A data driven approach to maritime anomaly detection

Zissis D.[a,b], Chatzikokolakis K. [b], Vodas M. [b] , Spiliopoulos G. [b] and Bereta K. [b],
[a] University of the Aegean, Syros, Greece
[b] MarineTraffic, Athens, Greece;

ABSTRACT

The operational community has long identified anomaly detection systems as vital for increasing the effectiveness of maritime surveillance systems, since the huge quantities of data produced today quickly reduce their effectiveness in the field. The limited range of currently available maritime anomaly detection systems rely heavily on expert knowledge and hardcoded rules; thus, limiting their scope and effectiveness only to known situations and patterns of behaviour. By contrast, data driven systems learn from the data itself and can thus generalise well to new tasks and previously unseen situations. In this paper, we present an overview of the anomaly detection system developed in the context of the European Commission H2020 funded project BigDataOcean, and specifically the "Maritime Security and Anomaly Detection" pilot. Through a combination of unsupervised machine learning methods and behavioural analytics, the developed system is capable of i) automatically modelling shipping routes at a global scale; ii) constructing vessel specific profiles and class baselines; and iii) detecting deviations from patterns of normalcy in real time.

Keywords: Anomaly Detection, Machine Learning, Big Data, Maritime Situational Awareness

1. INTRODUCTION

Anomaly detection in the maritime domain has been identified by the operational community as an important aspect requiring intensified research efforts and development [1][2]. Commonly, surveillance operators have to search and predict emerging conflict situations, for example, potential collisions, dangerous vessels, or suspicious activities from a large number of vessels within geographical regions. Today there are more than 23 mandatory ship reporting systems (e.g. the Automatic Identification System) and numerous surveillance systems (e.g. coastal radars), which produce constant streams of high-speed and high-volume data while tracking vessels at sea. Exploring and monitoring the data manually is a demanding task, not only due to the complexity and heterogeneous nature of the data itself but also due to other factors like uncertainty, fatigue, cognitive overload, or other time constraints. Early detection of dangerous situations provides critical time to take appropriate action, possibly before potential incidents evolve [2]. Increasing automation through advanced machine learning and artificial intelligence methods enables the system and the operator to spot complex situations by correlating various events from all surveillance sensors and classify them into important incidents.

As such, Maritime Situational Awareness (MSA) involves developing the ability to identify patterns emerging within huge amounts of data, fused from various uncertain sources and generated from monitoring thousands of vessels a day, so as to act proactively and minimise the impact of possible threats. At the core of this process is data mining, an essential step in the process, consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data. Vessels conducting illegal behaviour often try to hide their intentions and follow a set of patterns depending on the activity they perform: deviation from standard routes, unexpected AIS activity, unexpected port arrival, close approach, and zone entry [3]. Within this context anomaly detection can be understood as a method that supports situational assessment by indicating objects and situations that, in some sense, deviate from the expected, known or "normal" behaviour.

Essential for effective anomaly detection is building an accurate model of normalcy. The understanding of the complex maritime environment and a vessel behaviour though, cannot be limited to simply connecting vessel positions as they travel across the seas. As such "Patterns of Life" are understood as observable human activities that can be described as patterns related to a specific action (e.g. fishing) taking place at a specified time and place. Essentially, vessel-based maritime activity can be described in space and time, while classified to a number of known activities at sea (e.g., fishing, dredging, etc.). The spatial element describes recognised areas where maritime activity takes place; thus, including ports, fishing grounds, offshore energy infrastructure, dredging areas and others. The transit paths to and from these areas also describe the spatial element, e.g. commercial shipping and ferry routes etc., while the temporal element often holds additional information for categorising these activities (e.g. fishing period, time of year).

