Estimated Time of Arrival Using Historical Vessel Tracking Data

Alfredo Alessandrini, Fabio Mazzarella⬤, and Michele Vespe

Abstract—The growing availability of data coming from ship reporting systems, such as Automatic Identification System (AIS) and Long Range Identification and Tracking (LRIT), is originating an unprecedented set of opportunities to enforce maritime surveillance, ensure the security of the traffic at sea, and manage maritime operations. In this paper, a data-driven methodology is proposed to estimate the vessel times of arrival in port areas. The developed approach exploits both AIS and LRIT historical maritime traffic data collected over a desired area of interest and is based on an optimized data-driven path-finding algorithm. The methodology is applied and validated to real scenarios with real data sets, showing how a list of times of arrival can be automatically computed for predefined ports and progressively refined. Such information is expected to increase port operational efficiency and safety.

Index Terms—Estimated time of arrival, maritime situational awareness, data-driven path-finding, vessel tracking data, port operations.

I. INTRODUCTION

THE understanding of maritime traffic has become an increasingly important topic of research due to the wide range of activities of vessels at sea and for its direct implications on security, safety, environmental and socio-economic factors. On the one hand, with over 80% of world merchandise trade by volume being carried by sea, maritime transport remains the backbone supporting international trade and globalisation. On the other hand, every year, 400 million passengers embark and disembark in European ports, which represent vital gateways linking transport corridors all around the world [1]. In addition, ports play a key role not only in moving goods but also in generating employment, with 1.5 million workers employed in European ports and the same amount employed indirectly across the EU maritime State Members [2]. As a consequence, maritime transport and ports demand greater monitoring and understanding (e.g. for safety, security and efficiency.)

The growing number of ship reporting technologies and remote sensing systems (e.g. Automatic Identification System (AIS), Long Range Identification and Tracking (LRIT), radar tracking, Earth Observation) is providing a wealth of spatiotemporal and geographically distributed vessel positioning data.

Amongst ship reporting systems, AIS technology [3], originally conceived for collision avoidance, is becoming a keystone of Maritime Situational Awareness (MSA). AIS is mandatory for ships of 300 gross tons and upwards in international voyages, 500 tons and upwards for cargoes not in international waters, passenger vessels [4] and EU fishing vessels of overall length exceeding 15 meters [5]. AIS messages can be received also from space and a growing number of Low Earth Orbit (LEO) satellite based data providers have emerged [6], [7]. The information content of AIS messages, besides the vessel state vector and other kinematic information (e.g. latitude, longitude, speed, course), includes voyage-related (e.g. destination, Estimated Time of Arrival (ETA), etc.) and static (e.g. size, ship type, etc.) information about the ship. The continuous increase of terrestrial networks and satellite constellation of receivers is providing global tracking data well-suited to be used in a wide range of applications beyond collision avoidance. Example applications include vessel tracking [8]–[10], knowledge discovery [11], vessel behaviour identification [12], [13], anomaly detection [14]–[17].

Based on satellite communications and introduced for safety and security applications, LRIT [18] is mandatory for all passenger ships, high speed craft, mobile offshore drilling units and cargo ships of 300 gross tonnage and upwards [19]. These vessels must report their positions (latitude and longitude) and identities to their flag administration at least four times a day. Being LRIT messages sent only to the competent authorities, special permissions are required to obtain and use such information.

Despite AIS and LRIT technologies were born for different purposes, and the data sharing philosophies are very different (peer to peer for AIS and centralized for LRIT), their complementary nature is highlighted by their technical specifications and individual tracking capabilities [20], [21]. These include temporal resolution or refresh rate; type, flag and amount of vessels regulated; spatial coverage and data latency. AIS provides better *temporal resolution* (higher refresh rate) and *vessel fleet coverage*, while LRIT provides continuous *spatial coverage* and greater *reliability* for security applications. Although satellite AIS could guarantee global coverage, in high traffic density areas the message transmissions in the same frequency range increase the probability of message collisions.

This paper focuses on the automatic estimation of vessel times of arrival in port areas in order to improve monitoring activities and eventually enhance the port operations efficiency and safety. Although vessel ETA is available in the voyage-related part of AIS messages, it is often unreliable because it is manually input. Nevertheless, in the EU, operators, agents or masters of ships have to notify authorities with information about their time of arrival in port areas 72 hours before their arrival [22]. The estimation of accurate vessel times of arrival can be therefore considered a challenging problem in the maritime domain for several reasons. First, the maritime authorities need to double-check the trustworthiness of the declared notifications. Second, the notice period of 72 hours defined in [22] is sometimes too long to make accurate predictions and avoid the occurrence of undeclared or possibly undesired operations in port areas. Finally, maritime authorities would like to know the volume of vessels approaching their ports at a desired time in order to make port operations more effective.

The accurate and timely estimation of times of arrival is a well-studied topic in literature for a wide range of application areas. They include air traffic control arrival sequencing and scheduling [23], [24], airport gate assignment methods [25], [26], emergency services [27], in-car navigation systems, elevator control [28], vehicles traffic management [29]–[31], cargo-tracking systems [32]. In regard to the maritime domain, there is an extended bibliography dealing with the ship routing and scheduling issues. An exhaustive review of the research on these topics and related problems can be found in [33]. At our state of the knowledge, literature lacks of studies on the estimation of vessel times of arrival in port areas by exploiting historical vessel tracking data.

Differently from other transportation domains (e.g. car traffic, public transport, air traffic etc.) where roads and corridors strictly limit manoeuvrability, vessels movements are indeed regulated only in highly congested areas. Ship route optimisation approaches are often based on shortest-path, land-avoidance and meteo-oceanographic conditions based methodologies [34]. Nevertheless, the mix of both regulated and unregulated sea traffic makes historical movements at sea a fundamental prior knowledge to estimate and predict travel and arrival times.

