SCIENCE AND TECHNOLOGY ORGANIZATION
CENTRE FOR MARITIME RESEARCH AND EXPERIMENTATION

Reprint Series **CMRE-PR-2019-053**

Scalable distributed change detection and its application to maritime traffic

Leonardo M. Millefiori, Paolo Braca, Gianfranco Arcieri

June 2019

Originally published in:

2017 IEEE International Conference on Big Data (Big Data), 11-14 December 2017, Boston, MA, USA, pp. 1650-1657, doi: 10.1109/BigData.2017.8258101

About CMRE

The Centre for Maritime Research and Experimentation (CMRE) is a world-class NATO scientific research and experimentation facility located in La Spezia, Italy.

The CMRE was established by the North Atlantic Council on 1 July 2012 as part of the NATO Science & Technology Organization. The CMRE and its predecessors have served NATO for over 50 years as the SACLANT Anti-Submarine Warfare Centre, SACLANT Undersea Research Centre, NATO Undersea Research Centre (NURC) and now as part of the Science & Technology Organization.

CMRE conducts state-of-the-art scientific research and experimentation ranging from concept development to prototype demonstration in an operational environment and has produced leaders in ocean science, modelling and simulation, acoustics and other disciplines, as well as producing critical results and understanding that have been built into the operational concepts of NATO and the nations.

CMRE conducts hands-on scientific and engineering research for the direct benefit of its NATO Customers. It operates two research vessels that enable science and technology solutions to be explored and exploited at sea. The largest of these vessels, the NRV Alliance, is a global class vessel that is acoustically extremely quiet.

CMRE is a leading example of enabling nations to work more effectively and efficiently together by prioritizing national needs, focusing on research and technology challenges, both in and out of the maritime environment, through the collective Power of its world-class scientists, engineers, and specialized laboratories in collaboration with the many partners in and out of the scientific domain.

NOTE: The CMRE Reprint series reprints papers and articles published by CMRE authors in the open literature as an effort to widely disseminate CMRE products. Users are encouraged to cite the original article where possible.

Scalable Distributed Change Detection and its Application to Maritime Traffic

Leonardo M. Millefiori*, Paolo Braca*, and Gianfranco Arcieri*
*NATO STO Centre for Maritime Research and Experimentation (CMRE), La Spezia, Italy
Email: {leonardo.millefiori, paolo.braca, gianfranco.arcieri}@cmre.nato.int

***Abstract*—Building on a novel methodology based on the Ornstein-Uhlenbeck (OU) process to perform accurate long-term predictions of future positions of ships at sea, we present a statistical approach to the detection of abrupt changes in the process parameter that represents the desired velocity of a ship. Proceeding from well-established change detection techniques, the proposed strategy is also computationally efficient and fit well with big data processing models and paradigms. We report results with a large real-world Automatic Identification System (AIS) data set collected by a network of terrestrial receivers in the Mediterranean Sea from June to August 2016.**

***Keywords*-statistical change detection, long-term target state predicion, AIS, big data, Apache Spark**

I. Introduction

Primarily based on the Automatic Identification System (AIS) with which ships, among other related information, broadcast their position, speed and course details, maritime traffic monitoring networks are increasingly being used to achieve maritime surveillance capabilities. In parallel, AIS data analysis and research [1]–[6], have proven to be valid methods for monitoring vessels and extracting valuable information regarding their behavior, patterns and statistics. Also, maritime traffic and global compliance with international AIS requirements are steadily increasing, and worldwide networks of AIS receivers consequently growing and producing larger and larger volumes of AIS data. To give an example, the NATO Science and Technology Organization (STO) Centre for Maritime Research and Experimentation (CMRE) continuously receives, stores and analyses quasi-real-time streams from multiple aggregation services. Every month, this amounts to approximatively 800 million AIS messages from aggregation sources, produced by over 100 000 unique vessels [7]. Such a volume of data clearly poses harsh challenges when it comes to any real-world application in this field. Indeed, larger volumes of data bring scalability, complexity, generalization and interpretability challenges to this field, which remain largely unaddressed in the open literature, where small dataset are mostly used to demonstrate feasibility and effectiveness of algorithms that are not usually designed to scale to big data regimes.

Given that current networks are largely based on land receivers, coverage and persistence is challenging with AIS. In other words, vessels in open seas can be seldom *continuously* observed, resulting in highly intermittent and sparse data. The problem of long-term vessel state estimate and prediction is therefore crucial to achieve an effective surveillance capability of ships at sea with AIS.

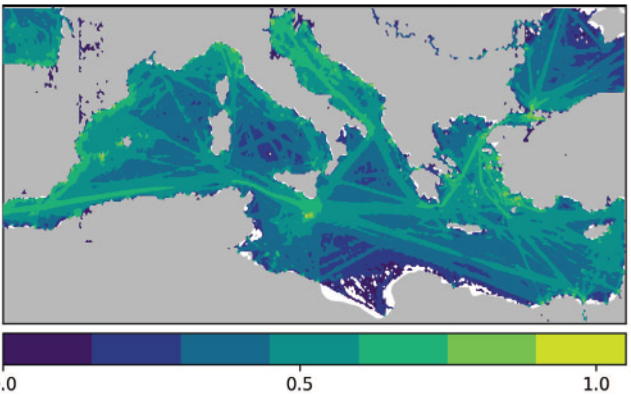


Figure 1. Density of AIS messages collected by a network of terrestrial AIS base stations from June to August 2016. Each pixel covers a 6-by-6 nmi (one-tenth-degree) square on the ground and its color is (logarithmically) proportional to the number of AIS messages whose reported positions fall within its footprint.

A novel method for the long-term state prediction of non-maneuvering targets has been recently proposed [8], which models the target kinematic state with an Ornstein-Uhlenbeck (OU) process on the velocity. Validated against a large real-world dataset [3], [8], the proposed modeling enables a more accurate representation of long-term target state estimates, when the target is not maneuvering. The main difference between the OU process and other conventional nearly-constant velocity models is a feedback loop, which ensures that the velocity of the target does not diverge with time, but is instead bounded around a finite value, representing the desired (cruise) velocity of the target.