This paper summarises the work completed in the context of the BigDataOcean H2020 project, which developed a fully-automatic method for first defining and then classifying vessel activities to "Patterns Of Life" in a data driven way, with no reliance on external information or need for a priori knowledge. The proposed framework is capable of dealing with big data volumes such as those produced in the maritime domain and specifically by the Automatic identification System at a global scale. In sum, the steps followed include first detecting port and activity areas, then trade routes and traffic patterns between these, before detecting deviations in true real time. The system has been developed in a step-wise approach and is currently in its final development phase. It relies on a number of state of the art and beyond state of the art frameworks for big data processing such as Spark[1] for batch and Akka[2] for stream processing, which are combined into a modified purpose built Lambda architecture [18]. The following sections present a short literature overview before describing the overall architecture and each element of the framework. In the final section a short discussion of results regarding real work incident detection is provided.

2. RELATED WORK

In recent years there has been a growing output of scientific papers and systems dealing with anomaly detection in the maritime domain and attempting to apply automated techniques to anomaly discovery. Throughout the literature anomalies are studied at different levels; both at a vessel information level (such as that provided in relevant registries and databases) and at a mobility (positional/trajectory) level. Often these two approaches are complementary.

In the first case, any available information about a vessel that can be found is used to build a behavioural profile for the ship, chartering company, etc. Such information is often static, unvarying or slowly changing over time. Two classes of solutions are dominant in this perspective; the ones relying on probabilistic risk assessment and the ones using fuzzy logic as a relaxation approach to the definite boundaries of probabilistic approaches. Probabilistic risk assessment has been introduced as a solution for the assessment of risk in the maritime domain in [6]. In [7] the authors applied a Bayesian simulation for the occurrence of situations with accident potential and a Bayesian multivariate regression analysis of the relationship between factors describing these situations and expert judgments of accident risk, to perform a full-scale assessment of risk and uncertainty. A fuzzy approach that evaluates the maritime risk assessment when applied to safety at sea and more particularly, the pollution prevention on the open sea is introduced in [8].

In the second case the focus is on mobility analytics. Some of the typical mining tasks in the spatiotemporal context include, frequent pattern discovery, trajectory pattern clustering, trajectory classification, forecasting, and outlier detection. Specifically, trajectory classification, includes constructing a model capable of predicting the class labels of moving objects based on their trajectories and other features [4]. The majority of works here focus on a data driven definition of normalcy from AIS data which then is used as baseline information to detect critical deviations. A number of publications rely on statistics to generate simple analytics of ship traffic and frequencies [5] [6]. More complex approaches can be categorised into (i) grid-based methods and (ii) methods of using vectorial representations of traffic. In grid-based approaches, the area of coverage is split into cells characterised by the motion properties of the crossing vessels to create a spatial grid (e.g., [16-21]). For example in [7], a data-driven methodology is proposed to estimate the vessel times of arrival in port areas. Grid-based anomaly detection algorithms include Fuzzy ARTMAP [8] , Holst Model [9], [10], Support Vector Machine and others. [11]. In the second category, vessel trajectories are modelled as a set of connected waypoints. Thus, vessel motions in large areas (e.g., at a global scale) can be managed thanks to the high compactness of the waypoint representation [12][13]. For example, the authors in [14] apply a Bayesian vessel prediction algorithm based on a Particle Filter (PF) on AIS data. The authors of [17] present an approach for detecting route deviations based on the Ornstein-Uhlenbeck (OU) mean-reverting stochastic process.

3. OVERALL APPROACH AND ARCHITECTURE

An efficient anomaly detection tool needs to support practitioners and operators along the complete lifecycle of monitoring situations at sea, from the observation of elements in the environment up to detection of anomalies and aids to planning. Following a number of dedicated workshops with representatives from the operational community a number of system requirements were identified, including but not limited to: (i) Providing a robust understanding of what to

[1] https://spark.apache.org/
[2] https://akka.io/docs/

expect in an area. (e.g. Patterns of Life), (ii) the system needs to prioritise information to the operator, filtering down data to a manageable cohort that may be considered for further investigation, using intelligent algorithms, (iii) real-time vessel risk profiling and activity classification should be provided, and (iv) anomaly detection should be performed in real-time, providing operators with justifiable alerts of critical unfolding incident and activities.

