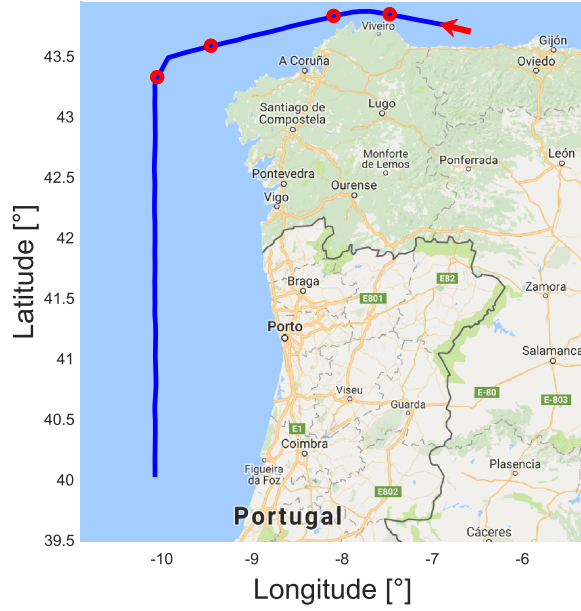


Fig. 2. Example of the change detection procedure applied to a vessel's track. On the left, the red circles are the points where a change is detected while the red arrow represents the starting point and the vessel's direction. On the right, the red dots represent the instantaneous values of the velocity along the x and y components and the track orientation while the circles the detected changes of the parameters. The parameter δ of the change detection procedure is set to 5° and $N = 15$.

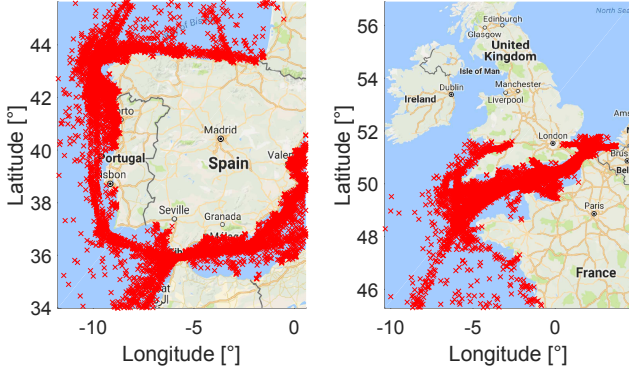


Fig. 3. Result of the change detection procedure along the Iberian Coast (on the left) and the English Channel (on the right) applied to the whole dataset. The red crosses represent the points where a vessel changes its direction of about $\pm 5^\circ$ (at least). The threshold h of the CUSUM algorithm is set to 4.

different sets of parameters are reported and compared. The resulting density and fragmentation of waypoint areas, and the consequent complexity of the graph depend to the specific application, and the calibration of the procedure is eventually left to the end user that should adjust the parameters to achieve a result that is acceptable to their needs. Ideally, minPts can be approximatively chosen as the minimum desired cluster size and ϵ based on the technique described in [26]; small values of ϵ are generally preferable, so that not so many samples are within this distance of each other.

C. Graph construction

Once the waypoint clusters have been computed, the connections between them (sea lanes or navigational legs) have

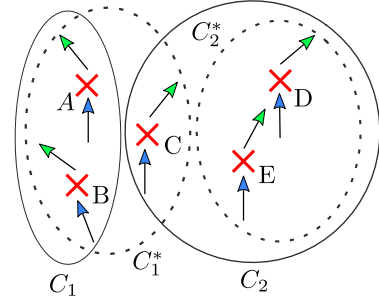


Fig. 4. Synthetic example of the use of the Mahalanobis distance and the Euclidean distance as a measure distance for the DBSCAN algorithm. The blue arrows represent the direction before the change and the green arrows the direction after the change. The points A, B and C could be grouped in a cluster, say C_1^* , while the points D and E in a different cluster, say C_2^* , if the Euclidean distance is used (dashed lines). The consideration of the direction changes leads to a better result. The different direction changes are taken into account with the Mahalanobis distance metric resulting in a better partitioning of the data (solid lines).

to be established. More specifically, if N_C is the number of clustered regions, we propose a representation as a directed graph with an adjacency matrix $T = (t_{ij}) \in \mathbb{N}^{N_C \times N_C}$, whose generic element t_{ij} is non-zero if there is a connection between cluster i and cluster j . One connection between cluster i and j is instantiated if there is at least a vessel, belonging to the training data set that have visited the i -th cluster and—immediately after—the j -th one, in the exact order; finally, diagonal elements of T are set to zero. Also, edge of the graph is weighted by a factor w_{ij} equal to the number of detected vessels which transitioned from the i -th to the j -th cluster.

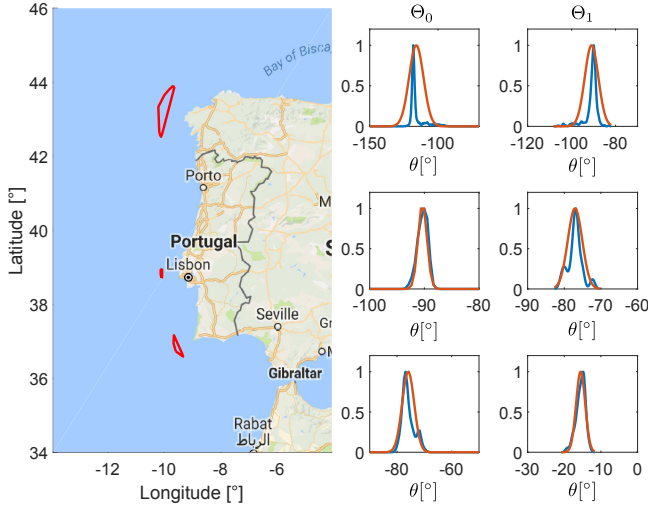


Fig. 5. Three clusters detected with the DBSCAN algorithm (on the left) and the corresponding estimated PDFs (on the right) of the input and output directions extracted as $\theta_{0,1} = \arctan(v_{y0,1}, v_{x0,1})$. The blue curves represent the result applying a Kernel Density Estimation (KDE) technique with Gaussian basis functions, while the red curves are the Gaussian approximations using the parameters estimated with the MLE technique. In a view of a compact representation, the distributions Θ_0 and Θ_1 could be therefore described by two synthetic parameters, μ and σ , due to the limited differences between the results of the two estimation approaches.

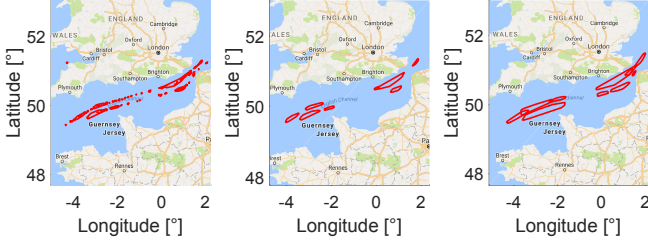


Fig. 6. Some DBSCAN results along the English Channel. Three combinations of the parameters $\minPts$ and ϵ are considered: (4, 0.05), (25, 0.1) and (45, 0.18), respectively (from left to right). Different combinations of such parameters lead to a different number of clusters and hence of routes.