However, such modeling is relevant for ships that do not maneuver (meaning that their desired velocity does not change), which is already the case for a significant portion of commercial maritime traffic, simply because ships seek to optimize fuel consumption. Obviously, vessels do not *always* move in a straight line, and will have, at some point, to maneuver. Interestingly, what we have observed is that even this behavior is somehow very regular, as most of the maritime traffic —and again especially the commercial traffic— tends to navigate by waypoints, as can be deduced also from 1. Put differently, the majority of ships' trajectories can be decomposed in a series of waypoints, where the orientation of the track or the nominal velocity changes, and legs, whereon ships show non-maneuvering behavior. Due to traffic regulations and fuel consumption optimization, these waypoints are also concentrated in specific regions, as can be again seen from Fig. 1.

Based on the OU model proposed in [8], and taking advantage of the performance study in [9], we have developed an efficient statistical change detection procedure to identify the time instants of change in one of the OU process parameters. This procedure constitutes a stepping stone towards the relaxation of the non-maneuvering target assumption, and enables us to build a more realistic motion model, based on a desired velocity which is a piecewise-constant function of time.

As in [10], the real-word case study that we consider is focused on the detection of ships approaching or leaving stationary areas. At the same time, the approach presented in this paper is fundamentally different from the data-driven procedure developed therein, which was meant to identify the extent of stationary areas from the higher observed density of AIS messages in a given region of interest. In this paper, we address instead the problem of statistically detecting abrupt changes in one of the OU process parameters with a large volume of data; change points that we expect to be more concentrated in specific regions, that can be either waypoint regions or stationary areas. Specifically, we apply the designed technique to the entirety of positional AIS message broadcast by ships in the Mediterranean Sea from June to August 2016 and collected by a network of terrestrial AIS base stations, with a big-data-oriented implementation of the above technique that naturally scales out to the large amounts of data at hand.

II. Mean-reverting process model

In this section we introduce the mean-reverting process model for the target dynamics. Let us indicate the target state at time $t \in \mathbb{R}_0^+$ with

$$\boldsymbol{s}\left(t\right) = \begin{bmatrix} \boldsymbol{x}\left(t\right) \\ \dot{\boldsymbol{x}}\left(t\right) \end{bmatrix},$$

where $\boldsymbol{x}\left(t\right)$ and $\dot{\boldsymbol{x}}\left(t\right)$ denote the target position and velocity, respectively, in a two-dimensional Cartesian reference system

$$\boldsymbol{x}(t) \overset{def}{=} \left[x(t),\, y(t)\right]^{\mathsf{T}}, \quad \dot{\boldsymbol{x}}(t) \overset{def}{=} \left[\dot{x}(t),\, \dot{y}(t)\right]^{\mathsf{T}}. \quad (1)$$

In general, the target dynamics can be modeled with a set of linear stochastic differential equations (SDEs) [11]. In [8] it is shown how the movement of real non-maneuvering vessels in the open sea can be represented by a mean-reverting stochastic process. Specifically, the velocity of the target can be modeled as an OU process, and hence its position with an Integrated Ornstein-Uhlenbeck (IOU) process. Under this assumption, the SDE for the target motion model has the following form

$$\mathrm{d}\boldsymbol{s}(t) = \boldsymbol{A}\,\boldsymbol{s}(t)\,\mathrm{d}t + \boldsymbol{G}\,\boldsymbol{v}\,\mathrm{d}t + \boldsymbol{B}\;\mathrm{d}\boldsymbol{w}(t), \qquad (2)$$

where $\boldsymbol{v} = \left[v_x, v_y\right]^T$ is the long-run process mean, and $\boldsymbol{w}(t)$ is a standard bi-dimensional Wiener process. The matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{G}$ are defined as:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{0} & -\boldsymbol{\Theta} \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{C} \end{bmatrix}, \quad \boldsymbol{G} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\Theta} \end{bmatrix}, \qquad (3)$$

being $\boldsymbol{0}$ the null matrix and $\boldsymbol{\Theta}$ and $\boldsymbol{C}$ generic bi-dimensional matrices. The matrix $\boldsymbol{\Theta}$ quantifies the mean-reversion effect, meaning the rate at which the target will tend back to the desired speed after a perturbation; its diagonal terms refer to the x and y components, while the off-diagonals quantify the coupling effect.

Let us denote the set of observation time instants t_k, with $t_K > \cdots > t_k > \cdots > t_1$. Unless otherwise stated, we will use hereafter subscripted indices to denote time dependency, i.e. $\boldsymbol{x}_k = \boldsymbol{x}(t_k)$, $\dot{\boldsymbol{x}}_k = \dot{\boldsymbol{x}}(t_k)$, $\boldsymbol{s}_k = \boldsymbol{s}(t_k)$ by definition. Let us also define the inter-observation times as $\Delta_k = t_k - t_{k-1}$.

We also assume that $\boldsymbol{\Theta}$ has positive and distinct eigenvalues, so that an affine transformation can be found that projects the matrix $\boldsymbol{\Theta}$ onto another space, i.e. $\boldsymbol{\Theta} = \boldsymbol{R}\boldsymbol{\Gamma}\boldsymbol{R}^{-1}$, where $\boldsymbol{\Gamma}$ is diagonal [8]. Under these assumptions, the target state at time t_k, given the target state at time t_{k-1}, is provided by the first moment of the SDE solution [12], and can be rearranged in matrix form [8]:

$$\boldsymbol{s}_k = \widetilde{\boldsymbol{R}}\boldsymbol{\Phi}\left(\Delta_k, \gamma\right)\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{s}_{k-1} + \widetilde{\boldsymbol{R}}\,\boldsymbol{\Psi}\left(\Delta_k, \gamma\right)\boldsymbol{R}^{-1}\boldsymbol{v} + \boldsymbol{w}_k, \ (4)$$

where $\boldsymbol{w}_k$ is a zero-mean Gaussian random variable with a covariance reported in [8], and $\widetilde{\boldsymbol{R}}$ is constructed as follows

$$\widetilde{\boldsymbol{R}} \overset{def}{=} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix}.$$

