SCIENCE AND TECHNOLOGY ORGANIZATION
CENTRE FOR MARITIME RESEARCH AND EXPERIMENTATION

Reprint Series **CMRE-PR-2019-122**

Mining maritime vessel traffic: promises, challenges, techniques

Luca Cazzanti, Giuliana Pallotta

June 2019

Originally published in:

OCEANS 2015, 18-21 May 2015, Genoa, Italy,
doi: 10.1109/OCEANS-Genova.2015.7271555

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.

Mining Maritime Vessel Traffic: Promises, Challenges, Techniques

Luca Cazzanti, Giuliana Pallotta

NATO STO Centre for Maritime Research and Experimentation (CMRE), La Spezia, Italy

Email: {Luca.Cazzanti, Giuliana.Pallotta}@cmre.nato.int

***Abstract*—This paper discusses machine learning and data mining approaches to analyzing maritime vessel traffic based on the Automated Information System (AIS). We review recent efforts to apply machine learning techniques to AIS data and put them in the context of the challenges posed by the need for both algorithmic performance generalization and interpretability of the results in real-world maritime Situational Awareness settings. We also present preliminary work on discovering and characterizing vessel stationary areas using an unsupervised spatial clustering algorithm.**

I. Introduction

The Automatic Identification System (AIS) was conceived primarily as a navigational safety tool for collision avoidance and for supporting vessel traffic services in ports and harbors [1]. The AIS is an international standard protocol that lets vessels communicate information about their identity and journey to other vessels and to ground-based stations. Vessels equipped with AIS systems periodically broadcast static information, which includes the Marine Mobile Service Identifier (MMSI), name, flag country, dimensions, ship type; and dynamic information. which includes position, speed-over-ground (SOG), course-over-ground (COG), next port of call [2]. Soon after the introduction of AIS in 2002, it became evident that the information in the transmitted messages could also support the Maritime Situational Awareness (MSA) activities of nations concerned with the security of their sea borders and with understanding the maritime activities within their territorial waters and globally [3].

In the past decade, maritime traffic, global compliance with the AIS standard, and new AIS base station installations worldwide have increased. The result has been larger and larger volumes of AIS data, which challenge the human operators of MSA systems with information overload. Recognizing the need to aid the operators cope with and make sense of the vast amounts of AIS data, researchers have studied how to apply computer-based techniques from machine learning and data mining to analyze the AIS data, extract the relevant information, and support the human analysts in their decision making. Today, automatically discovering and characterizing the activities of vessels through the use of machine learning and data mining techniques are key tasks for achieving MSA. A rich literature of published results now exists, demonstrating successful case studies in vessel traffic pattern discovery, route prediction, and anomaly detection [4]–[6].

At the same time, automatic vessel traffic analysis is maturing from the early stages, where demonstrating the applicability of individual techniques on specific problems is the primary goal, to studying more comprehensive data analytics systems that support a range of activities related to vessel traffic analysis for MSA [7]–[9]. Transitioning to real-world, operational systems poses challenges of data provenance, sampling, representation, scalability, generalization, and interpretability [10]. We defer the discussion of how Big Data technologies [11] can meet some of these challenges. We focus instead how machine learning approaches generalize to a broad range of real-world MSA scenarios and how their governing parameters and inferences maintain an interpretable link to physical and operational concepts. We discuss how scientists have partly addressed these generalization and interpretability and some possible future directions.

A particular problem of interest for MSA is detecting and characterizing vessel stationary areas, which can be thought of as small, geographically limited areas where one or more vessels proceed at slow speed. We present preliminary work on discovering stationary areas automatically based on the vessel behavior and show that both speed and direction could be used in a detection scheme for such areas.