Hence the edge, i.e. the navigational leg, from cluster i to j is defined by the set $\mathcal{S}_{ij} = \{p_{c_i}, p_{c_j}, w_{ij}\}$, where p_{c_i} and p_{c_j} are the 2-D position of the centroids of the two connected clusters and $w_{ij} = t_{ij}$ is the related weight. It is worth noting that the navigational leg's orientation can also be computed from the set $\mathcal{S}_{ij}$ in order to take into account the direction of motion between clusters.

Since the clusters' boundaries define convex regions, some clusters or sea lanes could cross lands leading to non-realistic graphs. To avoid this situation, we eliminate such clusters or graph edges in order to avoid meaningless intersections with lands. Examples of estimated graphs with the proposed procedure are shown in the simulation section in Fig. 13 and 14.

V. SEA-LANE PERFORMANCE CRITERIA

The objective of this section is to establish how well the constructed graph is able to represent the traffic motion and how it could be utilized for motion prediction. The proposed

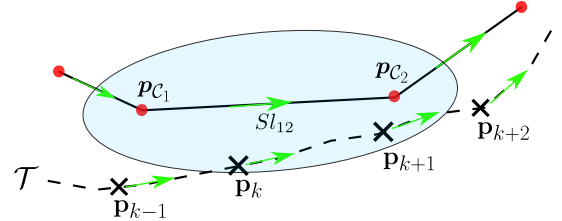


Fig. 7. Synthetic representation of the sea lane-based performance criterion. A track $\mathcal{T}$ (dashed line) and three graph branches are shown. The green arrows represent the vessel directions and the traffic motion (on the branches). The points p_k and p_{k+1} are associated to the sea lane Sl_{12} since they fall within its surrounding region while points p_{k-1} and p_{k+2} are labeled as non-associated. The total percentage of associated points is reported in Table II.

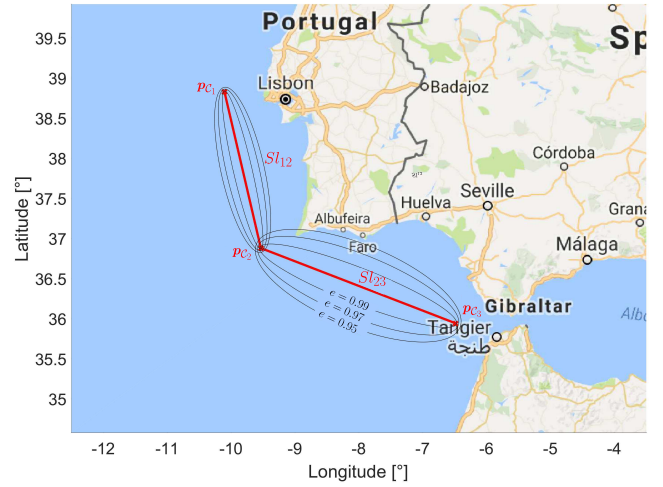


Fig. 8. Two navigational legs off the coast of the Iberian peninsula and elliptic areas defined by three different values of eccentricity: 0.95, 0.97, 0.99. The foci for the sea lane Sl_{12} are p_{c_1} and p_{c_2} while, for the sea lane Sl_{23} , are p_{c_2} and p_{c_3} respectively. The length of the semi-minor axes for Sl_{12} are 37.13 km, 28.32 km and 16.10 km while for Sl_{23} are 58.27 km, 44.43 km and 25.26 km. The length of each sea lane is 222.58 km (Sl_{12}) and 291.94 km (Sl_{23}), respectively.

cluster analysis aims at identifying distinct key points for shipping lanes, but an association criterion to match real trajectories to the estimated network is required as well. More specifically, a metric that quantifies how close real ship trajectories are to the derived graph.

Several metrics have been proposed in literature to compute trajectory distances that differ in shape and length; for example, the Hausdorff [27] or Fréchet [28] distances are among the firsts. A new metric, named *symmetrized segment-path distance* (SSPD), has been recently proposed in [29] to cluster geo-referenced points. However, to the best of our knowledge, little effort has been devoted to taking into account motion direction as well. In general track-to-graph association seems to be unrepresented in the literature; and certainly there is no maritime application of it.

Instead of treating trajectories as a whole, we consider the association between each track's point, or a subset of track's points and the network. In this regard, we introduce in the

following two different performance criteria for associating such points to the network.

A. Sea lane-based performance criterion

We introduce first a geometric criterion based on position and direction proximity, which considers only spatial information reported by the AIS transceiver.

In order to associate each track point to the graph, we consider an elliptical region around each navigational leg (graph edge), whose major axis is represented by the leg itself and whose foci are the graph nodes as shown Fig. 7. The idea is that if the region surrounding a point is circular, the region surrounding a segment should be elliptical. Therefore all the points falling into the region can be associated to that sea lane. The shape of the region can be controlled using an eccentricity parameter, as shown in Fig. 8, where two navigational legs and their elliptical regions off the Iberian Coast are depicted; the larger the eccentricity, the narrower the modeled sea lane will be. This also implies that, for longer routes, the total deviation is roughly constant across links.

1) *Ellipse Construction:* To define the ellipses around each branch, we augment the sea lane representation obtaining the set $Sl_{ij}^a = \{\mathbf{p}_{C_i}, \mathbf{p}_{C_j}, w_{ij}, a_{ij}, b_{ij}\}$, where a_{ij} and b_{ij} are the semi-major and the semi-minor axes, respectively. Fixing the eccentricity e_{ij} , the other parameters are obtained as follows:

$$a_{ij} = c_{ij}/e_{ij}, \quad b_{ij} = a_{ij}\sqrt{1 - e_{ij}^2}, \quad (8)$$

where $c_{ij} = \|\mathbf{p}_{C_j} - \mathbf{p}_{C_i}\|/2$ is the semi-focal length. The ellipse equation is obtained starting from an ellipse at the origin, translating it to the midpoint of the sea lane $\mathbf{p}_0^{ij} = [x_0^{ij}, y_0^{ij}]^T = (\mathbf{p}_{C_i} + \mathbf{p}_{C_j})/2$ and then rotating it by an angle α_{ij} , which is the orientation of the navigational leg. Dropping the indices i and j for simplicity, a point s that belongs to the ellipse, expressed in homogeneous coordinates

$$\mathbf{s} = \begin{bmatrix} s_x \\ s_y \\ 1 \end{bmatrix}, \quad (9)$$

satisfies the equation

$$E(\mathbf{s}) := \mathbf{s}^T \mathbf{Q}^T \mathbf{M} \mathbf{Q} \mathbf{s} = 0, \quad (10)$$

where

$$\mathbf{Q} = \begin{bmatrix} \cos \alpha & -\sin \alpha & x_0 \cos \alpha - y_0 \sin \alpha \\ \sin \alpha & \cos \alpha & x_0 \sin \alpha + y_0 \cos \alpha \\ 0 & 0 & 1 \end{bmatrix} \quad (11)$$

and

$$\mathbf{M} = \begin{bmatrix} 1/a^2 & 0 & 0 \\ 0 & 1/b^2 & 0 \\ 0 & 0 & -1 \end{bmatrix}. \quad (12)$$

2) *Track Association:* A track $\mathcal{T}$ is an ordered collection of spatial points, $\mathcal{T} = [\mathbf{p}_1, \dots, \mathbf{p}_n]^T$ where $\mathbf{p}_k \in \mathbb{R}^2$ are latitude and longitude coordinates provided by AIS messages, $k \in [1, \dots, n]$ and $n \in \mathbb{N}$ is its length. The association is evaluated only if the velocity vector $|\mathbf{v}_k| = |\mathbf{p}_k - \mathbf{p}_{k-1}|/\Delta_k > 0$, thus eliminating static vessels.

