SCIENCE AND TECHNOLOGY ORGANIZATION
CENTRE FOR MARITIME RESEARCH AND EXPERIMENTATION

Conference Proceedings **CMRE-CP-2018-002**

Proceedings of the Maritime Big Data Workshop

Elena Camossi, Anne-Laure Jousselme

January 2019

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.

CMRE-CP-2018-002

Proceedings of the Maritime Big Data Workshop

Elena Camossi, Anne-Laure Jousselme

This document, which describes work performed under the Project/the Programme Data Knowledge & Operational Effectiveness of the STO-CMRE Programme of Work, has been approved by the Director.

Intentionally blank page

CMRE-CP-2018-002

Proceedings of the Maritime Big Data Workshop

Elena Camossi, Anne-Laure Jousselme

Executive Summary: The NATO STO Centre for Maritime Research and Experimentation, as part of its mission to put forward the technological maritime research, with the support of the European Union's Horizon 2020 Programme has organized on May 9-10, 2018, the Maritime Big Data Workshop (MBDW). The workshop gathered together researchers, technological providers and members of the operational community to exchange their experience on Big Data innovations for maritime security, safety and security of maritime navigation and transport, sustainable fisheries and exploitation of ocean resources. For two days, 37 researchers and experts from Brazil, Canada, France, Germany, Greece, Italy, Portugal, South Africa, and Vietnam presented their work and main findings on Maritime and Big Data, including the outcomes of 6 ongoing Maritime Big Data projects and initiatives funded by the European Union: datAcron, MARISA, Ranger, EUCISE, AtlantOS, EMODnet.

The workshop's results enable to draw some preliminary conclusions on the current research and developments in Maritime Big Data. There is a general interest towards concrete societal and operational needs, coupled with an emerging tendency to develop methods combining heterogeneous, potentially complementary, information streams (mainly AIS, paired with SAR, Radar, METOC, acoustic), with an increasing attention towards source quality. The approaches adopted come from different areas of research, mainly machine learning and data mining, incorporating also techniques developed in Information and data fusion, but also data warehouse and online analytical processing.

The current trend towards experimenting open source Big Data technologies is challenged by the integration of diversified sources of information, which comes with an increased exigence of enhanced data management capabilities for harmonised data sharing and processing that can overcome the sole exploitation of kinematic data. Meanwhile, there is a prevailing requirement to reduce the uncertainty of detection and prediction results, entailing the development of capabilities to formally handle information and source quality. Analogously, the emergence of novel Artificial Intelligence approaches that, despite showing promising results, challenge results' interpretation, requires an increased involvement of experts in all the phases of the development (the so-called ``Human in the loop"), and the holistic incorporation of approaches addressing human factors' aspects.

Intentionally blank page

CMRE-CP-2018-002

Proceedings of the Maritime Big Data Workshop

Elena Camossi, Anne-Laure Jousselme

Abstract: The NATO STO Centre for Maritime Research and Experimentation, as part of its mission to put forward the technological maritime research, with the support of the European Union's Horizon 2020 Programme has organized on May 9-10, 2018, the Maritime Big Data Workshop (MBDW). The workshop gathered together researchers, technological providers and members of the operational community to exchange their experience on Big Data innovations for maritime security, safety and security of maritime navigation and transport, sustainable fisheries and exploitation of ocean resources. For two days, 37 researchers and experts from Brazil, Canada, France, Germany, Greece, Italy, Portugal, South Africa, and Vietnam presented their work and main findings on Maritime and Big Data, including the outcomes of 6 ongoing Maritime Big Data projects and initiatives funded by the European Union: datAcron, MARISA, Ranger, EUCISE, AtlantOS, EMODnet.

The workshop's results enable to draw some preliminary conclusions on the current research and developments in Maritime Big Data. There is a general interest towards concrete societal and operational needs, coupled with an emerging tendency to develop methods combining heterogeneous, potentially complementary, information streams (mainly AIS, paired with SAR, Radar, METOC, acoustic), with an increasing attention towards source quality. The approaches adopted come from different areas of research, mainly machine learning and data mining, incorporating also techniques developed in Information and data fusion, but also data warehouse and online analytical processing.