II. Existing Machine Learning and Data Mining Approaches to Vessel Traffic Analysis

Existing machine learning approaches to vessel traffic analysis can be divided in two categories: point-based and trajectory-based. Point-based approaches assume that an underlying stochastic process generates the maritime traffic, and treat AIS messages as independent, identically-distributed (iid) samples of the hypothesized process. Geographically disjoint grid cells are assumed independent of each other, so that the relevant quantities from AIS messages originating within a given cell, such as number of messages in a cell, vessel velocities, etc. are statistically independent of the quantities in neighboring cells. The independence assumption in point-based approaches simplifies the counting, averaging, and binning tasks that form the core of density estimation and prediction algorithms in support of maritime traffic analysis. Under the point-based approach, a fundamental data structure

that supports machine learning algorithms is the geographical grid size.

A drawback of point-based approaches is that they forgo the possible performance gains that could be obtained by accounting for the spatial correlations between the AIS-provided locations of each vessel. Trajectory-based approaches partly address this drawback by first estimating each vessel's trajectory from the spatio-temporal distribution of that vessel's AIS message stream. This approach requires more complex operations and careful bookkeeping in the first stages of the analysis, but the trajectory objects produced by these early operations can provide a richer knowledge base for downstream analyses. Thus, in the trajectory-based approach, a fundamental data structure that supports machine learning algorithms is the vessel trajectory.

A. Point-based Approaches

Ristic [12] proposes a point-based approach to modeling the normal maritime traffic patterns in a geographical area, in support of anomaly detection. The area is subdivided in independent, non-overlapping cells, and the number of AIS messages transmitted from each cell is modeled as a Poisson point process. The corresponding distributions of vessel velocities, as reported by the incoming AIS messages in each cell, is estimated by kernel density estimation (KDE).

Given the localized, cell-specific models of baseline traffic, anomalies in the number of messages and in the velocity profile can be detected statistically. In that study, the number of messages is flagged as unusually high (or low) if it surpasses (or remains below) a threshold derived from the quantiles of the underlying Poisson process. Similarly, a quantile-derived velocity threshold, based on the within-grid velocity distribution estimated with KDE, discriminates the anomalous velocities from a credible range of acceptable values. The author demonstrates the applicability of this approach with a dataset of AIS messages comprising two weeks of test data and four months of training data from Sydney Harbour.

A study that models the maritime traffic in Port Adelaide [5] also adopts KDE to estimate the joint location and velocity densities and quantile-based thresholds for flagging anomalies. In that study, vessel trajectories originating from a common location form the training data. In this case, velocities and locations from raw AIS messages form feature vectors used for estimating multivariate densities. The feature vectors are assumed iid, so no correlation between subsequent AIS messages that form each vessel's trajectory is considered. Instead, trajectories are intended as sequences of AIS messages grouped by MMSI, and are used only to define traffic patterns originating from a common grid cell. The proposed approach is appropriate for ports, where traffic originates from well-defined locations.

Rhodes et al. [13] build on their earlier work in associative learning and neural networks applied to maritime surveillance [14], [15], and present an algorithm that learns normal traffic patterns in support of predicting future vessel locations within a temporal horizon of 15 minutes. The proposed algorithm uses *outstar learning* to learn associations between the current geographic location of a vessel and the likely future vessel positions, on a pre-defined grid. The stronger the association between a current and predicted cell, the more likely a vessel will occupy the predicted cell. The authors give a probabilistic interpretation of the cell associations as empirical conditional probabilities of future vessel locations conditioned on the current vessel location report, so the proposed technique does not rely on explicitly modeling the local densities.

A similar, but independent, point-based method [16] applies *association rule mining* [11], a classical data mining technique used in E-commerce, to the problem of estimating future vessel locations from current AIS data. The algorithm identifies frequently-occurring pairs of time-consecutive vessel locations, thereby learning the frequency of association between the grid cells. The association can then be used to predict future vessel locations conditionally on current locations. This approach is prone to producing geographically discontinuous vessel tracks, because it does not enforce any spatial continuity of the co-occurring locations. The authors address this issue by introducing a Markov transition matrix that smoothes the predicted vessel locations. Nonetheless, discontinuities may persist, as only the frequent location co-occurrences are identified.