The state transition matrix and the control input function, $\boldsymbol{\Phi}\left(t, \gamma\right)$ and $\boldsymbol{\Psi}(t, \gamma)$, respectively, are defined as

$$\boldsymbol{\Phi}\left(t, \gamma\right) = \begin{bmatrix} \boldsymbol{I} & \left(\boldsymbol{I} - \mathrm{e}^{-\boldsymbol{\Gamma}\,t}\right)\boldsymbol{\Gamma}^{-1} \\ \boldsymbol{0} & \mathrm{e}^{-\boldsymbol{\Gamma}\,t} \end{bmatrix}, \qquad (5)$$

$$\boldsymbol{\Psi}(t, \gamma) = \begin{bmatrix} t\,\boldsymbol{I} - \left(\boldsymbol{I} - \mathrm{e}^{-\boldsymbol{\Gamma}\,t}\right)\boldsymbol{\Gamma}^{-1} \\ \boldsymbol{I} - \mathrm{e}^{-\boldsymbol{\Gamma}\,t} \end{bmatrix}. \qquad (6)$$

Considering only the velocity terms in (4) we have

$$\dot{\boldsymbol{x}}_k = \boldsymbol{v} + \boldsymbol{J}_k\left(\dot{\boldsymbol{x}}_{k-1} - \boldsymbol{v}\right) + \boldsymbol{\omega}_k, \qquad (7)$$

where $\boldsymbol{J}_k = \boldsymbol{R}\,\mathrm{e}^{-\boldsymbol{\Gamma}\Delta_k}\boldsymbol{R}^{-1}$, and the $\boldsymbol{\omega}_k$ are independent zero-mean Gaussian random variables with bi-dimensional covariance $\boldsymbol{\Xi}_k$ given by [8], [9]

$$\boldsymbol{\Xi}_k = \begin{cases} \boldsymbol{\Sigma}\left(\Delta_k\right) & \text{if } k > 1, \\ \boldsymbol{\Sigma}_\infty & \text{if } k = 1, \end{cases} \qquad (8)$$

where $\boldsymbol{\Sigma}\left(t\right) = \boldsymbol{\Sigma}_\infty \circ \boldsymbol{\Omega}\left(t\right)$, with the $\circ$ operator denoting the Hadamard product. The matrices $\boldsymbol{\Sigma}_\infty$ and $\boldsymbol{\Omega}\left(t\right)$ are defined as in [8], [9]

$$\boldsymbol{\Sigma}_\infty = \boldsymbol{R}\,\boldsymbol{\Sigma}\,\boldsymbol{R}^{-1}, \qquad \boldsymbol{\Sigma} = \frac{1}{2}\begin{bmatrix} \frac{\sigma_x^2}{\gamma_x} & \frac{2\sigma_{xy}}{\gamma_x+\gamma_y} \\ \frac{2\sigma_{xy}}{\gamma_x+\gamma_y} & \frac{\sigma_y^2}{\gamma_y} \end{bmatrix}, \tag{9}$$

$$\boldsymbol{\Omega}\left(t\right) = \begin{bmatrix} 1-\mathrm{e}^{-2\gamma_x t} & 1-\mathrm{e}^{-(\gamma_x+\gamma_y)t} \\ 1-\mathrm{e}^{-(\gamma_x+\gamma_y)t} & 1-\mathrm{e}^{-2\gamma_y t} \end{bmatrix}. \tag{10}$$

III. Detection of a single change of the long-run mean parameter value

We propose to model waypoints in the target trajectory as abrupt changes of the long-run mean parameter $\boldsymbol{v}$ of the OU process (2). At an unknown time instant t^*, the velocity abruptly changes from $\boldsymbol{v}_0$ (hypothesis $\mathcal{H}_0$) to $\boldsymbol{v}_1$ (hypothesis $\mathcal{H}_1$), being $\boldsymbol{x}^c = \boldsymbol{x}(t^*)$ the position of the waypoint. *Stopping* and *starting* points are special cases of waypoints. A generic waypoint is simply a change from, and to, a non-null velocity, i.e. $\boldsymbol{v}_0 \neq \boldsymbol{0} \neq \boldsymbol{v}_1$, while in a starting point, the change happens from a null velocity $\boldsymbol{v}_0 = \boldsymbol{0}$ to a non-null one $\boldsymbol{v}_1 \neq \boldsymbol{0}$; vice versa for a stopping point. Our aim is to detect these change points in the target trajectory and the estimate the time instants of change.

We formalize this problem in a statistical framework, after the change detection theory [13]. We assume that the time change t^* is synchronized with observation time instants, i.e. $t^* = t_{k^*} \in \{t_k\}_{k=1}^n$. For the sake of convenience, we define the differences of velocities as

$$\boldsymbol{z}_k = \dot{\boldsymbol{x}}_k - \boldsymbol{J}_k \dot{\boldsymbol{x}}_{k-1}.$$

In this way, the $\boldsymbol{z}_k$ are time independent Gaussian random variables, i.e. $\boldsymbol{z}_k \sim \mathcal{N}\left(\boldsymbol{z}; \left(\boldsymbol{I}-\boldsymbol{J}_k\right)\boldsymbol{v}; \boldsymbol{\Xi}_k\right)$ with $\boldsymbol{v} \in \{\boldsymbol{v}_0, \boldsymbol{v}_1\}$. In our notation, $\mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}; \boldsymbol{\Xi}\right)$ denotes a Gaussian Probability Density Function (PDF) with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Xi}$. Thanks to the mean of each $\boldsymbol{z}_k$ being dependent on the long-term velocity, the waypoint (stopping point) can be found by detecting a change in the mean of $\boldsymbol{z}_k$. Formally, we have, $\forall k$:

$$\begin{array}{ll} f_0(\boldsymbol{z}_k) : \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{k^*-1} & \\ & \searrow \\ f_1(\boldsymbol{z}_k) : & \qquad \boldsymbol{z}_{k^*}, \boldsymbol{z}_{k^*+1}, \ldots \end{array} \tag{11}$$

where $f_{0,1}(\boldsymbol{z}_k) = \mathcal{N}\left(\left(\boldsymbol{I}-\boldsymbol{J}_k\right)\boldsymbol{v}_{0,1}; \boldsymbol{\Xi}_k\right)$.