The current trend towards experimenting open source Big Data technologies is challenged by the integration of diversified sources of information, which comes with an increased exigence of enhanced data management capabilities for harmonised data sharing and processing that can overcome the sole exploitation of kinematic data. Meanwhile, there is a prevailing requirement to reduce the uncertainty of detection and prediction results, entailing the development of capabilities to formally handle information and source quality. Analogously, the emergence of novel Artificial Intelligence approaches that, despite showing promising results, challenge results' interpretation, requires an increased involvement of experts in all the phases of the development (the so-called ``Human in the loop"), and the holistic incorporation of approaches addressing human factors' aspects.

Keywords: Maritime Big Data, , Maritime sensors networks, Maritime Intelligent Surveillance and Reconnaissance, Maritime Situational Awareness, Maritime Interoperability, Maritime Information Fusion, Maritime Cyber Security, Human factors, Maritime Open Data, Efficiency of Navigation, Sustainable fisheries

Contents

CMRE-CP-2018-002

1
Editorial Note

The increase of the global maritime traffic and of the activities exploiting the ocean resources have led to the development of a series of technological innovations aiming at ensuring the safety and security of maritime navigation and guaranteeing the "blue growth", i.e., the growth of the maritime sector, to be sustainable and inclusive. These include the development of automated monitoring systems and maritime sensors networks, boosting the automation technologies now tested for autonomous cargo shipping, which have produced a tremendous increase of the traffic and environmental data available and opened up new avenues to science-driven maritime operations and policy making. According to a report published by the European Maritime Security Agency in 2017, European waters are navigated daily by some 12,000 vessels, which share their positions to avoid collisions, generating a total of about 200,000,000 positional messages every month, which are received by 700 coastal stations in Europe. The National Coastguards and Navy officers constantly monitor the maritime traffic in European waters and analyse these messages, cooperating to detect potential threats and to adopt the necessary safety procedures in case of accidents.

The NATO STO Centre for Maritime Research and Experimentation, as part of its mission to put forward the technological maritime research, with the support of the European Union's Horizon 2020 Programme has organized on May 9-10, 2018, the Maritime Big Data Workshop (MBDW). The workshop gathered together researchers, technological providers and members of the operational community to exchange their experience on Big Data innovations for maritime security, safety and security of maritime navigation and transport, sustainable fisheries and exploitation of ocean resources. For two days, 37 researchers and experts from Brazil, Canada, France, Germany, Greece, Italy, Portugal, South Africa, and Vietnam presented their work and main findings on Maritime and Big Data, including the outcomes of 6 ongoing Maritime Big Data projects and initiatives funded by the European Union: datAcron, MARISA, Ranger, EUCISE, AtlantOS, EMODnet.

The workshop's results enable to draw some preliminary conclusions on the current research and developments in Maritime Big Data. There is a general interest towards concrete societal and operational needs, coupled with an emerging tendency to develop methods combining heterogeneous, potentially complementary, information streams (mainly AIS, paired with SAR, Radar, METOC, acoustic), with an increasing attention towards source quality. The approaches adopted come from different areas of research, mainly machine learning and data mining, incorporating also techniques developed in Information and data fusion, but also data warehouse and online analytical processing.

CMRE-CP-2018-002

The current trend towards experimenting open source Big Data technologies is challenged by the integration of diversified sources of information, which comes with an increased exigence of enhanced data management capabilities for harmonised data sharing and processing that can overcome the sole exploitation of kinematic data. Meanwhile, there is a prevailing requirement to reduce the uncertainty of detection and prediction results, entailing the development of capabilities to formally handle information and source quality. Analogously, the emergence of novel Artificial Intelligence approaches that, despite showing promising results, challenge results' interpretation, requires an increased involvement of experts in all the phases of the development (the so-called ``Human in the loop"), and the holistic incorporation of approaches addressing human factors' aspects.