B. Trajectory-based Approaches

A recent trajectory-based approach [8] uses similarity-based and kernel-based machine learning methods to cluster and classify vessel traffic and to detect anomalous vessel behavior. The approach first compresses the AIS messages from each vessel into piecewise linear trajectories using standard geometric operations. Then, kernels are defined based on appropriate similarity functions that measure how well any two given trajectories are aligned. These alignment-based trajectory similarities form the kernels that power trajectory clustering with kernel k-means, classification with support vector machines (SVMs), and anomaly detection with single-class SVMs.

The authors of that study explore the applicability of dynamic time warping and edit distances to the problem of measuring the similarity between trajectories defined in a multidimensional space by the locations and velocities of the vessels. These measures are based on the notion of geometrical alignment between piecewise linear (in the multidimensional space) trajectories. The authors extend these geometric alignment similarities to incorporate semantic similarity, which captures the similarity of the types of geographic areas visited by the vessels, as categorized in external knowledge bases. The authors show that the derived enriched similarity kernels can improve the clustering and classification, but anomaly detection does not benefit from the additional domain knowledge.

Another approach is Traffic Route Extraction for Anomaly Detection (TREAD) [7], which was co-developed by one of the authors of this article at the Centre for Maritime Research and Experimentation (CMRE). TREAD is a hybrid trajectory- and point-based approach that does not rely on fully

formed vessel trajectories, but at the same time considers the sequential ordering to the AIS messages from each vessel. For TREAD, the fundamental data structure that supports the learning algorithms is the vessel object, identified from the AIS messages by its MMSI, and described by its corresponding static and dynamic attributes contained in the AIS messages. TREAD is general framework for maritime traffic characterization that produces dictionary of historical vessel patterns-of-life, which can be used as prior information for activity-detection algorithms,

As a first step, TREAD uses the unsupervised spatial clustering algorithm DBSCAN [17], [18] to cluster *waypoints*, which are locations corresponding to stationary areas, entry, and exit points for a selected geographical region. The waypoints are linked together to form *route* objects, based on the AIS attributes of the vessels whose messages contributed to the formation of the waypoint clusters. The result of applying TREAD to a geographical area is a dictionary of waypoint and route objects, adorned with dynamic and static AIS properties. This enables further exploratory analysis of major routes and waypoints that is not limited to location, and instead spans informational attributes, such as flag state, direction, velocities, destination, etc. The resulting knowledge base can be used in support of MSA activities, such as anomaly detection. A distinguishing property of TREAD is that its clustering operates in an incremental mode, so that new data points can be assigned to existing clusters in real time and update the cluster definitions without extensive model retraining: the clusters can be merged, split, removed, generated as new data arrives. This approach incurs a greater cost in the early stages of the analysis, because of its bookkeeping complexity and its required integration with databases, but simplifies model training and updating when new data arrives.

III. Generalization and Interpretability: Performance Metrics, Hyperparameters, and Data Density

Researchers' interest in machine learning and data mining applied to maritime vessel traffic analysis has produced a range of approaches, published in the scientific literature. Taken individually, each of these published feasibility studies, concept demonstrations, or proposed frameworks has positively contributed to this field of research. As the research community progresses toward making these techniques operationally viable in the broader context of real-world MSA, it must contend with the challenges of generalization and interpretability.

Generalization is the ability to apply a proposed technique to a broad range of situations. In particular, a technique that generalizes well does not require extensive re-tuning of the relevant algorithms when it is applied to different data sets and supports a variety of operational scenarios for MSA.

Interpretability is the ability to relate the algorithmic performance metrics and the parameters of a proposed technique to given key performance indicators (KPIs) of operational effectiveness and to physical quantities. In particular, an interpretable technique aides the smooth handshaking between operations and experimentation, which favors the transition of research concepts to real-world scenarios.