This paper presents a novel methodology for estimating the vessel times of arrival in port areas by exploiting historical ship reporting systems data. The proposed approach is based on a data-driven path-finding algorithm, depending on a set of parameters conveniently optimised for the area under investigation. Path-finding between two geographical positions in a dynamic environment is a challenging and non-straightforward problem for maritime navigation. Maritime traffic is constrained by ship routing systems such as traffic separation schemes. In open sea, it can change dynamically with environmental factors (tide, current and wind conditions). All these aspects need to be taken into account when designing accurate route prediction tools and therefore automatically estimate times of arrival. One way that also factors in the seasonal variability of such information and deal with the difficulty in accessing such contextual information is a data driven methodology that exploits
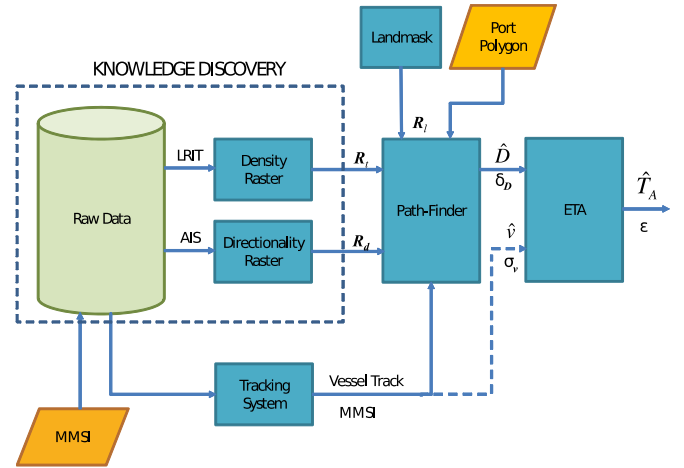

Fig. 1. General architecture for the estimation of vessel times of arrival in port areas. The system is fed with a list of MMSI numbers, identifying the vessels whose times of arrival need to be estimated, and a geo-referenced polygon of a defined port area. The output of the system are the estimated time of arrival, $\hat{T}_A$, and the associated uncertainty, ϵ.

historical real tracking data [35]. The architecture discussed in this paper is designed to benefit from the complementary characters of AIS and LRIT technologies. The joint use of historical AIS and LRIT data helps the search of an optimal route between two locations and thus the estimation of times of arrival in desired port areas.

The remainder of the paper is organised as follows. Section II provides an overview of the general approach followed in estimating the vessel times of arrival in port areas. Section III shows the application of the algorithm to real data and proves its reliability through validation. Section IV draws some conclusions and highlights future work.

II. METHODOLOGY OVERVIEW

The flowchart of the proposed methodology for the estimation of vessel times of arrival in port areas is depicted in Fig. 1. Input to the architecture are a set of $MMSI$ numbers,[1] identifying the ship tracks whose times of arrival have to be estimated, and a geo-referenced polygon of the port area of interest. It is worth noting that, in general, this step could be implemented through a data-driven approach by estimating the port area from ship reporting information (e.g. LRIT or, more accurately, AIS). The output of the system is the estimated time of arrival $\hat{T}_A$, and the associated uncertainty ϵ, for every input vessel MMSI.

The overall algorithm relies on two fundamental stages: the extraction of knowledge from historical vessel tracking data and the knowledge-based path-finding. In the former, AIS and LRIT historical vessel positioning data are selected and pre-processed in the direction of building a set of rasters (taking into account density and directionality of historical AIS messages) that will represent the data-driven baseline of the proposed architecture. In the latter, the path-finding is accomplished in order to obtain a route that connects a

---

[1] A Maritime Mobility Service Identity (MMSI) number is a nine-digits ship identifier as defined in AIS standard [3].
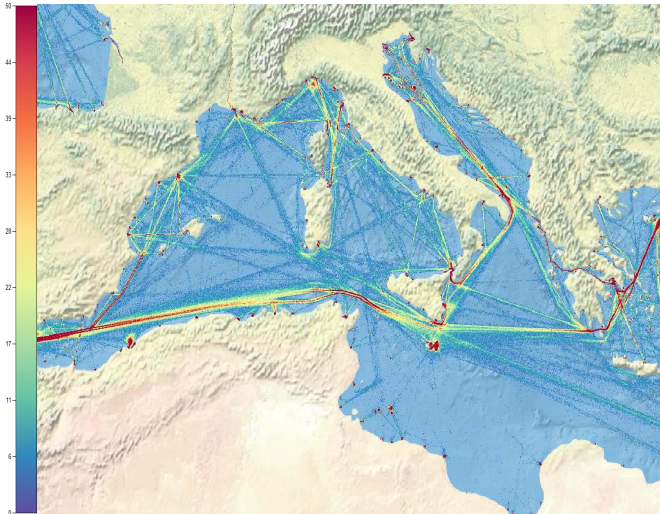
Fig. 2. Traffic density in the Mediterranean Sea, obtained by gridding historical 5-years LRIT data. The picture shows how LRIT data covers the scene and displays maritime lanes and traffic separation schemes. The traffic density information are fundamentals in the estimation of vessel times of arrival in port areas.

source location to a defined destination, while minimising a cost function. Vessel times of arrival in the defined port area, for the required MMSI numbers, are finally estimated by using the output of the path-finding process and a set of vessel dynamic features extracted by a tracking system.

A. Knowledge Discovery

Positioning data from ship reporting systems, such as AIS and LRIT, reveals vessels behaviour and motion patterns. Despite the oceans surface provides an enormous playground for maritime traffic, navigation is typically constrained to certain areas, delineated by economic, legal, logistic, security or safety reasons. Maritime traffic analysis focuses on the study of vessel historical positioning data to detect vessel motion patterns, and their exploitation to infer different levels of contextual information.

The knowledge of expected routes represents the core of the estimation of the vessel route to the destination. The *knowledge discovery* process in Fig. 1 consists of two different stages aiming at extracting density and directionality *rasters*, by exploiting both LRIT and AIS data, respectively. In geospatial analysis, rasters are rectangular, regular grids that represent continuous data over geographical space. Cells are arranged in rows and columns and each holds a value. A raster is accompanied by meta-data that indicate the resolution, extent and other properties.