Now, given a vessel in position $\mathbf{p}_k = [p_{x_k}, p_{y_k}, 1]^T$ at time t_k (in homogeneous coordinates), we verify if it is inside the ellipse surrounding the sea lane Sl_{ij}^a by checking if the condition $E_{ij}(\mathbf{p}_k) \leq 0$ is satisfied.

If a point is within an elliptical area, the vessel's direction is then compared to the orientation of the navigational leg α_{ij} to verify if the direction is compatible (in fact, a vessel could be close to or even within a sea lane, but it may be sailing in an opposite direction).

3) *Association cost:* Multiple associations are certainly possible, since a vessel could be positioned near multiple sea lanes and be part of two or more elliptical regions simultaneously. The vessel must be associated to the nearest navigational leg, taking into account also the direction of movement, that has to be compared with the orientation of the legs. Consequently, to determine the best association between a vessel in position $\mathbf{p}_k$, with velocity $\mathbf{v}_k$ and a navigational leg, we define an association cost c_{ij} w.r.t. the sea lane Sl_{ij}^a as $c_{ij} = w_{ij}(\Delta p_{ij} + \Delta \theta_{ij})$, where w_{ij} is the sea lane's weight and Δp_{ij} and $\Delta \theta_{ij}$ are defined as follows

$$\Delta p_{ij} = \frac{1}{Z_p} \frac{|E_{ij}(\mathbf{p}_k)|}{\|\mathbf{f}_{ij}\|}, \quad \Delta \theta_{ij} = \frac{1}{Z_\theta} \angle(\mathbf{v}_k, \mathbf{f}_{ij}), \quad (13)$$

where $\mathbf{p}_k$ and $\mathbf{v}_k$ are position and velocity at time k while $\mathbf{f}_{ij} = \mathbf{p}_{C_i} - \mathbf{p}_{C_j}$ is the vector connecting the two foci of Sl_{ij}^a ; finally, Z_p and Z_θ are normalization factors. In other words, we evaluate the distance from a navigational leg in terms of position (Δp_{ij}) and direction ($\Delta \theta_{ij}$). The normalization factor $\mathbf{f}_{ij}$ is used to provide a bias in favor of legs closer to the point $\mathbf{p}_k$. An association is eventually made if the condition $E_{ij}(\mathbf{p}_k) \leq 0$ is satisfied.

If at time t_k a vessel is associated to multiple sea lanes, we consider a vector $\mathbf{c}$ that groups the association costs for such edges as $\mathbf{c} = [c_{ij_1}, \dots, c_{ij_N}]^T$, where N is the total number of associations. In order to determine the closest edge Sl_{ij}^a from a vessel position $\mathbf{p}_k$, we search for the best association as

$$Sl_{ij}^a = \arg \min_{ij} \mathbf{c}. \quad (14)$$

The indices i and j identify the clusters and, consequently, the best leg association, i.e. the closest edge in terms of position and direction to the point $\mathbf{p}_k$. In particular, the association is evaluated for each track $\mathcal{T}$ and point $\mathbf{p}_k$.

A change in the number of clusters may lead to a completely different set of resulting navigational legs and therefore, in order to compare different graphs with the association cost above described, a comprehensive metric is required. We define a global cost, given a graph $\mathcal{G}$, as the following summation:

$$\mathcal{E}_{\mathcal{G}} = \frac{1}{N_{\mathcal{T}}} \sum_{n=1}^{N_{\mathcal{T}}} \frac{1}{N_{A_n}} \sum_{m=1}^{N_{A_n}} (\Delta p_m + \Delta \theta_m) \quad (15)$$

where $N_{\mathcal{T}}$ is the total number of tracks and N_{A_n} is the total number of associated points for the n^{th} track.

It is clear that, in this approach, a high number of connection are created in order to reduce the global cost. Moreover, the graph obtained in this way could also be used to predict future vessel positions. For example, once a significant number of points have been associated to the graph, future positions could

be predicted by traversing the graph and accounting for the motion direction and average speed, which are typically kept as constant as possible, as described in Section I.

B. Cluster-based performance criterion

There is another approach to assess the quality of the resulting graph, thanks to the way the change detection procedure is formulated. Specifically, since the detected points represent changes of the velocity v angle along the route, one could also use the proximity of change detections to the waypoint clusters. In particular, we consider if the detected points may—or may not—be considered part of the waypoint clusters. The more change points can be considered part of an existing cluster, the more likely that cluster is actually representative of an area of common directional changes of the maritime traffic. The association is evaluated in terms of the closest cluster, considering both position and directions before and after the velocity change.

For this reason, given the detected point d_i , we firstly identify the closest cluster C_i , i.e.

$$C_i = \arg \min_{j=1, \dots, N_C} \|\mathbf{w}_{p_i} - \mathbf{p}_{C_j}\|_2,$$

where $\mathbf{p}_{C_j}$ is the centroid of the j^{th} cluster. Then, since the proximity to a cluster is not sufficient to consider the point part of that cluster, we verify if the input and the output velocity directions θ_{0_i} and θ_{1_i} are *compatible* with the estimated input and output velocity directions of the cluster, Θ_{0_i} and Θ_{1_i} respectively. In other words, we verify if the value θ falls in the intervals $\mathcal{D} = [\mu - h\sigma, \mu + h\sigma]$, for $h = 1, 2, 3$, where μ and σ are the mean and standard deviation of the estimated Gaussians PDFs. The condition is evaluated for both θ_0 and θ_1 , checking if the detected direction changes fall within one, two or three standard deviations w.r.t. the mean value μ of the estimated input and output PDFs of the nearest cluster. Considering the three-sigma rule of thumb, and since the estimated PDFs are Gaussians, the condition takes into account $\sim 68.2\%$, $\sim 95.4\%$ and $\sim 99.7\%$ of the values.

The most representative graph of the traffic is the one with the greatest number of clusters *compatible* with the detected direction changes. To this purpose, we define the goodness of the graph $\mathcal{G}$ as follows:

$$\mathcal{G} = \frac{1}{N_{\mathcal{T}}} \sum_{i=1}^{N_{\mathcal{T}}} \frac{1}{N_{D_i}} \sum_{j=1}^{N_{D_i}} [\mathbb{1}_{\mathcal{D}}(\theta_{0_j}) \cdot \mathbb{1}_{\mathcal{D}}(\theta_{1_j})], \quad (16)$$

where $\mathbb{1}_{\mathcal{D}}(\theta) = 1$ if $\theta \in \mathcal{D}$ and otherwise is zero, N_{D_i} is the number of points determined by the *change detection* procedure for the i^{th} track and $N_{\mathcal{T}}$ is the total number of tracks. From the definition of $\mathcal{G}$ follows that $0 \leq \mathcal{G} \leq 1$ and $\mathcal{G} = 1$ if and only if the overall detected direction changes are *compatible* with the ones estimated for the clusters.