Generalization and interpretability are interconnected through performance metrics, parameter tuning, and their dependency on the data sources. Consider the performance of unsupervised learning algorithms. A difficulty with unsupervised techniques is that there is no ground truth for assessing the correctness of the discovered patterns, a task left to the researcher, who often "eyeballs" the results. This is fruitful in the beginning stages of any analytic task, and indeed the typical role of unsupervised learning in a larger data analysis system is exploratory and to provide a first glimpse at the hidden structure in a dataset. At the same time, this injects subjectivity in the performance assessment that can make the results difficult to interpret by end-users who did not originate a particular approach, and the techniques difficult to generalize to a broad range of real-world operational settings.

One standard objective strategy to assess the performance is to create manually a reference dataset for comparison. This strategy enables using standard supervised learning performance metrics, such as precision, recall, false positive, false negative, area-under-the-curve (AUC), etc. , to assess the performance of an unsupervised learning task. This strategy was adopted to assess the performance of the kernel k-means clustering algorithm applied to a set of vessel trajectories extracted from AIS data gathered around Texel Island, Netherlands [8]. In that work, a set of 714 vessel trajectories was clustered into 8 hand-labeled clusters, and the adopted performance metric was the $F1$-score, a standard objective measure of performance in data mining tasks.

The lack of ground truth also arises in anomaly detection tasks, for instance when a vessel veers off-course, or proceeds too fast compared to normal behavior. As for clustering, one could manually label a reference data set, and use standard supervised learning performance metrics. This strategy was adopted in a vessel trajectory anomaly detection task documented in the same study as the clustering task [8]. Single-class SVMs were trained on 747 normal trajectories, and used to detected 39 hand-labeled anomalous ones. Precision was the adopted standard performance metric. In both clustering and anomaly detection tasks, a remaining, unaddressed difficulty is that manually labeling the data is a laborious, time consuming, error prone and subjective task that does not easily generalize to different and larger datasets: For larger datasets, manually labeling data may not be an options

Another strategy to address the lack of ground truth in anomaly detection is to apply statistical outlier detection techniques. The decision is based on a threshold of statistical significance derived from estimated probabilistic models of the data. Thus, there is no need for ground truth as a term of comparison: instead, the standard performance metrics probability of detection. P_d [12] or probability of false alarm P_{fa} [5] can be used in this case. This strategy is particularly well suited for unsupervised statistical learning approaches

where a probabilistic model of vessel traffic normalcy is first learned, and then used as a term of comparison for detecting anomalies.

A recent maritime traffic anomaly detection study [12] adopts the statistical outlier detection approach. In the study, the number of AIS messages and the vessel velocities observed at grid locations are assumed random variables and modeled probabilistically using kernel density estimation (KDE). New data were declared anomalous if the number of messages or the velocity values surpassed a statistical significance threshold of 0.01, meaning that the probability of detection is chosen at 1%. In the study, the strategy gave detection rates of 2.2 anomalies/hour and 4.8 anomalies/hour, which the author judges an improvement compared to typical human operator load.

Generalization and interpretability challenges also arise from the need to tune the parameters that govern the machine learning algorithms. For example,consider the unsupervised clustering algorithm DBSCAN. It can flexibly cluster spatial points without relying on a labeled set of training data and without a priori knowledge of the number of clusters. Two parameters influence the outcome: the radius of the neighborhood around a given point for evaluating cluster membership of candidate, neighboring points, and the minimum number of points required for a cluster. In turn, these parameters are sensitive to the density of the points in a data set: more spatially dense datasets require a smaller radius and a lower minimum number of points.

An approach to selecting the radius and the minimum number of points could be cross-validation, which is a widely used, empirical technique for selecting the optimal parameter values for machine learning algorithms by testing each parameter from a range of possible values. Unfortunately it has limited applicability to unsupervised algorithms like DBSCAN, because it requires a performance metric against which to measure the efficacy of the selected values. We have already explained that performance metrics can be difficult to define in unsupervised learning settings.