Now, if all the process parameters are known, i.e. the $f_{0,1}(\boldsymbol{z}_k)$ are perfectly known, and the most suitable change detection procedure is Page's test [14], as it shows asymptotic optimality properties (further details in [13]). Page's test is based on the clipped Cumulative Sum (CUSUM) statistic:

$$S_k = \max\left\{0, S_{k-1} + \log\frac{f_1(\boldsymbol{z}_k)}{f_0(\boldsymbol{z}_k)}\right\}, \quad S_0 = 0. \tag{12}$$

In practice, a change from $\boldsymbol{v}_0$ to $\boldsymbol{v}_1$ is declared when S_k exceeds a threshold h; such time instant can be defined as

$$\mathcal{K} = \min\left\{k : S_k > h\right\}. \tag{13}$$

The time of the change can be estimated with a Maximum Likelihood (ML) procedure [13] that looks for the time of the last CUSUM reset. As in [13], we have, for our case:

$$\begin{aligned} \hat{k}^* &= \mathcal{K} - N_{\mathcal{K}} + 1, \\ N_k &= N_{k-1}\boldsymbol{1}_{\{S_{k-1}>0\}} + 1. \end{aligned} \tag{14}$$

Then, the estimated time of the change is the following $t_{\hat{k}^*}$.

IV. Detection of multiple changes of the long-run mean process parameter

In a real-world case, target trajectories show several waypoints and stopping points. In other words, the long-term velocity is a piecewise-constant function of elements $\boldsymbol{v}_j$, $j = 0, 1, \ldots, J-1$ where J is the unknown number of all the changes. Formally, (11) can be generalized, such as

$$f_j(\boldsymbol{z}_k) = \mathcal{N}\left(\left(\boldsymbol{I}-\boldsymbol{J}_k\right)\boldsymbol{v}_j; \boldsymbol{\Xi}_k\right).$$

In addition to this, in any real-world application, the value of each $\boldsymbol{v}_j$ is also unknown, except for starting or stopping points, in which the velocity is null either before or after (respectively) the change. A sequential procedure can be formulated to estimate the piecewise long-run velocity, by generalizing the Page's test (12). This procedure identifies: $i)$ $\boldsymbol{v}_j$; $ii)$ the changing times $t_{k_j^*}$ and $iii)$ the locations of the changes $\boldsymbol{x}_{k_j^*}$.

The long-run velocity can be efficiently estimated using a Sample Mean Estimator (SME) [9], provided that the underlying unknown long-run velocity value is constant. As in [13], a change detection algorithm is here used to check this assumption, assuming that the time duration between successive jumps is bounded from below. This assumption is necessary for the initial estimation of the long-run velocity to be used in the subsequent detection of change. The joint use of the estimation in [9] and the change detection results in cycles made of the following steps.

S0 Initialization. Set $j = 0$. Estimate the long-run velocity, in a fixed-size time interval N, during which the detection algorithm is switched off. The first piece of the long-run velocity function is provided by

$$\hat{\boldsymbol{v}}_0 = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n. \tag{15}$$

Define a set of P alternative hypotheses with P long-term velocities that are relatively close to $\hat{\boldsymbol{v}}_0$, and formally indicated by $\hat{\boldsymbol{v}}_{1,p}(\delta)$, $p = 1, 2, \ldots, P$.

S1 Activation of the change detection algorithm to find the j-th change against one of the P alternative hypotheses. Compute in parallel the CUSUM statistics for all the P alternative hypothesis defined in the previous step, using Gaussian distributions with long-term velocities $\hat{\boldsymbol{v}}_j$ and $\hat{\boldsymbol{v}}_{j+1,p}$, respectively, i.e. $f_0(\boldsymbol{z}_k) = \mathcal{N}\left(\left(\boldsymbol{I}-\boldsymbol{J}_k\right)\hat{\boldsymbol{v}}_0; \boldsymbol{\Xi}_k\right)$ and $f_{j+1,p}(\boldsymbol{z}_k) = \mathcal{N}\left(\left(\boldsymbol{I}-\boldsymbol{J}_k\right)\hat{\boldsymbol{v}}_{j+1,p}; \boldsymbol{\Xi}_k\right)$.