To fulfil this array of requirements we developed a number of components and novel algorithms, coupled together in a modified Lambda architecture. The architecture provides both support for batch processing of large volumes of spatiotemporal data and real time streaming capabilities. Each component uses the output of the previous step in an incremental fashion to produce its results. An overview of the workflow is given in the figure below

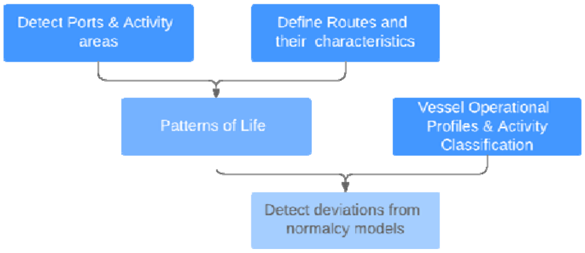
Detect Ports & Activity areas
Define Routes and their characteristics
Patterns of Life
Vessel Operational Profiles & Activity Classification
Detect deviations from normalcy models

Fig. 1 BigDataOcean framework for detecting deviations from normalcy

In the following sections we discuss the related parts of the framework which build up the holistic systems behaviour.

3.1 Port Area and Activity Detection

In our previous work, in collaboration with CMRE [15], [16], we developed a practical data driven implementation of a distributed method for calculating port operational areas from large amounts of spatiotemporal data. It is often an overlooked fact that seaport areas do not remain static over time; such spatial units often evolve according to environmental and seasonal patterns in size, connectivity, and operational capacity. Thus, accurately defining a seaport's exact location, operational boundaries, capacity, connectivity indicators, environmental impact and overall throughput is a challenging task. An adaptation of the well-known KDE algorithm to the map-reduce paradigm was developed and implemented on the distributed Spark framework so that it could cope with massive amounts of spatiotemporal data [15], [16]. In the context of the BigDataOcean project, the proposed approach was extended to other types of stationary areas besides ports: such as off-shore platforms, anchorage areas, and fishing grounds.

3.2 Route Definition

Sea going vessels follow specific paths when travelling across the vast blue ocean; these "roads" connecting major ports are some of the busiest places on earth, often only a few kilometres wide, scattered with many physical constraints (e.g., reefs), where enormous vessels perform risky manoeuvres under constantly changing environmental conditions (e.g., wind, sea currents). These waterways form a global maritime exchange network. More than often these connections are not direct lines (e.g., the shortest distance from the point of departure to destination), but "climatological routes" along which higher speeds can be achieved due to the existence of currents or the prevalence of wind, sea or swell. Sea roads though are not paved in concrete, as the location of the connector, its width and its content, can vary significantly over space and time, under the influence of various trade and carrier patterns, but also due to large infrastructure investments (e.g., canal expansions), climatic changes (e.g., global warming), new traffic restrictions (e.g., Emission Control Areas), political events and other international incidents (e.g., increase of piracy in specific regions).

In this context, a maritime big data modelling approach was developed, capable of accurately identifying the spatiotemporal dynamics of ship routes and most crucially their characteristics (such as route variable width, types of vessels, direction of travel etc.), thus deriving the maritime "patterns of life" at a global scale, without the reliance on any additional information sources or a priori knowledge. The proposed approach is based on the MapReduce programming model. The algorithm first filters and sorts all the positional data by assigning each position to a trip before "clustering" (reducing) each trip's data. In this sense the algorithm operates initially at a micro level, assigning data to trips, before developing a global network. The entire network can be produced on request in a few hours by using big data technologies (Spark and MapReduce) on a cluster of distributed computing nodes so as to depict any changes in the network due to external factors. In terms of performance and accuracy, experimental results on real world noisy datasets confirm an achieved overall accuracy of approximately 80%, while the entire processing duration is less than 3 hours for a terabyte of data.