For DBSCAN an effective strategy to set the parameters has been to use expert domain knowledge. Our own experience working with the DBSCAN-based TREAD algorithm has been that by following common practices and guidelines from subject matter experts in maritime traffic monitoring TREAD produces interpretable results [7]. Other researchers [4] have taken the same approach and adopted commonly-used parameter values to identify fishing vessels from AIS trajectory-derived signatures. In that study, a DBSCAN-based algorithm was also employed.

Another approach to selecting the parameters is to adopt their theoretically optimal values based on a hypothesized mathematical model of the data. For example, theoretically optimal settings exist for the bandwidth parameter for KDE, and indeed in two maritime anomaly tasks, the optimal bandwidths were selected for n-dimensional Gaussian kernels [5], [12], thereby eliminating the need for cross-validation.

Grid cell size is another parameter that critically affects the performance of point-based approaches to maritime traffic analysis, which rely on counting the number of AIS messages in each grid cell. The cell size must be large enough to capture a sufficient number of AIS messages for building robust local statistical models, but if it grows too large, the predictive usefulness of the models could be weakened by the lower spatial resolution. This imposes practical limits on the application of point-based approaches to areas where AIS messages are sparse. Trajectory-based approaches do not rely on grids and instead form trajectories, so are less influenced by data sparsity, but not completely immune: Sufficiently dense sub-areas must exist within a region of interest to spatially cluster the waypoints that define a trajectory [7] or to form robust piecewise linear approximations of vessel trajectories [8].

For the above reasons, we expect that the optimal cell size for a given geographical region depends on the AIS message rate and coverage in that region, as demonstrated by the variability of chosen cell sizes in the literature: a study adopts a cell size of $0.1\,\mathrm{deg} \times 0.1\,\mathrm{deg}$ [16]; another study adopts a much finer grid of $0.002\,\mathrm{deg} \times 0.002\,\mathrm{deg}$ [12]. In those studies adopted size, like the other parameters, are set manually based on expert knowledge of the region of interest or driven by computational and operational constraints.

A strategy for automatically setting the grid size is to adopt a multi-scale grid, where the cells of different sizes cover an area of interest [15]. Under this strategy, larger cells cover areas where AIS messages are more sparse, and smaller cells cover areas where AIS messages are more dense. This is the typical case of port and harbor areas, where maritime traffic converges from the open sea, where traffic is more sparse. In a maritime traffic prediction study of the Port of Miami [19], four grids of different cell sizes where chosen, achieving a closer approximation to a desired level of recall performance than a single-size grid. Although the study does not articulate a fully automated parameter learning procedure, this strategy suggests that cell sizes can be adaptively chosen according to a desired performance goal.

From the above discussions a more general point emerges. Parameters that represent physical quantities relevant to maritime vessel traffic (i.e. cell size, neighborhood radius, data density) and that influence the effectiveness of machine learning approaches could be automatically chosen as a function of performance parameters. Conversely, given the parameters relevant to a geographic region or dictated by operational constraints, the corresponding bounds on the achievable performance could be estimated prior to carrying out computationally expensive machine learning experiments. We hypothesize that the relationships between performance and parameters could take the form of mathematical functions, table lookups, or well-documented "rules-of-thumb," and that data mining techniques for massive data sets [10], [11] can help discover these relationships. The availability of this knowledge will provide operators in real-world MSA scenarios and machine learning scientists in the laboratory with a common reference for effectively navigating the often murky waters of general-

ization and interpretability.

IV. Work in Progress: Discovering and Characterizing Stationary Areas and Vessels

A problem of particular relevance to MSA is discovering and characterizing vessel stationary areas, which are small, geographically circumscribed areas where vessels proceed at slow speed or are stopped. Stationary areas are particularly important because they often correspond to the location of strategic and logistical nodes in the maritime traffic network. Ports are obvious stationary areas, and harbour entrances, offshore platforms, traffic choke-points, anchorage areas are other examples.