VI. EXPERIMENTAL RESULTS

We have performed experiments on the real-world dataset described in the introduction to model traffic of roughly 11,000 ships in the Atlantic Ocean off the coasts of Spain, France and in the English Channel, as shown in Fig. 1. Our results

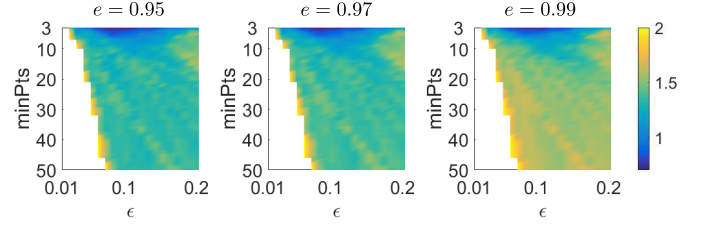


Fig. 9. Graphical representation of the error $\mathcal{E}_G$ for three values of eccentricity e . The white points represent the combination of parameters (minPts , ϵ) such that it is not possible to create a network due to no clusters or sea lanes being detected.

TABLE I
OPTIMAL VALUES OF THE TWO FREE PARAMETERS OF THE DBSCAN ALGORITHM, VARYING THE ECCENTRICITY VALUE.

	minPts	ϵ	$\mathcal{E}_G$
$e = 0.95$	3	0.11	0.5849
$e = 0.97$	3	0.11	0.6116
$e = 0.99$	3	0.11	0.7016

TABLE II
PERCENTAGE OF POINTS ASSOCIATED TO A SEA LANE, I.E. POINTS WITHIN ITS ELLIPTIC AREA. WE ALSO REPORT THE PERCENTAGE OF POINTS FOR WHICH THE ORTHOGONAL DISTANCE IS DEFINED. THE PERCENTAGE OF ASSOCIABLE POINTS, I.E. POINTS SUCH THAT $\mathbf{v}_k \neq 0$, IS EQUAL TO 98.3 AND IT IS THE SAME IN THE THREE CASES SINCE IT DOES NOT DEPEND ON THE ECCENTRICITY PARAMETER. THE PERCENTAGES REFER TO THE OPTIMAL VALUES OF minPts AND ϵ AS REPORTED IN TAB. I.

	Associated	Associated($d_{\perp}$)
$e = 0.95$	76.1%	64.5%
$e = 0.97$	71.9%	65.0%
$e = 0.99$	66.0%	64.5%

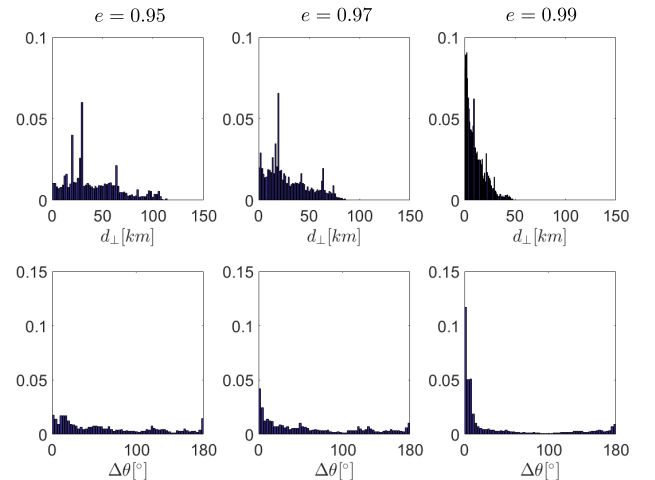


Fig. 10. Normalized histograms of the orthogonal distance $d_{\perp}(\mathbf{p}, \text{Sl}_{ij})$ between a point and the associated sea lane (first row) and the values $\Delta\theta$ which represent the distance in term of velocity direction w.r.t. the associated sea lane (second row) for the values of the eccentricity parameter.

are shown in Fig. 3 (estimated waypoints) and Fig. 5 and 6 (clusters).

In order to assess the graph model performance with the two proposed criteria, we use the 80% of the data ($\sim 8,000$ tracks) as training set and the remainder ($\sim 3,000$ tracks) for testing. Moreover, the DBSCAN algorithm is evaluated considering two intervals for the two free parameters: minPts from 3 to 50 and ϵ from 0.01 to 0.2.

Since for the first validation criterion the eccentricity e is also a free parameter, we define three possible values: 0.95, 0.97 and 0.99. As shown in Fig. 8, such values seem to capture reasonably well the most significant areas surrounding each navigational leg. Fig. 9 depicts the errors $\mathcal{E}_G$ varying the eccentricity and the DBSCAN parameters. In this case, the minimum number of points to define a cluster has to be very small to obtain a low error, which implies a greater number of sea lanes. In particular, Table I shows in detail the optimal values of the clustering parameters in the case of minimum error. The parameters minPts and ϵ are the same in the three cases which means that the ellipses' eccentricity does not affect much the association procedure. Qualitative representations of the networks are depicted in Fig. 13. The networks allow us to describe in great detail the traffic especially in high-density areas such as along the English Channel and the Strait of Gibraltar. In this regard, we consider the perpendicular distance $d_\perp(\mathbf{p}_k, Sl_{ij})$ from a point $\mathbf{p}_k$ to the navigational leg Sl_{ij} . In order to account for the Earth's curvature, in this case the great-circle distance in km is used to compute $d_\perp(\mathbf{p}_k, Sl_{ij})$, in place of the Euclidean distance. If the $\mathbf{p}_k$'s perpendicular base is outside the navigational leg's range, this distance is not evaluated. For this reason, we report in Table II the percentages of points associated to an edge and points for which the distance $d_\perp(\mathbf{p}_k, Sl_{ij})$ is defined. As expected, the percentage of points within the ellipses decreases as the eccentricity value increases. Fig. 10 shows the histograms of the distances $d_\perp(\mathbf{p}_k, Sl_{ij})$ and the values $\Delta\theta$ as defined in (13). The distributions of $d_\perp$ reach 100 km of distance, even though the higher the eccentricity parameter, the more the distributions reduce their support. The parameter $\Delta\theta$ appears much more variable even though most of the vessels move along the direction of the sea lane, especially for highest eccentricity value. Nevertheless, since $\Delta\theta$ assumes values between 0° and 180° , it is evident that there are a few vessels that move orthogonally to routes and even some that travel in the opposite direction.