3.3 Vessel operational profiles and activity classification

Every vessel has a unique Operational Profile. Each profile is a continuous learning model that employs machine learning to interpret behaviour in real-time. To construct the vessel profiles, we based our analysis on a number of features i.e., static vessel characteristics that, when combined, can give a representative view of a behaviour of a vessel. We observed that when the values of these features deviate considerably from the average values of the total vessel population, this can be an indication that the vessel could have an increased risk.

We collected data about vessels that have been involved in illegal activities, such as smuggling, illegal fishing or have been detained or even banned and we used them as ground truth. We observed that for a number of features, they deviate considerably from the respective average values for the vessel population of their category. Data driven methods are used to discover anomalies by learning statistical properties of the data and finding which data points do not fit.

Some of the features that we used are the following: Number of flag changes and name changes, number of transhipments, number of night port calls, number of port visits in countries with no anti-terrorism measures, the age of the vessel, number of days of coverage for a year, number of different countries visited, etc. In some cases, the importance of a feature makes sense when applied to the global fleet (e.g. number of flag changes or name changes) while in other the market segment plays significant role. For instance, number of transhipments may differ significantly for various vessel types (e.g., cargo or tankers) or completely irrelevant for other (e.g., passenger vessels or tugs). Thus, a subset of the features used have been calculated for each vessel type separately taking into account the different operations of shipping segments. Then, we employed a Random Forest Regressor to estimate the importance of each feature (shown in the figure below) and calculate the vessel's risk indicator.

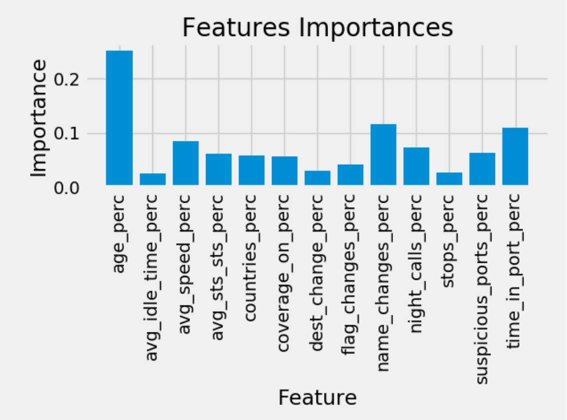

Fig. 2 Random Forest Regressor features' importance

3.4 Detecting deviations from normalcy models in real time

Time critical computations need to be completed in real time. Such systems have been described as systems which "control an environment by receiving data, processing it and returning the results sufficiently quickly to affect the environment at that time". In traditional system architectures the increasing volume of data ingested obviously negatively affects computation time. In our approach we introduce various architectural optimisations in order to achieve enhanced performance and low latency of computation. We have deployed a modified Lambda architecture, as this scheme allows the decoupling of batch processing (usually performed upon historical data) and real-time analysis, which typically exploits the knowledge extracted from the batch processing. The batch layer performs the analysis of historical positional data of vessels and extracts the required Patterns of Life. This is a long-running process which takes several hours to complete. Each Pattern of Life is a set of polygons that connect a departure port with a destination port for a specific vessel type. Taking into account the type of a vessel when producing the Patterns of Life is crucial, considering that vessels of different type may follow different routes for various reasons (e.g., due to vessel size limitations, sea depth, etc.). Once completed, the PoLs, together with the vessel operational profiles and activity classifications, are fed into the real-time layer in order to accommodate detection of vessel anomalies in real-time. Upon detection those incidents are displayed to the end user through the service layer. Furthermore, historical data are sent from this layer back to the batch layer at specific time intervals defined from the seasonality of the data, thus replacing previous processed patterns with new ones.

In the real-time layer, queries are performed on streaming and previously unseen data, thus enabling detection of security incidents in real time. More specifically, streaming vessel's positions and voyage-related AIS data are combined with static datasets and data mining models in real-time to detect anomalous events.