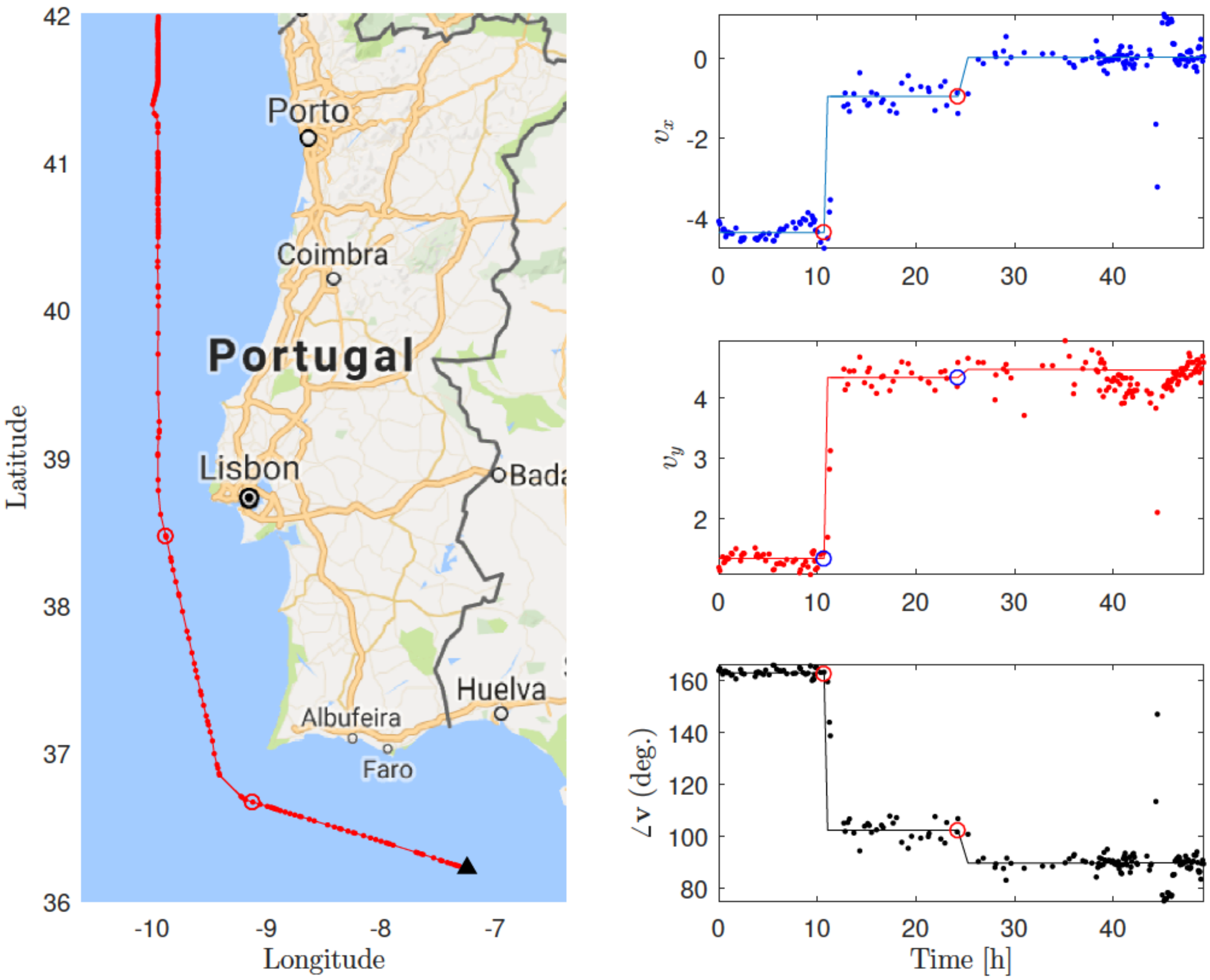


Figure 2. Example of the change detection procedure applied on a single track. The left-hand side panel shows the selected track, which comes from a cargo vessel. The black triangle denotes the first data sample of the track and therefore shows the motion direction. The red circles show the detected changes in the parameter. Right-hand side, from top to bottom: instantaneous value of the Cartesian components of the velocity over time, along x and y, respectively, and instantaneous track orientation; dots denote the instantaneous values, solid lines the value of the parameter, and circles the change detections.

S2 If none of the CUSUM statistics for the P alternative hypotheses reaches a threshold h, the procedure ends.

S3 When a change is detected, the time of change is estimated as in (14). The position of the change (i.e. the waypoint, stopping or starting point) is then provided by $\hat{x}^c_j = x(\hat{t}_{\hat{k}^*_j})$. Increment j.

S4 Update the estimation of the long-run velocity after the change has been detected:

$$\hat{v}_j = \frac{1}{N} \sum_{n=\hat{k}^*_{j-1}+M}^{k^*_{j-1}+M+N} x_n, \qquad (16)$$

where the delay parameter M is used to neglect the transitory points of the OU process that would deteriorate the estimation performance. The long-run velocities of the alternative hypotheses are updated accordingly to the newly estimated long-run velocity value and the parameter δ. Return to **S1**.

For the waypoints, the alternative hypotheses should be representative of the deviations, quantified by the parameter δ, from the null ($\mathcal{H}_0$) hypothesis. If a bigger deviation occurs, the change will still be detected but at the cost of a degraded detection performance; in other words, the smaller is δ, the larger will be the detection delay for a given false alarm rate.

Put differently, the choice of δ poses a trade-off between the change detection performance and the capability to detect small deviations from the null hypothesis. We propose the following strategies to construct meaningful deviations from the null hypothesis:

$i)$ two-side additive deviation on both the axes

$$\hat{v}_{j+1,p}(\delta) = \begin{cases} \hat{v}_j + \delta\, e_x & p = 1 \\ \hat{v}_j - \delta\, e_x & p = 2 \\ \hat{v}_j + \delta\, e_y & p = 3 \\ \hat{v}_j - \delta\, e_y & p = 4 \end{cases} \tag{17}$$

where $e_x = [1,0]^T$ and $e_y = [0,1]^T$.

$ii)$ two-side additive deviation on the direction $\angle\hat{v}_j$ of the long-run velocity of the null hypothesis:

$$\|\hat{v}_{j+1,p}(\delta)\| = \|\hat{v}_j\|\,, \tag{18}$$

$$\angle\hat{v}_{j+1,p}(\delta) = \begin{cases} \angle\hat{v}_j + \delta & p = 1 \\ \angle\hat{v}_j - \delta & p = 2 \end{cases} \tag{19}$$

Fig. 2 depicts an example of the change detection procedure applied on a single track. The left-hand side panel shows the trajectory of a cargo ship navigating off the cost of Portugal. The black triangle denotes the first data sample of the track, and the red circles show the detected changes in the long-run mean parameter. In the right-hand side, from top to bottom, the instantaneous values of the Cartesian components of the velocity are shown over time, along x and y, respectively, and the instantaneous track orientation; dots denote the instantaneous values, solid lines the value of the parameter, and circles the change detections.

V. Approach

A. Data description

The AIS is a collaborative, self-reporting system that, amongst other things, allows ships to broadcast their identity, position and other voyage-related information to nearby vessels and base stations. The system was originally conceived as a safety system for navigation, especially for collision avoidance, but soon after the International Maritime Organization (IMO) mandated ships of certain categories to have AIS transceivers installed onboard, the AIS quickly began the primary means, by coverage and volume of data, of maritime traffic surveillance.

The ITU 1371-4 standard defines 64 different types of AIS messages that can be broadcast by AIS transceivers. In this work we focus only on types 1, 2 and 3, which are most frequently broadcast [15]. These message types are position reports, meaning that they include the position of the ship (latitude and longitude) and other information related to the ship's motion (e.g. Speed Over Ground (SOG), Course Over Ground (COG), and more). Each vessel is identified by its MMSI number, which we will use to reconstruct the ship's trajectory. The AIS communication protocol is asynchronous and prescribes that different types of messages are to be transmitted with different frequencies: static information (type 5 messages) every 6 minutes; position information every 2 seconds to 3 minutes, depending on the speed, location, and navigational status of the vessel.

In the remainder of this paper, we apply our approach to a dataset of more than 73 million AIS messages collected by a worldwide collaborative network of AIS terrestrial receivers and recorded from June to August 2016, in an area of interest that spans more than 6×10^6 square km, approximatively from $-5°$ to $35°$ longitude and from $30°$ to $46°$ latitude. For reference, we have reported in Fig. 1 a density map of the dataset over the area of interest; each pixel in the figure covers a 6-by-6 nmi (one-tenth-degree) square on the ground and its color is proportional to the logarithm of the number of recorded AIS messages whose reported positions fall within its footprint.