Traffic density information is extracted by gridding historical LRIT positioning data, given that they are not biased by spatial coverage performance (see Sec. I). In particular, this process produces a geo-referenced raster $\mathbf{R}_t$, whose elements represent the number of vessels crossing the corresponding cells in a defined time interval. The values in the raster are scaled in such a way that ports and other stop areas do not play as attraction points for the path-finding algorithm. The traffic density in the Mediterranean Sea, calculated on 5-years of

LRIT data, is depicted in Fig. 2. The picture shows how LRIT data are capable to cover the scene under analysis and how the extracted traffic density can highlight maritime lanes and traffic separation schemes. The information provided by the LRIT-based traffic density plays a key role in the estimation of the optimal route connecting two geographical locations. Nevertheless, LRIT does not provide direction information and the 6-hour refresh rate does not allow for an accurate estimation of Course Over Ground (COG) between consecutive messages [36]. On the contrary, AIS messages contain direction information and for this reason historical AIS tracking data are used to extract a *directionality* layer. In particular, this step produces a geo-referenced raster $\mathbf{R}_d$, where every cell contains the median value of COGs of the vessels crossing that cell in a certain time interval.

Algorithm 1 Path Finding & ETA Estimation

Require: Port $\mathbf{x}_p$, $\{MMSI_i\}_{i=1}^N$, $\{\mathbf{S}_i\}_{i=1}^N$, $\mathbf{R}_t$, $\mathbf{R}_d$, $\mathbf{R}_l$, w_t, w_d, w_l

1: // Rasters to graphs conversion:
  $\mathbf{G}_t \leftarrow$ transition($\mathbf{R}_t$, *mean*)
  $\mathbf{G}_d \leftarrow$ transition($\mathbf{R}_d$, *mean*)
  $\mathbf{G}_l \leftarrow$ transition($\mathbf{R}_l$, *mean*)
2: // Graphs geocorrection:
  $\mathbf{T}_t \leftarrow$ geocorrection($\mathbf{R}_t$)
  $\mathbf{T}_d \leftarrow$ geocorrection($\mathbf{R}_d$)
  $\mathbf{T}_l \leftarrow$ geocorrection($\mathbf{R}_l$)
3: // Weighted geocorrected graph:
  $\mathbf{G} = w_t \mathbf{T}_t + w_d \mathbf{T}_d + w_l \mathbf{T}_l$
4: **for** $i = 1$ to N **do**
5:   // Extraction of newest state variables from track i
    $[\mathbf{x}_{i,S}, time_{i,S}] \leftarrow last\_obs(\mathbf{S}_i)$
6:   // Optimum path between last observation and port
    $[\mathbf{P}_i, \hat{D}_i] \leftarrow Dijkstra(\mathbf{G}, \mathbf{x}_{i,S}, \mathbf{x}_p)$
7:   // Speed profile for vessel i
    $sog_i \leftarrow speed\_profile(\mathbf{S}_i)$
8:   // Speed expected value from Gaussian fitting on speed profile
    $[\hat{v}_i, \sigma_{v,i}] \leftarrow gauss\_fitting(sog_i)$
9:   // ETA estimation
    $\hat{T}_{A,i} = time_{i,S} + \frac{\hat{D}_i}{\hat{v}_i}$
10: **end for**
11: // Return the list of times of arrival
12: **return** $\{\hat{T}_{A,i}\}_{i=1}^N$, $\{\hat{D}_i\}_{i=1}^N$

Both traffic density and directionality layers are clearly less necessary in areas where traffic is neither regulated nor constrained by, *e.g.*, the bathymetry and therefore vessels navigate along great circles.

B. Tracking System

AIS and, if available, LRIT messages for a desired Area Of Interest (AOI) and over time interval of interest are ingested and integrated in order to identify and manage separate ship tracks. The tracking system in Fig. 1 provides thus a set of *vessel tracks*, namely spatio-temporal trajectories $\{\mathbf{S}_i\}_{i=1}^N$, containing dynamic (latitude, longitude, Course Over

Ground (COG), Speed Over Ground (SOG)) and static/voyage-related (MMSI, name, destination, ship type) features for every single reporting ship. The input data are also cleaned against outliers and decimated according to a defined interval as described in [9]. The percentage of outliers varies depending on the quality of the AIS transmitter and GPS device onboard, usually resulting in wrong positions. Such errors can be filtered out by gating: if the velocity estimated between two consecutive AIS positions is beyond a plausible threshold, then the latest position can be considered unreliable and discarded as outlier. The tracking system can operate both in real-time and off-line. Given the availability of vessel tracks, it is possible to query the tracking system every time in order to get details for all those vessels contained in the ship pre-arrival notice list $\{MMSI_i\}_{i=1}^{N}$ of the port of interest and representing one of the input set of the architecture in Fig. 1.
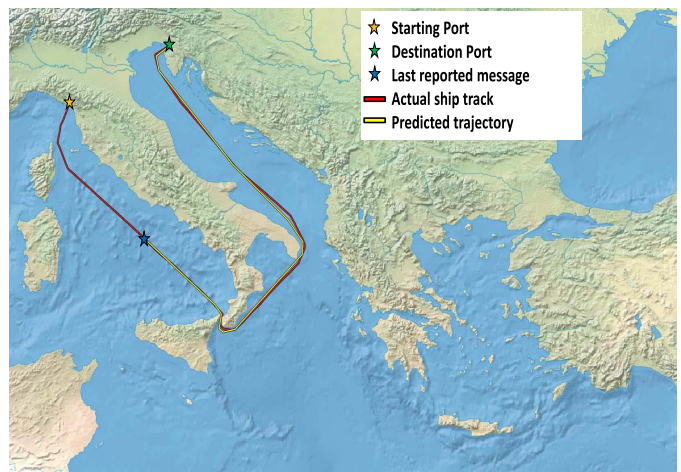
C. Path-Finding Algorithm

The main goal of the path-finding algorithm is to come up with an optimal route between two geographical locations, while minimising a cost function. The most commonly used geographic distance measure is the great-circle distance, which represents the shortest line between two points, taking into account the curvature of the earth. This distance could be conceived of as the distance measured along a route of a very efficient traveller who knows where to go and has no obstacles to deal with. As described in the introductory part, in a dynamic environment like the sea, the traffic is strongly constrained by several factors and a least-cost criterion should be adopted. In particular, the approach proposed in this paper relies on the grid-based solution in [37] and the algorithm is summarised in Algorithm 1.