The same analysis has been conducted for the second performance criterion. Fig. 11 depicts the value $\mathcal{G}$ varying the confidence interval defined by the parameter h . Table III shows the quantitative results and the optimal values for minPts and ϵ . Unlike the previous case, the number of points required to create a cluster for such a criterion is much larger than the previous one (more than 34). As expected, the number of points d_i associated to a cluster grows by increasing the confidence interval. Fig. 12 depicts the histograms for the three confidence intervals of the distance between $\mathbf{w}_{p_i}$ and the clusters' centroids $\mathbf{p}_{C_i}$ and the absolute deviations of the input and output velocity directions from the mean values of the PDFs of each cluster, i.e. $d = \text{dist}(\mathbf{w}_{p_i}, \mathbf{p}_{C_i})$, where $\mathbf{w}_{p_i}$ is

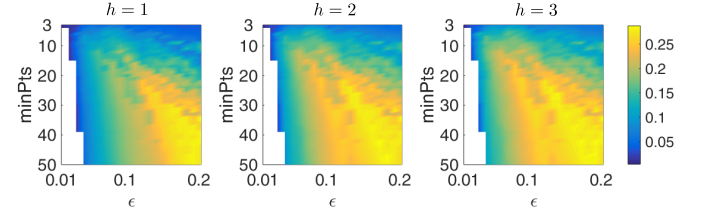


Fig. 11. Goodness $\mathcal{G}$: the maximum value needs to be considered in this case to obtain the optimal parameters. The white points represent the parameters for which it is not possible to construct a graph due to no clusters or sea lanes being detected.

TABLE III
OPTIMAL VALUES OF THE DBSCAN PARAMETERS FOR THE THREE CONSIDERED INTERVALS OF CONFIDENCE. WE ALSO REPORT THE PERCENTAGE OF POINTS ASSOCIATED TO A CLUSTER. A POINT IS ASSOCIATED TO THE NEAREST CLUSTER IF BOTH θ_0 AND θ_1 BELONG TO THE SET $\mathcal{D}$.

	minPts	ϵ	$\mathcal{G}$	Associated
$h = 1$	34	0.20	0.1883	36.5%
$h = 2$	42	0.17	0.2561	40.9%
$h = 3$	35	0.20	0.2873	43.5%

the point detected by the change detection procedure and $\mathbf{p}_{C_i}$ is the cluster's centroid, and $\Delta\theta = |\theta - \mu|$. The distance is less than 50 km while the directions typically differ in the order of 10° . Finally, a qualitative representation of the networks is depicted in Fig. 14 where a less dense network results from the elimination of some significant traffic lanes.

The qualitative analysis shows how the second approach is able to capture the underlying patterns of the traffic mainly in heavy density areas. Conversely, it is not able to highlight the most regular regions of traffic, for example along the Iberian coast. The opposite happens for the first approach mainly due to the large number of clusters which provides a much more detailed description of the monitored regions.

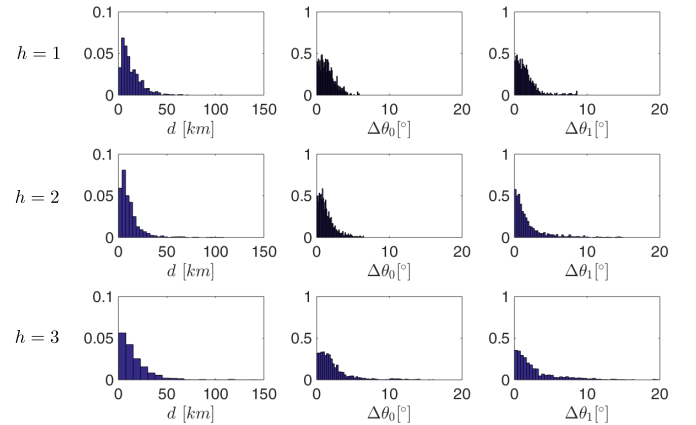


Fig. 12. Normalized histograms of the distance between the waypoints $\mathbf{w}_{p_i}$ and the clusters' centroids $\mathbf{p}_{C_i}$ and the absolute deviations $\Delta\theta = |\theta - \mu|$ for the input and output velocity directions. The rows represent the results varying the confidence interval $\mathcal{D}$.

VII. CONCLUSION

In this work we have presented an approach for compactly and automatically representing commercial maritime traffic in form of a graph, this graph with nodes are represented by clusters of waypoints, connected together by a network of navigational legs (graph edges). Two performance criteria are proposed to assess the suitability of the graph to represent the maritime traffic in a given area, one of which seems to be more suitable than the other to represent the desired level of detail. Multiple OU processes, combined with a change detection procedure, have been proven to be a valid approach to discover regions where vessels recurrently exhibit direction changes. Given the large amount of real-world data in our tests, it would seem fair to claim that our method is robust against outliers, such as abnormal or non-standard behaviors, which do not need to be previously filtered out from the data. The proposed approach could be easily integrated with data collected by other type of sensors, e.g. high-frequency surface wave (HFSW) radar, to enhance the awareness of the monitored areas.

Future work will include the development of a global-scale representation of the maritime traffic that can be leveraged for several applications in different contexts, from long-term prediction to safety and security domains. The next step will also include stop areas, in order to create a complete network which could model a greater number of situations that may happen during navigation.

ACKNOWLEDGMENT

This project was supported by NATO Allied Command Transformation (ACT) through 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, and P. Willett, "Consistent estimation of randomly sampled Ornstein-Uhlenbeck process long-run mean for long-term target state prediction," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1562–1566, Nov 2016.
- [2] L. 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, 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 Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6533–6546, Nov 2017.
- [4] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors*, vol. 37, no. 1, pp. 32–64, 1995.
- [5] —, "Designing for situation awareness in complex systems," in *Proceedings of the Second International Workshop on Symbiosis of Humans, Artifacts and Environment*, 2001.
- [6] A. N. Steinberg, C. L. Bowman, and F. E. White, "Revisions to the JDL data fusion model," pp. 430–441, 1999.
- [7] V. F. Arguedas, G. Pallotta, and M. Vespe, "Automatic generation of geographical networks for maritime traffic surveillance," in *17th International Conference on Information Fusion (FUSION)*, July 2014, pp. 1–8.
- [8] G. Pallotta, M. Vespe, and K. Bryan, "Traffic knowledge discovery from AIS data," in *Proceedings of the 16th International Conference on Information Fusion*, July 2013, pp. 1996–2003.