B. Processing chain

In Fig. 3 we report a block diagram representation of the data processing chain that has been applied to the aforementioned real-world data set. As it is formulated, the problem becomes embarrassingly parallel[1], once that the AIS messages are aggregated in tracks, and perfectly fits to be distributed using Apache Spark to multiple computing nodes. The only criticality is that Spark heavily relies on data locality, which makes Spark jobs particularly sensitive to where the data is located. It is therefore mandatory to partition and distribute wisely the input dataset in order to minimize data transfer among the nodes that carry out the computation and consequently maximize the throughput.

A pre-processing stage is required in order to prepare the data for the change detection procedure described in Sect. IV. We assume that the AIS data is stored on disk in raw NMEA format. As shown in Fig. 3, the first operation of this initial pre-processing phase (*Stage 0*) is therefore to read the raw data files and decode the information. This leaves us with a collection of dictionary objects, one for each received AIS message, among which only those that contain positional information (specifically AIS message types 1, 2 and 3) are retained, discarding all the other message types, as not relevant for our analysis.

The decoding stage is then immediately followed by a first validation routine, which discards messages with invalid MMSI number. In fact, it is not uncommon that invalid MMSI numbers are broadcast by ships, either intentionally or not. Even if seemingly of secondary importance, this step is nonetheless important for our analysis, as we will later use the MMSI number to *reconstruct* ships' trajectories. The positional information in the messages that satisfy the MMSI validation check is then used to project the position of the

[1] The proposed approach is preliminary to the problem of discovering recurrent waypoint areas with increasing volumes of data, which is not embarrassingly parallel *by definition*. Much the contrary, pattern knowledge discovery algorithms available in literature are usually either unable to process large volumes of data in reasonable time or not well suited to run in a parallel or distributed fashion. One important contribution and element of novelty of this work is precisely the formulation of the problem after a change detection procedure, which is what makes the problem embarrassingly parallel.

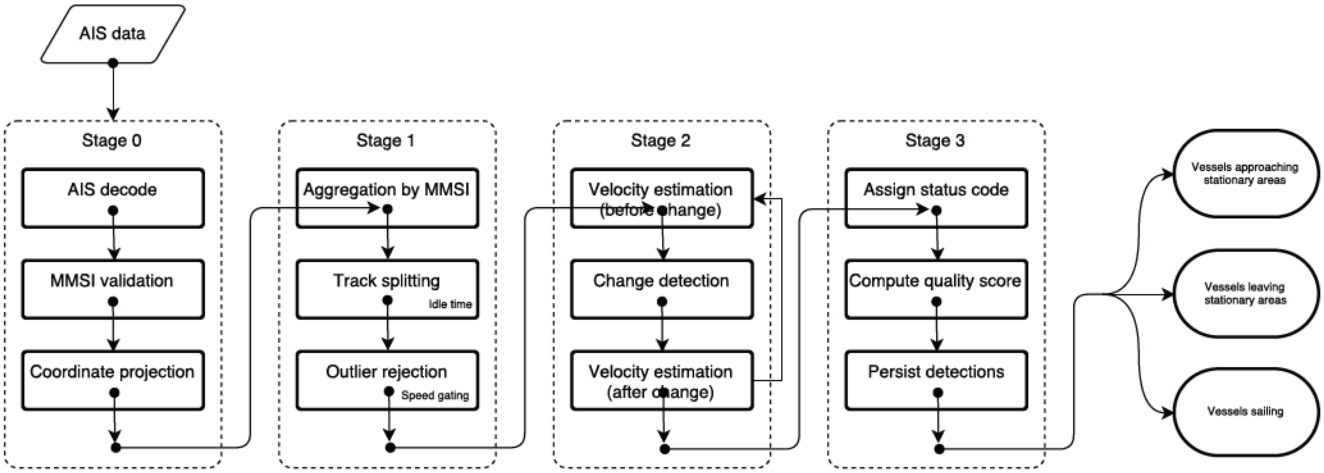


Figure 3. Diagram of the complete change detection pipeline. The input is represented by raw AIS messages in NMEA format, that are loaded in an Apache Spark RDD and immediately decoded. The blocks in stages 0 and 1 mostly deal with the pre-processing of the data set: messages with invalid MMSI numbers are discarded, the geographic coordinates are projected using the UTM projection. In stage 1, the messages are arranged first in ship tracks and then split in time-continuous segments. Outliers are rejected using a speed gating and stage 2 is devoted to the change detection procedure. In stage 3 a status label is assigned to each detection, a quality score computed, and the Spark DataFrame of detections eventually persisted.

ship on a Cartesian system, for which we use the UTM projection.

The next processing stage (*Stage 1* in Fig. 3) builds the ships' tracks, which essentially are time-ordered lists of AIS messages, with some additional operations. More specifically, the *Aggregation by MMSI* processing block rearranges the AIS messages in a key-value structure, having the MMSI numbers as keys and, as values, the related lists of AIS messages, ordered by time. Also, it is not uncommon to observe temporal and spatial *gaps* in the AIS data, happening because the AIS transceiver can be turned off or simply due to the ship exiting from the coverage area. The *Track splitting* block is responsible to split a ship's track into more tracks (track segments) if the communication from the ship was lost for more than a fixed threshold (*idle time*). Finally, a speed gating is applied to reject outliers, meaning AIS messages whose position cannot possibly be compatible with the observation time and position of the same ship at a previous time, which usually happens because of one or more incorrectly timestamped AIS messages (more on this problem can be found in [16]).

Processing *Stage 2* is finally devoted to the change detection technique. As also described in Sections III and IV, the change detection procedure that we have developed relies on the estimation of the long-run mean velocity parameter of the OU process, before and after a possible change. At its core—the *Change detection* block—the CUSUM is computed and, if a deviation from the desired threshold is observed, a change of the long-run mean velocity is declared.

In *Stage 3*, eventually, the detections are extracted and a status code is assigned to each of them, which is basically a category label. If the long-term mean velocity before the change was null and the one after the change was not null, the detection is categorized as a *starting point*; if vice versa the velocity before the change was not null and that after the change was null, the detection is categorized as a *stopping point*; if, finally, the velocities before and after the change were both non-null, the detection is labeled as a *waypoint*.