The calculations starts with raster data. As depicted in Fig. 1, the path-finder process relies on the traffic density and directionality rasters, $\mathbf{R}_t$ and $\mathbf{R}_d$, respectively, extracted from historical raw data as described in Sec. II-A. In addition, a land-masking layer $\mathbf{R}_l$ is used to avoid the paths to cross the coastline. Distance and route calculations on rasters are all based on graph theory. The first step of the algorithm consists thus of converting input rasters into graphs $\mathbf{G}_t$, $\mathbf{G}_d$ and $\mathbf{G}_l$. This is achieved by connecting cell centres to each other, which become the nodes in the graphs. More in detail, cells are connected with their eight orthogonal and diagonal nearest neighbours. During this conversion, weights are given to each *edge*, namely the connection between nodes. Weights correspond to different concepts, and in the approach in [37] they represent *conductance* values (1/resistance). The obtained graphs can then be mathematically represented as conductance matrices, whose values are evaluated according to a user-defined *transition function* $f(i, j)$ to get the transition value from cell i to cell j. In the Algorithm 1, the cell *mean value* is used as transition function for the whole set of rasters.

In the input rasters, diagonal neighbours are more remote from each other than orthogonal neighbours. In addition, on longitude-latitude grids, West-East connections are longer at the Equator and become shorter at the Poles. Therefore, values in the conductance matrices need to be corrected and this is accomplished by the function *geocorrection* in
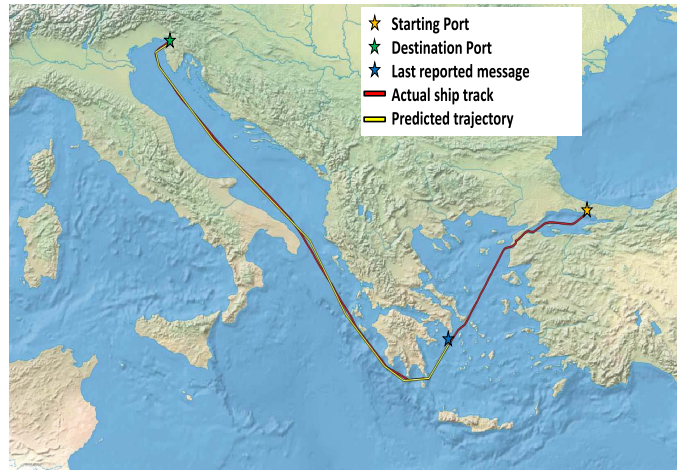
(a)

(b)

Fig. 3. Outcome of the proposed path-finding algorithm for a single vessel (a) starting from the port of Istanbul (yellow star) and going to the port of Trieste (green star) between May 20 and May 23 2016, and (b) starting from the port of La Spezia (yellow star) and going to the port of Trieste (green star) between April 20 and April 23 2016. For both panels, the red trajectory is the ship track reported by the AIS messages. The path-finding algorithm is simulated supposing that the last available AIS report is the one reported by the blue star. The yellow track represents the best path found by the proposed algorithm between blue and green stars, whose time distance is 48 hours.

Algorithm 1, producing the geo-corrected matrices $\mathbf{T}_t$, $\mathbf{T}_d$ and $\mathbf{T}_l$. At this point, a linear combination of geo-corrected conductance matrices is calculated, such that the path-finder takes into account the whole set of constraints in finding the best route between two points. The coefficients w_t, w_d and w_l are provided by an optimisation process that will be explained later.

For every vessel identified by the input list $\{MMSI_i\}_{i=1}^{N}$, the tracking system can be queried in order to extract the corresponding track, $\mathbf{S}_i$, over a desired time interval and isolate the last available vessel position, $\mathbf{x}_{i,S}$, and its timestamp, $time_{i,S}$. The best path between the last reported vessel position and the port location $\mathbf{x}_p$, is then calculated by exploiting the well-known Dijkstra algorithm [38] constrained by the conductance matrix $\mathbf{G}$. The output is made of the best path $\mathbf{P}_i$
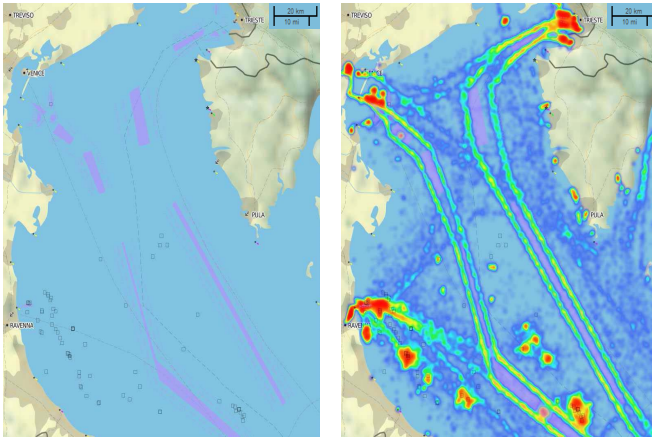
Fig. 4. Ships routing systems (left) and vessel traffic density (right) in the North Adriatic Sea close to the destination port analysed in this study (Trieste).
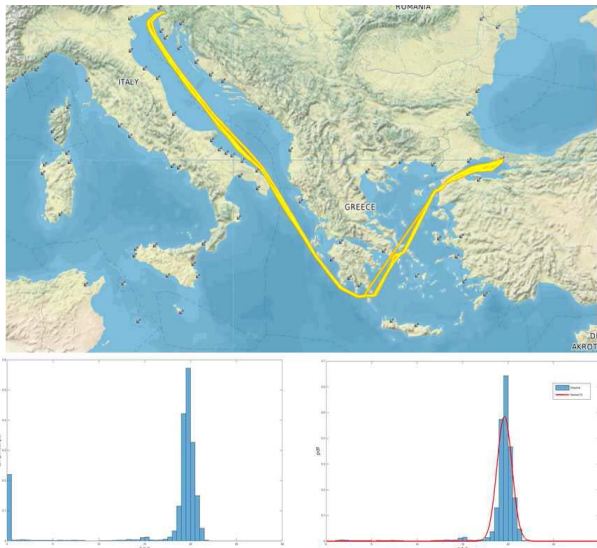

Fig. 5. The upper plot depicts the route experienced by a cargo vessel moving between ports of Istanbul, Turkey and Trieste, Italy, from February to May 2016. Lower left plot represents the histogram of the Speed Over Ground (SOG) reported by the vessel for the time interval considered. On the right lower plot, the SOG values related to port areas have been removed and the red curve is the result of a Gaussian fitting of SOG values out of ports. The expected value of the obtained normal distribution is used for the estimation of time of arrival, while its standard deviation is representative of the estimate uncertainty.

and the geographic distance between source and destination via the found trajectory, $\hat{D}_i$.