When vessels proceed at slow speed or stop in locations that do not match the a priori knowledge of stationary areas, further analysis is required. It could be that the slow moving vessel is loitering in pursuit of illegal activities or is having engine trouble. Another possibility is that the newly discovered stationary area reflects an emerging local traffic pattern, perhaps as the a result of newly-established waiting areas near a port, or physiological changes in regional maritime traffic trajectories, for example due to seasonal weather patterns, piracy activity, etc.

In this work we focus on the exploratory aspect of the analysis of stationary areas. We use the TREAD algorithm with AIS data to discover the stopping areas from maritime traffic patterns in a given region, and to characterize the traffic patterns specifically of the vessels that visit the stationary areas.

Once derived, stationary areas can be intersected with the context information. Stationary areas which do not correspond to any known stationary areas are of particular interest. As an example, stationary areas which fall far from the coastline and do not match with the location of off-shore platforms, can potentially be either anchorage areas, fishing areas or loitering areas.

The first step is to assign the labels "vessel is sailing" or "vessel has stopped" to segments of each vessel track. This is achieved by Algorithm 1, which is a speed change point detector that operates on the available AIS data for each vessel in a given region of interest and for a given time window.

For each vessel of interest v, the velocity $v.avg\_speed$ is estimated from the reported AIS positions and timestamps. Simple thresholds on the speed determine if the vessel is moving or stopped (*i.e.* is in port). The detected stopped events are then clustered using incremental DBSCAN [18], which is embedded in TREAD, producing coherent spatial clusters of "vessel stopped" points: these are the discovered stationary areas.

Fig. 1 shows an example of discovered stationary areas in the Persian Gulf. For this analysis, we considered terrestrial and satellite AIS data from the MSSIS network, spanning the period of February 3, 2013 to May 7 2013. There were 12,051 unique vessels observed in the area. The most common vessel types were 35.2% cargoes, 20.8% tankers, and 5.2% tugs. The stationary areas were formed by clustering the contribution of 1,779 vessels which passed the speed gating threshold.

Algorithm 1 *Vessel Motion Status Detector*

Require: V, v // list of all the vessels entered in the area, V, and Vessel of interest, v with a given $MMSI$
Require: $v.track_{last24H}$, $v.SOG$, $SPEED_{min}$, $SPEED_{max}$, $sample_{width}$
while $v.Track_{last24H} > sample_{width}$ **do**
  $Dist(v.track(end - sample_{width}), v.track_{(end)}) \leftarrow \Delta S$
  $v.Time(end) - v.Time(end - sample_{width}) \leftarrow \Delta T$
  $\frac{\Delta S}{\Delta T} \leftarrow v.avg\_speed$ // observed average speed shown by the vessel
  if $v.avg\_speed > SPEED_{max}$ **then**
    $Vessel\ is\ sailing$
  else
    if $v.avg\_speed < SPEED_{min} | v.SOG < SPEED_{min}$ **then**
      $Vessel\ has\ stopped$
    end if
  end if
end while
return $Vessel\ Motion\ Status$

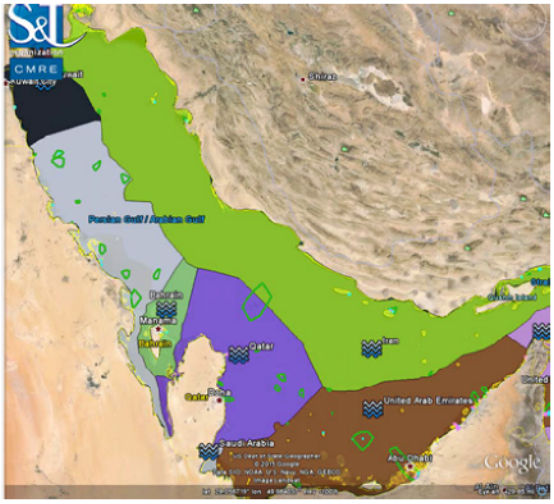

Fig. 1. Stopping areas, denoted by green polygons discovered in the Persian Gulf with TREAD. The polygons are overlaid with local Exclusive Economic Zones. Only the stationary areas a distance away form the shoreline are shown, so as to not overcrowd the figure with known ports.