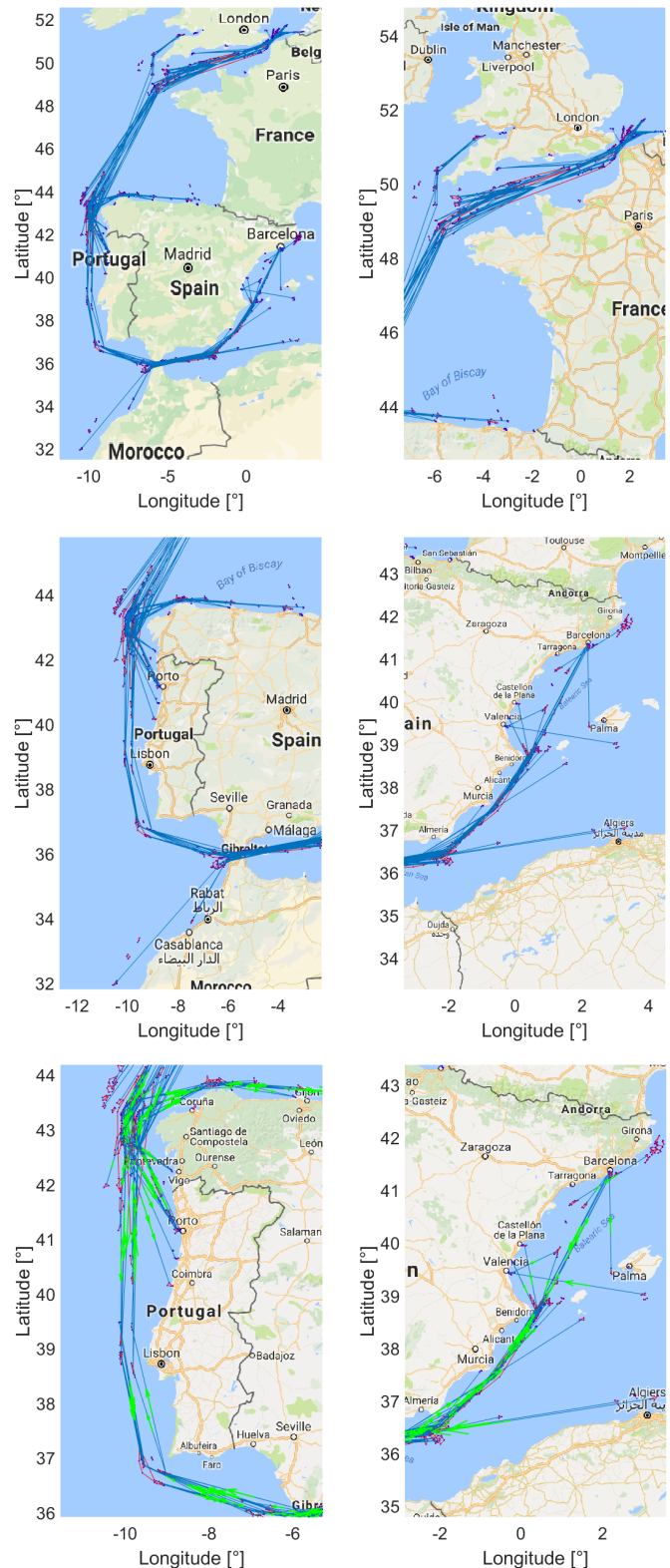


Fig. 13. Resulting network for the optimal values of minPts and ϵ using the first performance criterion. The first four pictures depict clusters and sea lanes for the analysed domain along with three detailed areas. On the bottom, two particulars of the Iberian Peninsula are reported where the traffic directions are highlighted with green arrows. The number of nodes and edges are equals to 602 and 706, respectively.

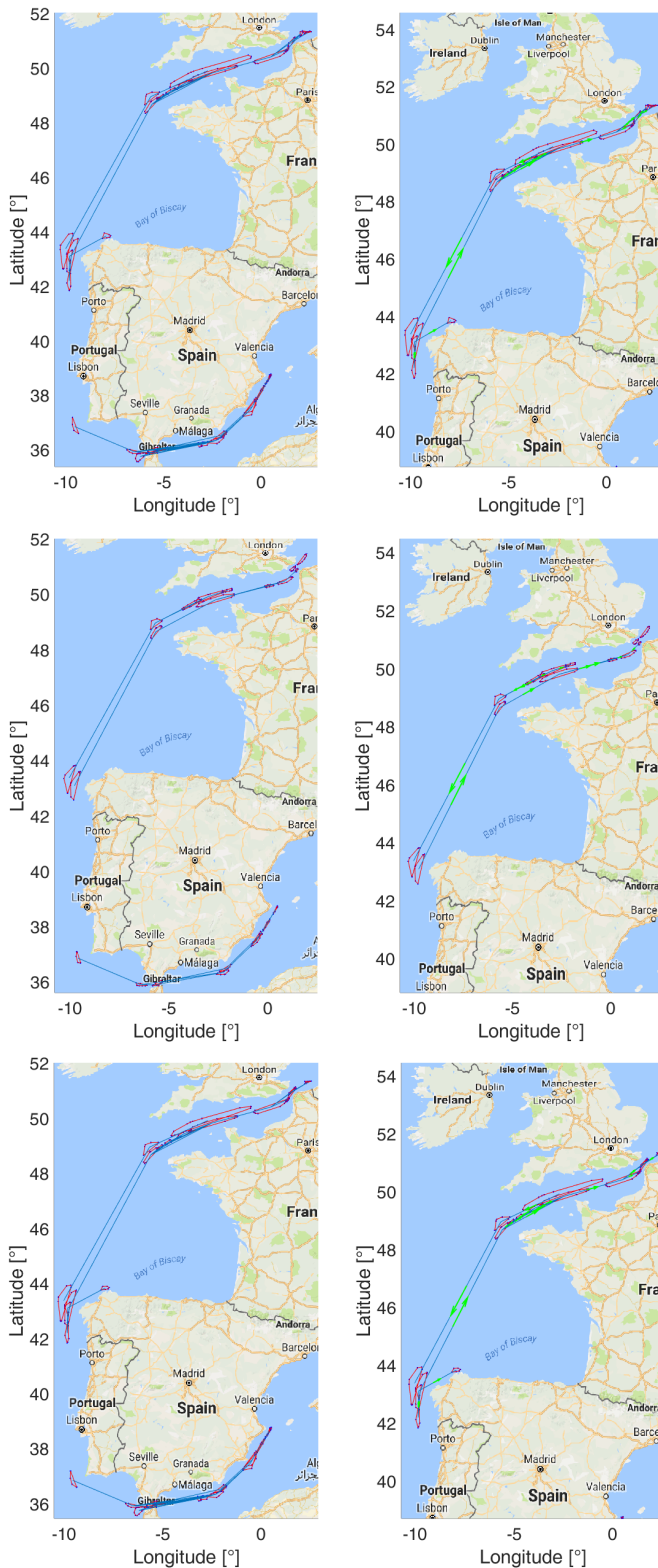


Fig. 14. Resulting networks using the second performance criterion for the three optimal DBSCAN parameters (from top to bottom). The main differences are visible along the English Channel where the size and the number of clusters are different. The green arrows indicate the ships' directions and demonstrates the effect of the Traffic Separation Scheme (TSS) which defines confined waterways for two principal directions. This approach appears unable to detect the sea lanes along the coast of the Portugal. The number of nodes and edges, for the three graphs, are equals to 36, 24, 36 and 44, 22, 50, respectively.