The path-finding process output is thus the route between the source point and the destination port, together with the corresponding distance. Examples of application of the path-finding algorithm are depicted in Fig. 3 for two different scenarios. A basic land-avoidance path-finding algorithm would perform poorly in such area, given the high level of routing schemes, as can be seen in Fig. 4, which depicts ships' routing systems and vessel traffic density in the North Adriatic Sea close to the destination port analysed in this study. In Fig. 3a, a cargo vessel going from the port of Istanbul to the port of Trieste is considered. The proposed algorithm is applied between the starting point represented by the blue star

and the port of Trieste. The traffic density in Fig. 2 highlights that the route between La Spezia and Sicily is not affected by prior knowledge as much as the route along the Adriatic sea, where traffic separation schemes appears clearly and a greater gain in the prediction performance is expected. Fig. 3b depicts instead a cargo covering the route between ports of La Spezia and Trieste. For both figures, the yellow tracks, which represent the 48-hours predicted trajectories, show a significant match to the actual trajectories, in red.

1) Optimisation: The optimisation process, aimed at finding the coefficients w_t, w_d and w_l, can be summarised as follows. Given a set of training tracks $\mathcal{S} = \{\mathbf{S}_i\}_{i=1}^L$ for a certain AOI, the path-finding procedure in Algorithm 1 is applied to each trajectory while evaluating the averaged Hausdorff distance [39] between the true track and the estimated ones, $\mathcal{P} = \{\mathbf{P}_i\}_{i=1}^L$. This averaged metric is used as the objective function for the coefficients optimisation step. Formally, the optimisation problem can be stated as:

$$\min_{w_t, w_d, w_l} [d_H(\mathcal{S}, \mathcal{P})],$$
$$\text{subject to } w_t + w_d + w_l = 1,$$
$$w_t > 0,$$
$$w_d > 0,$$
$$w_l > 0. \tag{1}$$

The solution of this problem allows to use the best coefficients for the area under analysis, while minimising the objective function in (1). The values of coefficients will clearly reflect the complexity of the traffic schemes for a certain area. In particular, in areas with a low level of routing schemes, the coefficient w_t will be negligible with the respect to w_d and w_l.

D. Estimation of Vessel Times of Arrival

Once the route and the distance between the geographical locations of the source and the port are known, the estimation of the time of arrival is calculated by exploiting speed information on the vessel under test made available from the tracking system. As summarised in Algorithm 1, the most straightforward way to do this is considering, for each vessel, the expected reported vessel cruise speed, say $\hat{v}_i$, and evaluate the time employed by the vessel to cover the estimated distance $\hat{D}_i$ between source and port locations as $\hat{T}_i = \hat{D}_i / \hat{v}_i$. The estimated time of arrival in port of the vessel under investigation, $\hat{T}_{A,i}$, is thus obtained by summing $\hat{T}_i$ up to the vessel reported time at the source location, $time_{i,s}$. An example of the procedure to estimate the vessel cruise velocity is shown in Fig. 5, where the upper plot depicts the behaviour of a cargo vessel tracked for three months, from February to May 2016 by using AIS data. In particular, the ship makes repeated trips between the ports of Istanbul and Trieste. The vessel speed profile is represented in the lower left panel of Fig. 5, where the highly frequent occurrences of low SOG values (e.g. less than 1 knot) are associated to the port areas. The vessel cruise velocity can be estimated by removing those values and then fitting a Gaussian distribution to the SOG samples, as appears from the lower right plot

of Fig. 5. The expected value of this distribution is the vessel speed, $\hat{v}$, used in the estimation of time arrival in port areas.

The performance of the estimated times of arrival can be assessed by computing its maximum absolute error, ϵ_{max}, with the respect to the available actual values, namely

$$
\begin{aligned}
\epsilon_{max,i} &= |T_{A,i} - \hat{T}_{A,i}| \\
&= \left| time_{i,s} + \frac{D_i}{v_i} - time_{i,s} - \frac{\hat{D}_i}{\hat{v}_i} \right| \\
&= \left| \frac{D_i}{v_i} - \frac{D_i - \delta_{D,i}}{v_i + \sigma_{v_i}} \right| \\
&= \frac{\delta_{D,i}}{v_i + \sigma_{v,i}} + \frac{D_i}{v_i} \frac{\sigma_{v,i}}{v_i + \sigma_{v,i}}.
\end{aligned}
\tag{2}
$$

The maximum absolute error $\epsilon_{max,i}$ of the estimated time of arrival is strictly linked to the actual distance to destination D_i, the uncertainty of its estimation, $\delta_{D,i}$, and the standard deviation of the estimated vessel cruise velocity, $\sigma_{v,i}$. The term $\delta_{D,i}$, quantifying the error in estimating the distance between origin and destination, can be approximated as the minimum value achieved by the objective function in (1).

III. EXPERIMENTS

A. Data Set Description and Experimental Setup

The proposed algorithm for the estimation of the vessel times of arrival in port areas has been evaluated by considering the specific port of Trieste, an Italian city on the Northeastern Adriatic Sea. As regards the data-driven part of the architecture in Fig. 1, the directionality raster $\mathbf{R}_d$ is extracted from one month of terrestrial AIS data related to October, 2015. The traffic density information $\mathbf{R}_t$, instead, is extracted by gridding five years of historical LRIT data (from 2009 to 2014). The AIS and LRIT-based rasters have both a cell resolution of 0.02 degrees × 0.02 degrees.

The resolution of the density and transition layers has been chosen by taking into account different issues. First, the raster cell represents a summary picture of the real vessel traffic behaviour (intensity, direction, speed,..). That information summary is thus affected by the cell size, since a too small cell can describe only a single trajectory event, and accordingly a too large cell is not informative enough (in representing too many trajectories having different behaviour). We found out that the 0.02 degrees resolution is an empirical compromise that allows to maximise the information retrievable from the historical vessel traffic. In addition, since we extracted the network from the raster layers, a too small cell size cannot be used, since the computational and storage capacity could be a significant constraint factor.