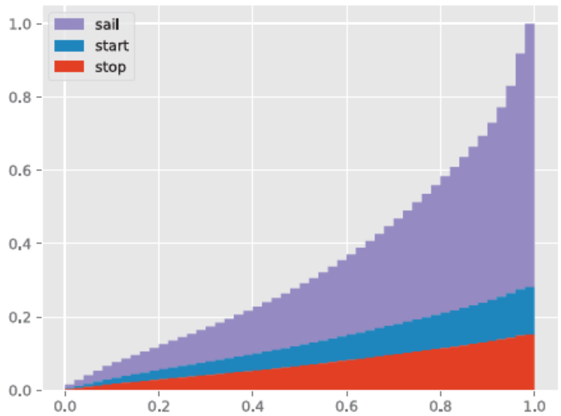
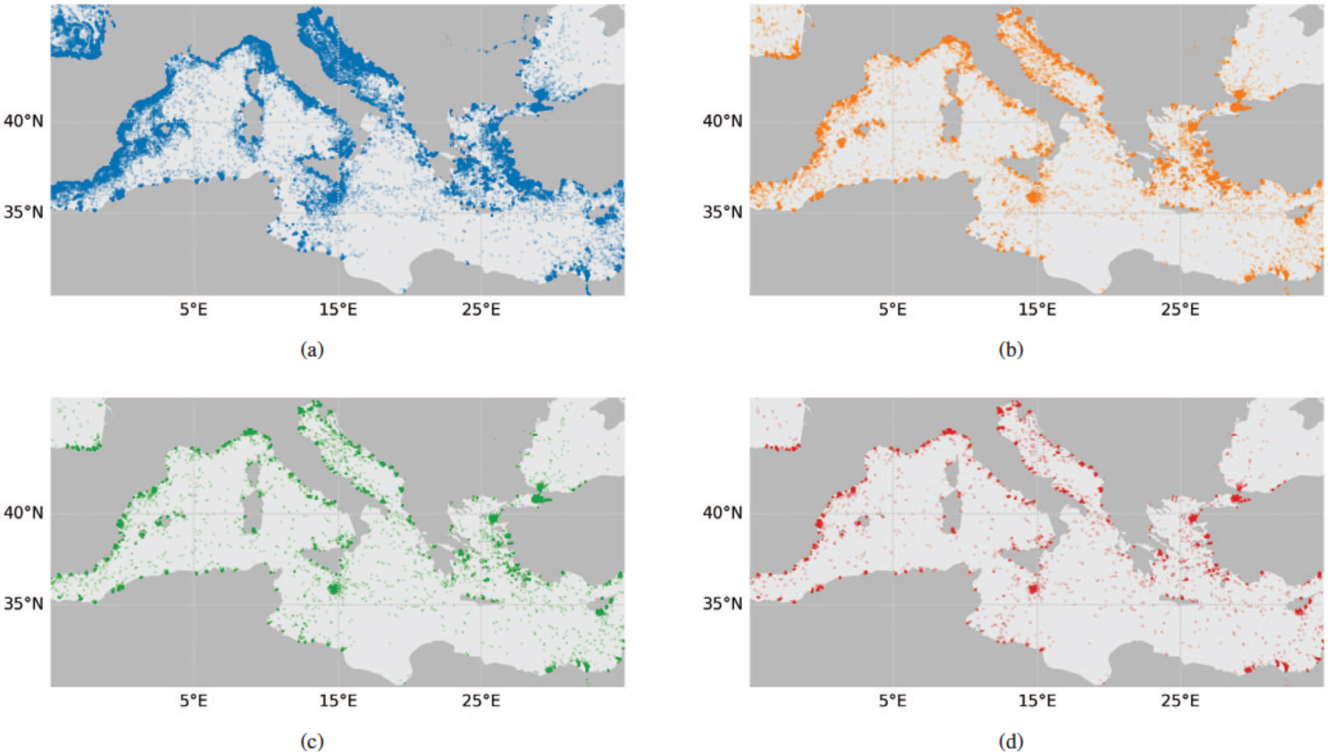


Figure 4. Empirical distribution—cumulative histogram—of the score for the three classes of detections that have been considered: *starting* and *stopping* points, and *waypoints* (sail). The data set included also few negative score values, which have been cropped out of the image.

When applied to real-world data, the proposed procedure has to necessarily deal with all the underlying non-idealities. Although some of them are accounted for and compensated in the preprocessing stages, residual sources of noise and errors should be anyway expected in the process. For this reason, before persisting the data set of detections, one additional block in *Stage 3* computes and assigns a quality score to each detection that has been revealed. The metric that we have used is the Normalized Cross-Correlation



(a) (b) (c) (d)

Figure 5. Detections revealed with the proposed change detection technique on a data set of 73 million AIS messages collected in the area of interest from June to August 2016. Panel (a) shows the entire data set of detections, while (b), (c) and (d) only the 20%, 10% and 5%, respectively, highest score detections. As the bound on the score gets more selective, it is apparent how fewer detections appear on the map, at the same time being more concentrated and closer to the shoreline or to islands.

(NCC) between the speed $\dot{s}_x = [\dot{x}_1, \dots, \dot{x}_n]$ recorded at time instants $t_1, \dots t_n$ and the corresponding long-term velocity mean value, which has been estimated with the change detection procedure, i.e. $\hat{v}_x = \left[\hat{v}_1^{(x)}, \dots, \hat{v}_n^{(x)}\right]$. The operation is repeated for both the Cartesian components and the minimum NCC value is taken as quality score of the detection. This amounts to compute, for each track, the quantity

$$q = \min\left\{\left\langle \frac{\dot{s}_x}{\|\dot{s}_x\|}, \frac{v_x}{\|v_x\|}\right\rangle, \left\langle \frac{\dot{s}_y}{\|\dot{s}_y\|}, \frac{v_y}{\|v_y\|}\right\rangle\right\},$$

which can take values in the interval $[-1, 0) \cup (0, 1]$. By construction, $q = 1$ only if the estimated long-run mean velocity values are all perfectly equal to those of the instantaneous velocity. Conversely, $q = -1$ means that the estimated long-run velocity is the exact opposite of the instantaneous velocity. In other words, the closer the value of q is to 1, the *better* the change detection routine has worked[2].