- [9] L. Snidaro, I. Visentini, K. Bryan, and G. L. Foresti, "Markov logic networks for context integration and situation assessment in maritime domain," in *2012 15th International Conference on Information Fusion*, July 2012, pp. 1534–1539.
- [10] V. F. Arguedas, G. Pallotta, and M. Vespe, "Unsupervised maritime pattern analysis to enhance contextual awareness," *Context-Awareness in Geographic Information Services (CAGIS 2014)*, p. 50, 2014.
- [11] B. Ristic, B. L. Scala, M. Morelande, and N. Gordon, "Statistical analysis of motion patterns in ais data: Anomaly detection and motion prediction," in *2008 11th International Conference on Information Fusion*, June 2008, pp. 1–7.
- [12] T. Fabbri, R. Vicen-Bueno, R. Grasso, G. Pallotta, L. M. Millefiori, and L. Cazzanti, "Optimization of surveillance vessel network planning in maritime command and control systems by fusing METOC and AIS vessel traffic information," in *OCEANS 2015 - Genova*, May 2015, pp. 1–7.
- [13] D. Makris and T. Ellis, "Path detection in video surveillance," *Image Vision Comput.*, vol. 20, no. 12, pp. 895–903, 2002.
- [14] L. Cazzanti, A. Davoli, and L. M. Millefiori, "Automated port traffic statistics: From raw data to visualisation," in *2016 IEEE International Conference on Big Data (Big Data)*, Dec 2016, pp. 1569–1573.
- [15] L. Cazzanti, L. M. Millefiori, and G. Arcieri, "A document-based data model for large scale computational maritime situational awareness," in *2015 IEEE International Conference on Big Data (Big Data)*, Oct 2015, pp. 1350–1356.
- [16] L. M. Millefiori, D. Zissis, L. Cazzanti, and G. Arcieri, "A distributed approach to estimating sea port operational regions from lots of AIS data," in *2016 IEEE International Conference on Big Data (Big Data)*, Dec 2016, pp. 1627–1632.
- [17] B. Liu, E. N. de Souza, S. Matwin, and M. Sydow, "Knowledge-based clustering of ship trajectories using density-based approach," in *2014 IEEE International Conference on Big Data (Big Data)*, Oct 2014, pp. 603–608.
- [18] N. L. Guillaume and X. Lerouvreur, "Unsupervised extraction of knowledge from S-AIS data for maritime situational awareness," in *Proceedings of the 16th International Conference on Information Fusion*, July 2013, pp. 2025–2032.
- [19] F. Castaldo, F. A. N. Palmieri, V. Bastani, L. Marcenaro, and C. Regazzoni, "Abnormal vessel behavior detection in port areas based on dynamic Bayesian networks," pp. 1–7, July 2014.
- [20] F. Castaldo, F. A. N. Palmieri, and C. S. Regazzoni, "Bayesian analysis of behaviors and interactions for situation awareness in transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 313–322, Feb 2016.
- [21] S. Coraluppi and C. Carthel, "Stability and stationarity in target kinematic modeling," in *IEEE Aerospace Conference*, March 2012, pp. 1–8.
- [22] —, "Advances in multi-target filtering of evasive targets," in *Aerospace Conference, 2015 IEEE*, March 2015, pp. 1–8.
- [23] O. E. Barndorff-Nielsen and N. Shephard, "Integrated OU processes and non-Gaussian OU-based stochastic volatility models," *Scandinavian Journal of Statistics*, vol. 30, no. 2, pp. 277–295, 2003.
- [24] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.
- [25] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.
- [26] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters 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*, ser. KDD'96. AAAI Press, 1996, pp. 226–231.
- [27] M. P. Dubuisson and A. K. Jain, "A modified Hausdorff distance for object matching," in *Proceedings of 12th International Conference on Pattern Recognition*, vol. 1, Oct 1994, pp. 566–568 vol.1.
- [28] T. U. Wien, T. Eiter, T. Eiter, H. Mannila, and H. Mannila, "Computing discrete Fréchet distance," *Tech. Rep.*, 1994.
- [29] P. C. Besse, B. Guillouet, J. M. Loubes, and F. Royer, "Review and perspective for distance-based clustering of vehicle trajectories," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 11, pp. 3306–3317, Nov 2016.



Pasquale Coscia earned the Laurea degree (*cum laude*) in Ingegneria Informatica in 2014 from Seconda Università degli Studi di Napoli (now Università degli Studi della Campania “Luigi Vanvitelli”), Aversa, Italy. In November 2015, he joined the PhD program in Industrial and Information Engineering at Seconda Università degli Studi di Napoli. His main research interests are in the areas of computer vision, machine learning, and probabilistic models applied to multi-agents systems.



Paolo Braca (M’14–SM’17) received the Laurea degree (*summa cum laude*) in electronic engineering, and the Ph.D. degree (highest rank) in information engineering from the University of Salerno, Italy, in 2006 and 2010, respectively. In 2009, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. In 2010–2011, he was a Postdoctoral Associate with the University of Salerno, Italy. In 2011, he joined the NATO Science & Technology Organization Centre for Maritime

Research and Experimentation (CMRE), and currently is a *Senior Scientist* with the Research Department. He conducts research in the general area of statistical signal processing with emphasis on detection and estimation theory, wireless sensor network, multi-agent algorithms, target tracking and data fusion, adaptation and learning over graphs, distributed radar (sonar) processing. He is coauthor of about one hundred publications in international scientific journals and conference proceedings. In 2017 Dr. Braca was awarded with *National Scientific Qualification* to function as associate professor in Italian Universities. Dr. Braca serves as Associate Editor of *IEEE Transactions on Signal Processing*, *IEEE Transactions on Aerospace and Electronic Systems*, *ISIF Journal of Advances in Information Fusion*, *EURASIP Journal on Advances in Signal Processing*. In 2017 he was Lead Guest Editor of the special issue “Sonar Multi-Sensor Applications and Techniques” in *IET Radar Sonar and Navigation*. He served as Associate Editor of *IEEE Signal Processing Magazine* (E-Newsletter) from 2014 to 2016. He is in the Technical Committee of the major international conferences in the field of signal processing and data fusion. He was the recipient of the Best Student Paper Award (first runner-up) at the *12th International Conference on Information Fusion* in 2009. He received the NATO STO Scientific Achievement Award 2017 for its contribution to “Development and Demonstration of Networked Autonomous ASW” from the NATO Chief Scientist.



Leonardo M. Millefiori (S’12–M’14) attended the Sapienza University of Rome, Italy, where he received his B.Sc. degree in Aerospace Information Engineering and his M.Sc. degree (*summa cum laude*) in Communication Engineering with focus on radar systems and remote sensing, in September 2010 and January 2013, respectively. From June to December 2013, he was Visiting Researcher at the NATO Science & Technology Organization Centre for Maritime Research and Experimentation (CMRE), that he joined in 2014 as Scientist with

the Research Department under the Maritime Security programme. His current research interests include target motion modeling, statistical signal processing, target tracking and data fusion and radar systems.



Francesco A. N. Palmieri (M’09) received the Laurea degree in Ingegneria Elettronica (*cum laude*) from Università degli Studi di Napoli Federico II, Naples, Italy, in October 1980, and the M.S. degree in applied sciences and the Ph.D. degree in electrical engineering from the University of Delaware, Newark, DE, USA, in 1985 and 1987, respectively. From October 1980 to March 1982, he served as a 2nd Lieutenant in the Italian Army in fulfillment of draft duties. In 1982 and 1983, he was with the ITT firms: FACE SUD Seletttronica in Salerno (currently Alcatel), Italy, and Bell Telephone Manufacturing Company in Antwerpen, Belgium, as a designer of digital telephone systems. In 1983, he was awarded a Fulbright scholarship to conduct graduate studies at the University of Delaware. He was appointed as an Assistant Professor in Electrical and Systems Engineering at the University of Connecticut, Storrs, CT, USA, in 1987, where he was awarded tenure and promotion to Associate Professor in 1993. In the same year, after a national competition, he was awarded the position of Professore Associato at the Dipartimento di Ingegneria Elettronica e delle Telecomunicazioni at Università degli Studi di Napoli Federico II, Italy, where he has been until October 2000. In February 2000, he was nominated Professore Ordinario di Telecomunicazioni after a national competition and appointed in November 2000 at Dipartimento di Ingegneria dell’Informazione, Seconda Università di Napoli (now Università degli Studi della Campania Luigi Vanvitelli), Aversa, Italy. He is also a Visiting Research Scholar at the University of Connecticut. His research interests include the areas of signal processing, communications, information theory, neural networks, Bayesian models, and graphical models.