As described in Sec. II, the input of the system is a list of *MMSI* numbers, that can be retrieved from the pre-arrival notice list available at the port of Trieste for a desired time interval. Since at this stage we aim at evaluating the performance of the proposed architecture, the following experimental setup is considered. A list of vessels entering the port of Trieste during the whole month of April 2016 is obtained from the historical AIS messages. In order to assess the accuracy of the Algorithm 1, for each of the
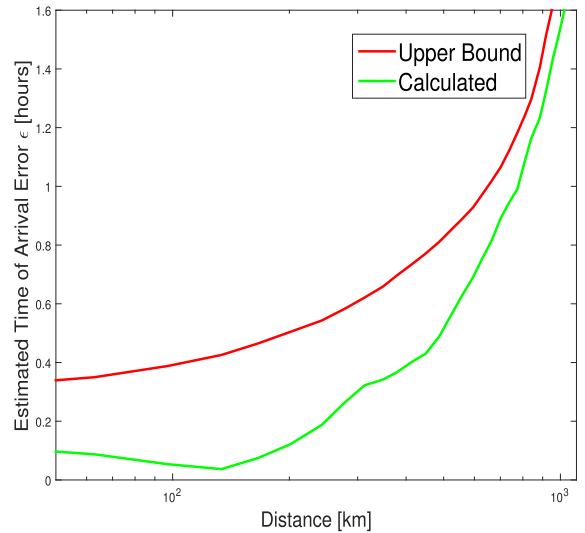

Fig. 6.   Behaviour of the absolute error on the estimated time of arrival estimation error versus the distance between the source and destination locations, for a single track chosen amongst the data set described in Sec. III-A. The green line is the absolute error obtained experimentally, while the red line represents the upper bound in (2).

identified vessels only those trips approaching Trieste from other locations without any intermediate stops are retained. The times reported from AIS messages when the vessels enter the geo-referenced port polygon for the first time with a declared speed lower than 0.5 knots, are taken as *ground truth* times of arrival in port, T_{GT}. The coefficients w_t, w_d and w_l have been calculated by using a numerical constrained global optimisation over a set of trajectories related to the area [10° E 24° E; 35° N 46° N] during the month of May, 2016.

The Algorithm 1 has thus available all the required input: the list of MMSI numbers and the corresponding trajectories, the rasters, the port location and the coefficients w_t, w_d and w_l.

B. Results

The performance of the algorithm is evaluated through the following conceptual experiment, iterated over all the vessels of interest, identified as described in Sec. III-A. Given a single vessel track, the path-finding algorithm is applied by keeping the destination point fixed on port location and changing the source point with the observations of the vessel track. For each point of the track, the time of arrival can be estimated as described in Sec. II and an absolute error, ϵ, between the estimated time of arrival and the ground truth can be computed. Accordingly, for every single vessel we get a set of errors together with the corresponding time differences between the source point and the ground truth time of arrival, $\{\epsilon_1, \Delta_{t,1}, \ldots, \epsilon_N, \Delta_{t,N}\}$, where N is the number of points considered in a ship track and $\Delta_{t,i}$, with $i = 1, \ldots, N$, is the time distance between the reporting time of the i-th source point and the ground truth time of arrival in the port of interest.

Fig. 6 depicts the behaviour of the absolute error ϵ versus the geographical distance between the source and destination locations, for a single track amongst those under analysis. In particular, green line represents the absolute error obtained experimentally, while the red line is obtained by using the upper bound in (2). The absolute error thus grows with the
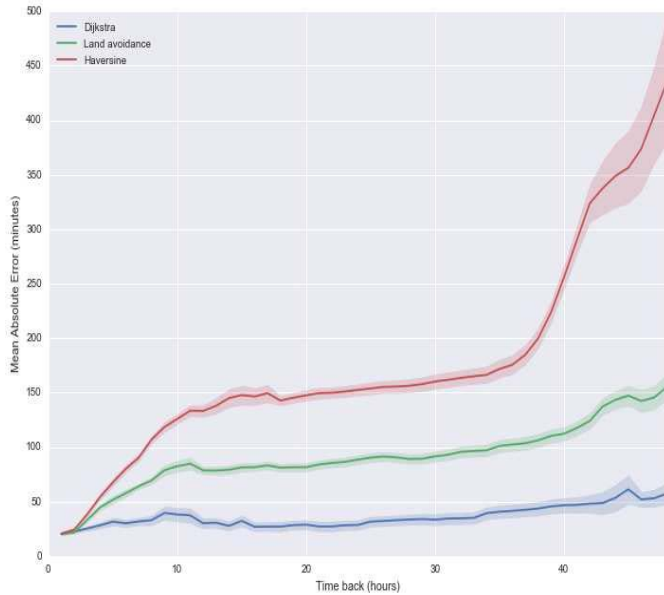
Fig. 7. Absolute time of arrival estimation error averaged over all the ship tracks entering the port of Trieste during the month of April, 2016. The behaviour of the averaged absolute error ϵ is depicted as the time distances Δ_t change between 1 and 48 hours, by using the proposed path-finding algorithm (blue), the land avoidance strategy (green), and the Haversine approximation for distances (red).

distance with a rate strongly affected by the accuracy of the cruise velocity estimation and the performance of the path-finding strategy, through δ_D and σ_v, respectively. In particular, we approximate the former with the Hausdorff distance between the actual trajectory and the one predicted through the proposed path-finder. The latter, instead, is approximated with the standard deviation of the Gaussian distribution fitting the SOG characterising the trajectory under observation.

The proposed algorithm is compared against two different methodologies. The first one can be considered as a benchmark and uses the well-known Haversine formula to approximate the great circle distance between source location and port. The second one is a land avoidance strategy, where the Algorithm 1 is applied withe $w_t = w_d = 0$ and $w_l = 1$. The absolute error ϵ averaged over the whole set of tracks, versus the time distances Δ_t, is depicted in Fig. 7, where the blue plot refers to the proposed algorithm in Fig. 1, the red plot refers to the Haversine approximation, and the green plot is the outcome of the land avoidance strategy. Time distances Δ_t changes between 1 and 48 hours and the results displayed in Fig. 7 prove that the presented approach greatly outperforms the other two methodologies. In particular, the expected gain with the respect to the Haversine baseline strategy in the mean absolute error is around 6 hours for estimations made 48 hours before the arrival in port. On the other hand, the gain with the respect to the land avoidance strategy is around 100 minutes for 48 hours estimations.