VI. Results

The approach described in this paper has been applied to a data set of more than 73 million AIS messages collected by a worldwide collaborative network of terrestrial receivers and recorded from June to August 2016, in the area of the Mediterranean Sea (6×10^6 square km), approximatively from $-5°$ to $35°$ longitude and from $30°$ to $46°$ latitude. The size of the original AIS was around 140 GB on disk; the change detection routine[3], leaves us with 997 624 detections, in great part waypoints (711 831); detected starting and stopping points amount to, respectively, 132 506 and 153 287.

Remark: This sample dataset is limited in space and time and is for demonstration purpose only. That being said, rather than on the real-world case study, the focus of this work is on the algorithm and its implementation, which has been entirely designed to fit and run in a parallel and distributed fashion, thus effectively addressing the issue of processing large volumes of data by scaling out the number of computing nodes.

The map in Fig. 5 shows the entire data set of detected starting and stopping points (Fig. 5a), along three derived subsets that only show the 20% (Fig. 5b), 10% (Fig. 5c) and 5% (Fig. 5d) highest scored detections. Another perspective is offered by Fig. 4, which shows the empirical cumulative

[2] In any practical application, $q = 1$ is clearly not achievable, as that would mean to have no dispersion of the instantaneous velocity samples around the long-run mean which, for the OU is a limit and trivial case.

[3] We ran the proposed processing on a tiny Apache Spark cluster running in local mode with 8 executors, having each 1 GB of memory available. In this setup, the total processing time is about 1 hour per month of data.

(normalized) distribution of the detection score for the three detection categories.

Unsurprisingly, the distribution of the detected starting and stopping points is more dense in close proximity of the shoreline, and revealed points get fewer and fewer as we move away from it. Interestingly though, as the bound on the quality score gets more selective and less detections are shown on the map, the detections revealed with the proposed technique, rather than being spread all over the shoreline, concentrate in more constrained locations, presumably corresponding to ports.

VII. Conclusion and future work

Long-term prediction of future positions of ships at sea is an important capability that can help to —at least partially— fill the technological gaps that currently prevent the achievement of global surveillance of ships at sea, especially if complemented with contextual and historical information about ship movements at sea.

Based on a novel ship motion model to issue accurate long-term predictions, we have presented an efficient, statistically sound technique to detect not only starting and stopping points, but also generic waypoints in ships' trajectories. As it is formulated, the problem is also embarrassingly parallel and can easily distributed with multiple processing nodes, needing only to take some precautions when partitioning and distributing the data.

Future steps of this work include the building of an historical knowledge base of ships' movements at sea from the detections obtained with the presented approach. This knowledge base, which might be seen as a network graph of the maritime traffic, will be then leveraged to further refine the prediction technique, in order to issue even more accurate predictions of future positions of ships at sea.

Acknowledgment

This work is supported by the NATO ACT under project SAC000708 – "Data Knowledge Operational Effectiveness."

References

[1] 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 (FUSION)*, June 2008, pp. 1–7.

[2] 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.

[3] L. M. Millefiori, G. Pallotta, P. Braca, S. Horn, and K. Bryan, "Validation of the Ornstein-Uhlenbeck route propagation model in the Mediterranean Sea," in *OCEANS 2015 - Genova*, May 2015, pp. 1–6.

[4] 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.

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

[6] 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.

[7] G. Cimino, G. Arcieri, S. Horn, and K. Bryan, "Sensor data management to achieve information superiority in Maritime Situational Awareness," *CMRE Formal Report, NATO UNCLASSIFIED*, 2014.

[8] L. M. Millefiori, P. Braca, K. Bryan, and P. Willett, "Modeling vessel kinematics using a stochastic mean-reverting process for long-term prediction," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 5, pp. 2313–2330, October 2016.

[9] 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.

[10] 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.

[11] X. Rong Li and V. Jilkov, "Survey of maneuvering target tracking. part I. Dynamic models," vol. 39, no. 4, pp. 1333–1364, Oct. 2003.

[12] O. E. Barndorff-Nielsen and N. Shephard, "Integrated OU processes and non-Gaussian OU-based stochastic volatility models," *Scand. J. Statist.*, vol. 30, no. 2, pp. 277–295, Jun. 2003.

[13] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application.* Englewood Cliffs, N.J: Prentice-Hall, 1993.

[14] E. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, Jan. 1954.

[15] P. Last, C. Bahlke, M. Hering-Bertram, and L. Linsen, "Comprehensive analysis of automatic identification system (AIS) data in regard to vessel movement prediction," *The Journal of Navigation*, vol. 67, pp. 791–809, 9 2014.

[16] L. M. Millefiori, P. Braca, K. Bryan, and P. Willett, "Adaptive filtering of imprecisely time-stamped measurements with application to AIS networks," in *2015 18th International Conference on Information Fusion (FUSION)*, July 2015, pp. 359–365.

Document Data Sheet

Security Classification		*Project No.*
Document Serial No. CMRE-PR-2019-053	*Date of Issue* June 2019	*Total Pages* 8 pp.

Author(s)

Leonardo M. Millefiori, Paolo Braca, Gianfranco Arcieri

Title

Scalable distributed change detection and its application to maritime traffic

Abstract

Building on a novel methodology based on the Ornstein-Uhlenbeck (OU) process to perform accurate long-term predictions of future positions of ships at sea, we present a statistical approach to the detection of abrupt changes in the process parameter that represents the desired velocity of a ship. Proceeding from well-established change detection techniques, the proposed strategy is also computationally efficient and fit well with big data processing models and paradigms. We report results with a large real-world Automatic Identification System (AIS) data set collected by a network of terrestrial receivers in the Mediterranean Sea from June to August 2016.

Keywords

Statistical change detection, long-term target state prediction, AIS, big data, Apache Spark

Issuing Organization

NATO Science and Technology Organization
Centre for Maritime Research and Experimentation

Viale San Bartolomeo 400, 19126 La Spezia, Italy

[From N. America:
STO CMRE
Unit 31318, Box 19, APO AE 09613-1318]

Tel: +39 0187 527 361
Fax:+39 0187 527 700

E-mail: library@cmre.nato.int