Peter Willett (F’03) received the B.A.Sc. in engineering science from the University of Toronto, Toronto, Ontario, Canada, in 1982 and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 1986. He has been a faculty member at the University of Connecticut, Storrs, CT, USA, ever since, where since 1998, has been a Professor. He has published in the areas of change/abnormality detection, optical pattern recognition, communications and industrial/security condition monitoring. His research interests include

statistical signal processing, detection, machine learning, data fusion and tracking. Prof. Willett was editor-in-chief of *IEEE Signal Processing Letters* (2014–2016) and the *IEEE Transactions on Aerospace and Electronic Systems* (2006–2011), and was the Vice President for Publications for AESS (2012–2014). He was a member of the IEEE AESS Board of Governors (2003–2009 and 2011–2016). He was the General Co-Chair for the 2006 IEEE/ISIF Fusion Conference in Florence, Italy, and again for both 2008 in Cologne, Germany and 2011 in Chicago, IL, USA. He was the Program Co-Chair for the 1999 Fusion Conference in Sunnyvale, CA, USA, the 2003 IEEE Conference on Systems, Man and Cybernetics in Washington, DC, USA, and the 2016 ISIF/IEEE Fusion Conference in Heidelberg, Germany.

Fig. 1. - Density map of AIS messages collected from cargo ships during two months in the area of interest. Several waypoints, i.e. localized spatial regions where vessels show velocity or orientation changes are observable. Linear piecewise lanes that make up standardized routes followed by vessels are clearly visible. The color is proportional to the number of ships whose reported positions fall within the ground footprint, from green (low traffic density) to red (high traffic density).

Fig. 2 - Example of the change detection procedure applied to a vessel's track. On the left, the red circles are the points where a change is detected while the red arrow represents the starting point and the vessel's direction. On the right, the red dots represent the instantaneous values of the velocity along the x and y components and the track orientation while the circles the detected changes of the parameters. The parameter δ of the change detection procedure is set to 5° and $N = 15$.

Fig. 3 - Result of the change detection procedure along the Iberian Coast (on the left) and the English Channel (on the right) applied to the whole dataset. The red crosses represent the points where a vessel changes its direction of about $\pm 5^\circ$ (at least). The threshold h of the CUSUM algorithm is set to 4.

Fig. 4 - Synthetic example of the use of the Mahalanobis distance and the Euclidean distance as a measure distance for the DBSCAN algorithm. The blue arrows represent the direction before the change and the green arrows the direction after the change. The points A, B and C could be grouped in a cluster, say C_1^* , while the points D and E in a different cluster, say C_2^* , if the Euclidean distance is used (dashed lines). The consideration of the direction changes leads to a better result. The different direction changes are taken into account with the Mahalanobis distance metric resulting in a better partitioning of the data (solid lines).

Fig. 5 - Three clusters detected with the DBSCAN algorithm (on the left) and the corresponding estimated PDFs (on the right) of the input and output directions extracted as $\theta_{0,1} = \arctan(v_{y_{0,1}}, v_{x_{0,1}})$. The blue curves represent the result applying a Kernel Density Estimation (KDE) technique with Gaussian basis functions, while the red curves are the Gaussian approximations using the parameters estimated with the MLE technique. In a view of a compact representation, the distributions Θ_0 and Θ_1 could be therefore described by two synthetic parameters, μ and σ , due to the limited differences between the results of the two estimation approaches.

Fig. 6 - Some DBSCAN results along the English Channel. Three combinations of the parameters minPts and ϵ are considered: (4, 0.05), (25, 0.1) and (45, 0.18), respectively (from left to right). Different combinations of such parameters lead to a different number of clusters and hence of routes.

Fig. 7 - Synthetic representation of the sea lane-based performance criterion. A track $\mathcal{T}$ (dashed line) and three graph branches are shown. The green arrows represent the vessel directions and the traffic motion (on the branches). The points $\mathbf{p}_k$ and $\mathbf{p}_{k+1}$ are associated to the sea lane Sl_{12} since they fall within its surrounding region while points $\mathbf{p}_{k-1}$ and $\mathbf{p}_{k+2}$ are labeled as non-associated. The total percentage of associated points is reported in Table II.

Fig. 8 - Two navigational legs off the coast of the Iberian

peninsula and elliptic areas defined by three different values of eccentricity: 0.95, 0.97, 0.99. The foci for the sea lane Sl_{12} are $\mathbf{p}_{C_1}$ and $\mathbf{p}_{C_2}$ while, for the sea lane Sl_{23} , are $\mathbf{p}_{C_2}$ and $\mathbf{p}_{C_3}$ respectively. The length of the semi-minor axes for Sl_{12} are 37.13 km, 28.32 km and 16.10 km while for Sl_{23} are 58.27 km, 44.43 km and 25.26 km. The length of each sea lane is 222.58 km (Sl_{12}) and 291.94 km (Sl_{23}), respectively.

Fig. 9 - Graphical representation of the error $\mathcal{E}_G$ for three values of eccentricity e . The white points represent the combination of parameters (minPts, ϵ) such that it is not possible to create a network due to no clusters or sea lanes being detected.

Fig. 10 - Normalized histograms of the orthogonal distance $d_\perp(\mathbf{p}, Sl_{ij})$ between a point and the associated sea lane (first row) and the values $\Delta\theta$ which represent the distance in term of velocity direction w.r.t. the associated sea lane (second row) for the values of the eccentricity parameter.

Fig. 11 - Goodness $\mathcal{G}$: the maximum value needs to be considered in this case to obtain the optimal parameters. The white points represent the parameters for which it is not possible to construct a graph due to no clusters or sea lanes being detected.

Fig. 12 - Normalized histograms of the distance between the waypoints w_{p_i} and the clusters' centroids $\mathbf{p}_{C_i}$ and the absolute deviations $\Delta\theta = |\theta - \mu|$ for the input and output velocity directions. The rows represent the results varying the confidence interval $\mathcal{D}$.

Fig. 13 - Resulting network for the optimal values of minPts and ϵ using the first performance criterion. The first four pictures depict clusters and sea lanes for the analysed domain along with three detailed areas. On the bottom, two particulars of the Iberian Peninsula are reported where the traffic directions are highlighted with green arrows. The number of nodes and edges are equals to 602 and 706, respectively.

Fig. 14 - Resulting networks using the second performance criterion for the three optimal DBSCAN parameters (from top to bottom). The main differences are visible along the English Channel where the size and the number of clusters are different. The green arrows indicate the ships' directions and demonstrates the effect of the Traffic Separation Scheme (TSS) which defines confined waterways for two principal directions. This approach appears unable to detect the sea lanes along the coast of the Portugal. The number of nodes and edges, for the three graphs, are equals to 36, 24, 36 and 44, 22, 50, respectively.