Route deviations: A route deviation is triggered when a ship is found travelling outside the "normal route" it is expected to follow in a given area and time, out of normal limits (e.g. speed patterns) or there is a mismatch in its type of activity reported for the given area or time period (e.g. fishing in a forbidden area). The first case (i.e., vessel travelling outside the "normal route") is triggered when a spatial intersection query of the vessel's position and the geometry of the PoL that the vessel should follow (based on its departure and destination ports) is empty. The second case (i.e., vessel travelling out of normal limits) is triggered when the first case is not triggered and in addition the vessel does not follow the PoL's statistical information in terms of SoG, CoG, Heading or Travelling Time. Finally, the third case is triggered when the vessel's identified activity (e.g. fishing) spatially intersects with an area that is semantically annotated as forbidden for that type of activity.

Proximity events: A proximity event is triggered when at least two vessels are detected sailing within close proximity to each other. This could be interpreted as an imminent collision, Ship to Ship Transfer or other. For each new message received at time *t* the real-time layer filters out former messages that are neither close spatially (i.e., messages with position more than 200 m away), nor temporally (i.e., messages with timestamp more than 10 minute before *t*), and projects the rest at time t based on their reported SoG and CoG. Then, for each of the projected positions we assume an area of probable movement (based on its SoG and CoG). If any of those areas intersects spatially with the area of probable movement of the vessel, then a proximity event is triggered.

Sailing in shallow/dangerous waters. Sailing in shallow waters are events that occur when a vessel is travelling in areas where the sea depth is less than the ship's draught. This information is correlated with bathymetry data and vessel's draught reported through AIS to increase the accuracy of detected groundings. Additionally, these events are fired when a ship is travelling in previously reported dangerous areas (such as high piracy areas etc.). In some cases, sailing in shallow waters or in dangerous areas would also mean that the vessel is sailing out of its Pattern of Life, and one or multiple route deviation events could also be triggered. Furthermore, when investigating potential groundings, relying only on bathymetry data may lead to poor accuracy, as most of the available dataset offer a coarse grain analysis of the sea depth providing sea depth values for large areas (i.e., several km^2) and do not take into account tides or seasonality that affect the sea depth. Thus, in order to increase the confidence level of our algorithm the real-time layer takes into account also the navigational status that is reported through AIS and vast decrease in vessel's SoG and CoG which may indicate that the vessel has ran aground.

Frequent or vessel specific AIS field Changes. Alerts are generated for vessels that highly exceed the number of frequent changes in the static parameters of AIS or for user defined vessels. The vessel may report through the navigation status field of AIS messages that is not be under command, or its ability to manoeuvre is restricted, or even explicitly report it is aground; thus an alert is generated.

Loss of AIS signal. A loss of AIS signal happens when the service stops receiving data from a ship whilst the ship is within the network coverage of AIS receivers, suggesting a possible "dark target". The reporting rate of an AIS transponder varies from 2 seconds to 3 minutes depending on the class of the transponder, the vessel's speed and changes of speed and/or course. Thus, using a time threshold the real-time layer can roughly estimate when to expect a message from each vessel and based on its last known position, SoG and CoG it can also roughly estimate where to expect that message. Thus, if the estimated location is within network coverage (i.e., we have received messages from other vessels in the vicinity) at the estimated timestamp and no message from the vessel has been received, the real-time layer yields a Loss of AIS Signal event.

Vessels of Interest (high risk ships). Ships that have been previously identified as High Risk or Vessels of Interest, trigger a number of alerts as they visit ports, user defined areas or conduct activities at sea. These are displayed on a map as a layer providing insights to end-users for potential illegal activities.

4. CONCLUSION

Situational awareness and anomaly detection tools are of outmost importance for the operational community in the maritime domain. In the context of the BigDataOcean project we developed a framework and system prototype capable of detecting maritime anomalies at a global scale in a data driven way, thus without the requirement of user defined rules or a priori knowledge. This paper provides an overview of the work conducted in the context of this project, highlighting important findings and breakthroughs.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732310.