The discovered stationary areas may be further characterized in terms of their directional characteristics. To do this, we consider the vessels that visited a given stationary area and study how the COG changes over time. The idea is that sailing vessels typically hold a steady course and therefore the reported COG in the AIS message is approximately constant. Conversely, when vessels are anchored or moving slowly within a confined area, the reported COG changes significantly.

As an example, we analysed the track reports form the 24 hours preceding a vessel ending a trip in a stationary area. Fig. 2 shows the track of a cargo vessel in the Persian Gulf, along with its kinematic features. As we can see, it is possible

to characterize the transition phases when the vessel stops or starts sailing again after stopping. COG changes which are more significant in absolute values and more persistent in time in correspondence of SOG values higher than $SPEED_{min} = 1knots$ but lower than $SPEED_{max} = 5knots$ which is a maximum threshold for speed provided by operators to identify loitering vessels.

This analysis has been focused on measurements received from AIS sensors but can be easily extended to data provided from other sensors (*e.g.*, coastal radars) from which the vessel kinematic features can be estimated.

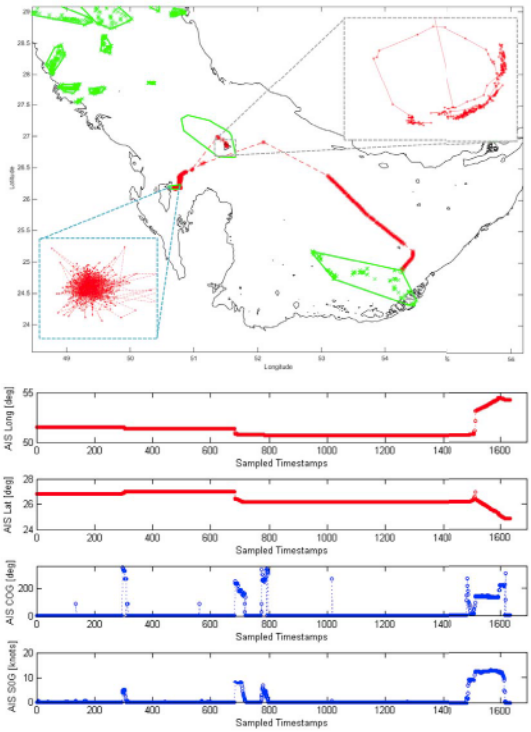

Fig. 2. Example AIS real track of a stationary vessel in the Persian Gulf (top), and its kinematic components (bottom). The COG changes more significantly in correspondence of SOG values lower than 5 knots.

V. Summary

We discussed the challenges posed by the need to apply machine learning and data mining techniques to maritime traffic vessel analysis. We highlighted how generalization and interpretability can help transition research to real-world operational scenarios. We presented preliminary work on detecting and characterizing stationary areas by using a spatial clustering approach.

Acknowledgements

This work was funded by NATO Allied Command Transformation (NATO-ACT) as part of the Data Knowledge Operational Effectiveness (DKOE) programme at CMRE.

References

[1] International Maritime Organization, "International Convention for the Safety of Life at Sea (SOLAS)."

[2] International Telecommunications Union, "Technical characteristics for an automatic identification system using time division multiple access in the vhf maritime mobile band (Recommendation ITU-R M.1371-4)," 2012.