The performance in Fig. 7 highlight three different regions in the mean absolute error behaviour. Within the first one, for predictions made until to 10 hours before the time of arrival in port, the gain of Dijkstra strategy with the respect to the other two increases with the time distance because of the strong influence of the traffic patterns on the vessel track

prediction, as already observed in [40]. For time intervals between 10 and 38 hours, instead, the behaviours of the three different strategies are quite similar. This region corresponds to the Adriatic Sea traffic where, as highlighted in Fig. 2, routes are straight and the use of a data-driven strategy does not improve the prediction performance. In the last region, especially for time intervals greater than 40 hours, the non linear trend of the mean absolute error of the great circles baseline shows how the influence of the routing systems on track prediction is again fundamental in providing an improvement in terms of estimation error.

The curves in Fig. 7 show some bumps when the time interval Δ_t is between 10 and 20 hours and after 40 hours. As described in the above, the performance analysis is conducted on a set of tracks going towards the port of Trieste from another port with no intermediate stops. The effects in Fig. 7 could be ascribed to few outlier vessel tracks that experience one or more intermediate stops before arriving in the port of destination. For those trajectories, the error on the estimated time of arrival can increase even though the path-finding is accurate.

IV. CONCLUSIONS

In this paper, a novel methodology has been proposed to estimate the vessel times of arrival in port areas. The presented approach is based on a data-driven path-finding algorithm exploiting historical ship reporting systems data. In particular, both historical AIS and LRIT maritime traffic data over a desired area of interest have been used. The methodology has been applied to a real scenario with real data sets, and has been compared to other two strategies: the first one is based on the well-known Haversine formula to approximate the great circle distance between two geographical positions; the other one connects the source to the destination by avoiding the land. The experimental results show that making use of the data-driven algorithm allows the system to achieve higher accuracy in terms of time of arrival estimation error. In addition, the performance gain obtained by taking into account the knowledge extracted from historical data reveals to be greater where the traffic is more structured.

Despite the encouraging performance achieved by the proposed algorithm, it is constrained by the knowledge of the historical vessel speed information. In particular, the more accurate the estimate of the vessel speed is, the more accurate the estimation of time of arrival will be. The exploitation of traffic density and other meta-information, which are not individual vessel dependent, to estimate the vessel speed is a topic which could be looked at in future works.

The promising results presented in this paper can be exploited to build a system that automatically computes and progressively refines a list of times of arrival for predefined ports. The information provided by such a system can be used in order to improve monitoring activities and eventually enhance the efficiency of port operations.

REFERENCES

[1] *Ports: An Engine for Growth. COM/2013/0295*, Eur. Commission, Brussels, Belgium, 2013.

[2] *Europe's Seaports 2030: Challenges Ahead. MEMO/13/448*, European Commission, Brussels, Belgium, 2013.

[3] *Technical Characteristics for an Automatic Identification System Using Time-Division Multiple Access in the VHF Maritime Mobile Band*, Int. Telecommun. Union, Geneva, Switzerland, Apr. 2010.

[4] *Safety of Life at Sea Convention Chapter V, Regulation 19*, SOLAS, London, U.K., 2000.

[5] *Directive 2002/59/EC Establishing a Community Vessel Traffic Monitoring and Information System, as Emended by Directive 2009/17/EC and Commission Directive 2011/15 EU*, Eur. Parliament Council, Brussels, Belgium, 2009.

[6] T. Eriksen, G. Høye, B. Narheim, and B. J. Meland, "Maritime traffic monitoring using a space-based AIS receiver," *Acta Astron.*, vol. 58, no. 10, pp. 537–549, 2006.

[7] G. K. Høye, T. Eriksen, B. J. Meland, and B. T. Narheim, "Space-based AIS for global maritime traffic monitoring," *Acta Astron.*, vol. 62, nos. 2–3, pp. 240–245, 2008.

[8] G. Pallotta, S. Horn, P. Braca, and K. Bryan, "Context-enhanced vessel prediction based on Ornstein–Uhlenbeck processes using historical AIS traffic patterns: Real-world experimental results," in *Proc. 17th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2014, pp. 1–7.

[9] F. Mazzarella *et al.*, "Data fusion for wide-area maritime surveillance," in *Proc. COST MOVE Workshop Moving Objects Sea*, Brest, France, 2013, pp. 1–5.

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

[11] 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. 6, pp. 2218–2245, 2013.

[12] F. Natale, M. Gibin, A. Alessandrini, M. Vespe, and A. Paulrud, "Mapping fishing effort through AIS data," *PLoS ONE*, vol. 10, no. 6, p. e0130746, 2015.

[13] F. Mazzarella, M. Vespe, D. Damalas, and G. Osio, "Discovering vessel activities at sea using AIS data: Mapping of fishing footprints," in *Proc. 17th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2014, pp. 1–7.

[14] B. Ristic, B. La Scala, M. Morelande, and N. Gordon, "Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction," in *Proc. 11th Int. Conf. Inf. Fusion*, Jun./Jul. 2008, pp. 1–7.

[15] 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 *Proc. 12th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2009, pp. 756–763.

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

[17] F. Mazzarella, M. Vespe, A. Alessandrini, D. Tarchi, G. Aulicino, and A. Vollero, "A novel anomaly detection approach to identify intentional AIS on-off switching," *Expert Syst. Appl.*, vol. 78, pp. 110–123, Jul. 2017.

[18] *Revised Performance Standards and Functional Requirements for LRIT*, Int. Maritime Org., London, U.K., 2008.

[19] *Safety of Life at Sea Convention Chapter V, Regulation 19-2*, SOLAS, London, U.K., 2000.

[20] M. McIntyre, L. Genik, P. Mason, and T. Hammond, "Towards an understanding of security, privacy and safety in maritime self-reporting systems," in *Proc. IFIP Int. Conf. Trust Manage.*, vol. 238. 2007, pp. 185–206.