REFERENCES

[1] E. Martineau and J. Roy, “Maritime Anomaly Detection: Domain Introduction and Review of Selected Literature,” Defence Research and Development Canada, 2011.

[2] M. Riveiro, G. Pallotta, and M. Vespe, “Maritime anomaly detection: A review,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, no. 5, p. e1266, 2018.

[3] “Maritime Anomaly Detection Based on Mean-Reverting Stochastic Processes Applied to a Real-World Scenario - Semantic Scholar.” . Available: https://www.semanticscholar.org/paper/Maritime-Anomaly-Detection-Based-on-Mean-Reverting-d’Afflisio-Braca/9606a9958d70ce5686e3c285dd07b3681fd604b0. [Accessed: 16-Mar-2019].

[4] J.-G. Lee, J. Han, X. Li, and H. Gonzalez, “TraClass: Trajectory Classification Using Hierarchical Region-based and Trajectory-based Clustering,” Proc VLDB Endow, vol. 1, no. 1, pp. 1081–1094, Aug. 2008.

[5] L. Zhang, Q. Meng, and T. Fang Fwa, “Big AIS data based spatial-temporal analyses of ship traffic in Singapore port waters,” Transp. Res. Part E Logist. Transp. Rev., Aug. 2017.

[6] Q. Meng, J. Weng, and S. Li, “Analysis with Automatic Identification System Data of Vessel Traffic Characteristics in the Singapore Strait,” Transp. Res. Rec. J. Transp. Res. Board, vol. 2426, pp. 33–43, Sep. 2014.

[7] A. Alessandrini, F. Mazzarella, and M. Vespe, “Estimated Time of Arrival using Historical Vessel Tracking Data,” IEEE Trans. Intell. Transp. Syst., pp. 1–9, 2018.

[8] N. A. Bomberger, B. J. Rhodes, M. Seibert, and A. M. Waxman, “Associative Learning of Vessel Motion Patterns for Maritime Situation Awareness,” in 2006 9th International Conference on Information Fusion, 2006, pp. 1–8.

[9] R. Laxhammar, “Anomaly detection for sea surveillance,” in 2008 11th International Conference on Information Fusion, 2008, pp. 1–8.

[10] 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, 2008, pp. 1–7.

[11] L. Wu, Y. Xu, Q. Wang, F. Wang, and Z. Xu, “Mapping Global Shipping Density from AIS Data,” J. Navig., vol. 70, no. 01, pp. 67–81, Jan. 2017.

[12] Y. Li, R. W. Liu, J. Liu, Y. Huang, B. Hu, and K. Wang, “Trajectory compression-guided visualization of spatio-temporal AIS vessel density,” in 2016 8th International Conference on Wireless Communications & Signal Processing (WCSP), 2016, pp. 1–5.

[13] M. Fiorini, A. Capata, and D. D. Bloisi, “AIS Data Visualization for Maritime Spatial Planning (MSP),” Int. J. E-Navig. Marit. Econ., vol. 5, pp. 45–60, 2016.

[14] 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), 2015, pp. 1–6.

[15] 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 IEEE Int. Conference on Big Data (Big Data), 2016, pp. 1627–1632.

[16] L. M. Millefiori, D. Zissis, L. Cazzanti, and G. Arcieri, “Scalable and Distributed Sea Port Operational Areas Estimation from AIS Data,” in 16th International Conference on Data Mining Workshops (ICDMW), 2016, pp. 374–381.

[17] E., P. Braca, L. M. Millefiori, P. Willett: Detecting Anomalous Deviations From Standard Maritime Routes Using the Ornstein-Uhlenbeck Process. IEEE Trans. Signal Processing 66(24): 6474-6487 (2018)

[18] Amazon Web Services, Lambda Architecture for Batch and Stream Processing on AWS, May 2015