[3] B. Tetreault, "Use of the automatic identification system (AIS) for maritime domain awareness (MDA)," in *OCEANS, 2005. Proceedings of MTS/IEEE*, Sept 2005, pp. 1590–1594 Vol. 2.

[4] F. Mazzarella, M. Vespe, D. Damalas, and G. Osio, "Discovering vessel activities at sea using AIS data: Mapping of fishing footprints," in *Information Fusion (FUSION), 2014 17th International Conference on*, July 2014, pp. 1–7.

[5] B. Ristic, B. La Scala, M. Morelande, and N. Gordon, "Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction," in *Information Fusion, 2008 11th International Conference on*, June 2008, pp. 1–7.

[6] K. Kowalska and L. Peel, "Maritime anomaly detection using Gaussian Process active learning," in *Information Fusion (FUSION), 2012 15th International Conference on*, July 2012, pp. 1164–1171.

[7] 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. [Online]. Available: http://www.mdpi.com/1099-4300/15/6/2218

[8] G. K. D. de Vries and M. van Someren, "Machine learning for vessel trajectories using compression, alignments and domain knowledge," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 426 – 13 439, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412007762

[9] N. L. Guillarme and X. Lerouvreur, "Unsupervised extraction of knowledge from s-ais data for maritime situational awareness," in *Information Fusion (FUSION), 2013 16th International Conference on*, July 2013, pp. 2025–2032.

[10] National Research Council, *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press, April 2013. [Online]. Available: /catalog/18374/frontiers-in-massive-data-analysis"

[11] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Data Sets*. Cambride University Press, November 2014.

[12] B. Ristic, "Detecting anomalies from a multitarget tracking output," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 50, no. 1, pp. 798–803, January 2014.

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

[14] N. A. Bomberger, B. J. Rhodes, M. Seibert, and A. M. Waxman, "Associative learning of vessel motion patterns for maritime situation awareness," in *Proc. Intl. Conf. on Information Fusion*, Florence, 2006, pp. 1–8.

[15] M. Zandipour, B. Rhodes, and N. Bomberger, "Probabilistic prediction of vessel motion at multiple spatial scales for maritime situation awareness," in *Information Fusion, 2008 11th International Conference on*, June 2008, pp. 1–6.

[16] F. Deng, S. Guo, Y. Deng, H. Chu, Q. Zhu, and F. Sun, "Vessel track information mining using ais data," in *Multisensor Fusion and Information Integration for Intelligent Systems (MFI), 2014 International Conference on*, Sept 2014, pp. 1–6.

[17] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR-US, 1996.

[18] M. Ester, H. Kriegel, J. Sander, M.Wimmer, and X. Xu, "Incremental clustering for mining in a data warehousing environment," in *24th International Conference on Very Large Data Bases*, New York, USA, 1998.

[19] B. J. Rhodes, N. A. Bomberger, M. Zandipour, L. H. Stolzar, D. Garagic, J. R. Dankert, and M. Seibert, *Anomaly Detection and Behavior Prediction: Higher-Level Fusion Based on Computational Neuroscientific Principles*, N. Milisavljevic, Ed. InTech, 2009.

Document Data Sheet

Security Classification		Project No.
Document Serial No.	**Date of Issue**	**Total Pages**
CMRE-PR-2019-122	June 2019	6 pp.

Author(s)

Luca Cazzanti, Giuliana Pallotta

Title

Mining maritime vessel traffic: promises, challenges, techniques

Abstract

This paper discusses machine learning and data mining approaches to analyzing maritime vessel traffic based on the Automated Information System (AIS). We review recent efforts to apply machine learning techniques to AIS data and put them in the context of the challenges posed by the need for both algorithmic performance generalization and interpretability of the results in real-world maritime Situational Awareness settings. We also present preliminary work on discovering and characterizing vessel stationary areas using an unsupervised spatial clustering algorithm.

Keywords

Trajectory, data mining, ports (computers), measurement, standards, machine learning algorithms

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