[21] A.-L. S. Lapinski, A. W. Isenor, and S. Webb, "Simulating surveillance options for the Canadian North," *J. Navigat.*, vol. 69, no. 5, pp. 940–954, 2016.

[22] *Directive 2009/16/EC on Port State Control*, Eur. Parliament Council, Brussels, Belgium, 2009.

[23] G. C. Carr, H. Erzberger, and F. Neuman, "Fast-time study of airline-influenced arrival sequencing and scheduling," *J. Guid., Control, Dyn.*, vol. 23, no. 3, pp. 526–531, 2000.

[24] K. Roy, B. Levy, and C. J. Tomlin, "Target tracking and estimated time of arrival (ETA) prediction for arrival aircraft," in *Proc. AIAA Guid., Navigat., Control Conf. Exhib.*, Aug. 2006, pp. 1–22.

[25] A. Lim, B. Rodrigues, and Y. Zhu, "Airport gate scheduling with time windows," *Artif. Intell. Rev.*, vol. 24, no. 1, pp. 5–31, 2005.

[26] M. E. Narciso and M. A. Piera, "Robust gate assignment procedures from an airport management perspective," *Omega*, vol. 50, pp. 82–95, Jan. 2015.

[27] D. A. Propp and C. A. Rosenberg, "A comparison of prehospital estimated time of arrival and actual time of arrival to an emergency department," *Amer. J. Emergency Med.*, vol. 9, no. 4, pp. 301–303, 1991.

[28] B. Xiong, P. B. Luh, and S. C. Chang, "Group elevator scheduling with advanced traffic information for normal operations and coordinated emergency evacuation," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2005, pp. 1419–1424.

[29] A. Karbassi and M. Barth, "Vehicle route prediction and time of arrival estimation techniques for improved transportation system management," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2003, pp. 511–516.

[30] E. Mazloumi, G. Rose, G. Currie, and M. Sarvi, "An integrated framework to predict bus travel time and its variability using traffic flow data," *J. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 75–90, 2011.

[31] D. Fagan and R. Meier, "Intelligent time of arrival estimation," in *Proc. IEEE Forum Integr. Sustain. Transp. Syst. (FISTS)*, Jun. 2011, pp. 60–66.

[32] C. Heywood, C. Connor, D. Browning, M. C. Smith, and J. Wang, "GPS tracking of intermodal transportation: System integration with delivery order system," in *Proc. Syst. Inf. Eng. Design Symp. (SIEDS)*, Apr. 2009, pp. 191–196.

[33] M. Christiansen, K. Fagerholt, B. Nygreen, and D. Ronen, "Ship routing and scheduling in the new millennium," *Eur. J. Oper. Res.*, vol. 228, no. 3, pp. 467–483, 2013.

[34] G. Mannarini, G. Coppini, P. Oddo, and N. Pinardi, "A prototype of ship routing decision support system for an operational oceanographic service," *TransNav, Int. J. Marine Navigat. Safety Sea Transp.*, vol. 7, no. 1, pp. 53–59, 2013.

[35] A. Alessandrini *et al.*, "Data driven contextual knowledge from and for maritime situational awareness," in *Proc. Context-Awareness Geograph. Inf. Services (CAGIS)*, 2014, pp. 39–49.

[36] M. Vespe, H. Greidanus, and M. A. Alvarez, "The declining impact of piracy on maritime transport in the Indian Ocean: Statistical analysis of 5-year vessel tracking data," *Marine Policy*, vol. 59, no. 30, pp. 9–15, 2015.

[37] J. van Etten, "R package gdistance: Distances and routes on geographical grids," *J. Stat. Softw.*, vol. 76, no. 13, pp. 1–21, 2017. [Online]. Available: https://www.jstatsoft.org/v076/i13, doi: 10.18637/jss.v076.i13.

[38] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numer. Math.*, vol. 1, no. 1, pp. 269–271, Dec. 1959.

[39] H. Alt, B. Behrends, and J. Blömer, "Approximate matching of polygonal shapes (extended abstract)," in *Proc. 7th Annu. Symp. Comput. Geometry (SCG)*, 1991, pp. 186–193.

[40] F. Mazzarella, M. Vespe, and C. Santamaria, "SAR ship detection and self-reporting data fusion based on traffic knowledge," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 8, pp. 1685–1689, Aug. 2015.

Alfredo Alessandrini received the degree in forestry and environmental management from University of Tuscia in 2005 and the Ph.D. degree in forestry management in 2008. From 2011 to 2012, he was a Post-Doctoral Researcher with the EMMAH INRA Unit, Avignon, France, where he was involved in remote sensing techniques and field experimental surveys to model the functioning of the Mediterranean forest ecosystem. From 2012 to 2015, he was a Post-Doctoral Researcher with the Joint Research Centre (JRC) Maritime Unit, to support the implementation and monitoring of EU maritime policies mainly through the application of data mining, statistics, and geospatial analysis techniques. He has been a Researcher with the JRC, Ispra, since 2012. He is currently a Geographical Data Scientist with the Demography, Migration, and Governance Unit, a European Commission Centre of expertise created to support EU policies related to migration and demography.

Fabio Mazzarella received the M.S. degree in electronics engineering and the Ph.D. degree in information engineering from University of Salerno, Italy, in 2008 and 2012, respectively. He is currently a Data Scientist with Vodafone Automotive Italia SpA, where he is involved in AI and Machine Learning. From 2012 to 2017 he was a Research Fellow with the European Commission Joint Research Centre, where he conducted research on information fusion, knowledge discovery, and data mining. His research interests include statistical signal processing, machine learning, and data mining.

Michele Vespe received the degree in telecommunications engineering from University of Florence in 2003 and the Ph.D. degree in signal processing from University College London in 2006. From 2008 to 2011 he was a Post-Doctoral Researcher with the Joint Research Centre (JRC) developing synthetic aperture radar based pre-operational applications in the maritime domain. From 2011 to 2013, he was a Senior Scientist with the NATO Centre for Maritime Research and Experimentation, involved in maritime data mining, anomaly detection, and networked radar systems. He is currently a Scientific Officer with the European Commission—JRC, where he coordinates the activities of a team of researchers in transforming data into knowledge and scientific evidence for policy and decision-making. Prior to this, he spent two years in industry as a Project Engineer designing small- and medium-area surveillance and data fusion tools.