The BigDataOcean maritime anomaly detection service

Dimitris Zissis

University of the Aegean and Marine Traffic

1. Project Overview

The main objective of H2020 BigDataOcean[1] EU project is to enable maritime big data scenarios for EU-based companies, organisations and scientists, through a multi-segment platform that will combine data of different velocity, variety and volume under an inter linked, trusted, multilingual engine to produce a big-data repository of value and veracity back to the participants and local communities. BigDataOcean aims to capitalise on existing modern technological breakthroughs in the areas of the big data driven economy, and roll out a completely new value chain of interrelated data streams coming from diverse sectors and languages and residing on cross technology innovations being delivered in different formats (as well in different states, e.g. structured/unstructured, real-time/batches) in order to revolutionise the way maritime-related industries work, showcasing a huge and realistic economic, societal and environmental impact, achieved by introducing an economy of knowledge into a traditional sector. The main outputs of the project are novel services and applications that allow maritime-related industries, organisations and stakeholders in general to generate more (a) factual and evidence-based analytics, (b) decision support models, and (c) new business services focused on real-time collaboration, knowledge sharing amongst the key stakeholders, based on both (i) real-time data streams taking into consideration the data and temporal granularity aspects, and (ii) on batch processed data to extract analytics and intelligence to influence strategic mid-term and long-term operations planning. BigDataOcean aims to constitute the central node of a dynamic and expanding network.

The BDO consortium consists of 10 partners including industrial partners, research institutes and technology providers and integrators. BDO partners include: NTUA (coordinator), UBITECH, MarineTraffic, ANEK, Foinikas, ISMB, HCMR, R&D Nester, Universitat Bonn, UNINOVA. The entire project is driven by four industrial pilots, exploring four major use cases regarding big data applications in the maritime environment: one referring to security and anomaly detection; one referring to maritime environment protection; one referring to electrical power productions through waves' energy; and one referring to vessels' fault prediction and proactive maintenance.

2. Anomaly Detection service on the H2020 BigDataOcean Platform

Maritime Domain Awareness (MDA) is the effective understanding of activities, events and threats in the maritime environment that could impact global safety, security, economic activity or the environment. Recent advancements in Information and Communications Technologies (ICT) have created opportunities for increasing MDA, through better monitoring and understanding of vessel movements. While in the past, surveillance had suffered from a lack of data, current tracking technologies have transformed the problem into one of an overabundance of information, leading to a need for automated analysis. The major challenge faced today by the security domain is developing the ability to identify patterns emerging within huge amounts of data, fused from various sources and generated from monitoring thousands of vessels a day, so as to act proactively to minimize the impact of possible threats. The understanding of the complex maritime environment though, can never be limited to simply adding up and connecting together various vessel positions as they travel across the seas. Achieving situational awareness, perceiving and comprehending elements and their contextual meaning in the environment within a given volume of time and space, while projecting their status into a future timeframe, is a critical element of Maritime Domain Awareness.

[1] http://www.bigdataocean.eu

Anomaly detection can be defined as a method that supports situation assessment process by indicating objects and situations that, in some sense, deviate from the expected, known or “normal” behaviour and thus may be of interest for further investigation. Specifically, the goal is to use the BigDataOcean platform to analyse vessels’ and voyages’ characteristics and classify the diverse types of anomalies in maritime shipping to either static or dynamic, to act proactively and minimise threats at sea (for more details [1], [2]).

Static anomalies are related to unforeseen changes or mismatching of vessel’s identity information (e.g. IMO, MMSI, vessel’s name, etc.), while dynamic anomalies are mostly related to abnormal voyage behaviour (e.g. stop at sea, deviation from route, etc.). Vessel behaviour can be defined as the sum of all characteristics defining vessels movement, such as vessel position, course, heading and speed, observed over a given period. By definition, a pattern is composed of recurring events that repeat in a predictable manner. Vessel behaviour monitored over a long period of time gives insights into the navigational patterns followed by each vessel on specific routes. On the other hand, behaviour and pattern analysis of vessel profiles exploiting vessel movement information has been used to identify common routes of vessels traveling on the same itinerary and thus, determine dynamic anomalies (e.g. stops at sea, deviation from the common route, unforeseen speed changes, etc.). Our approach for dynamic anomaly detection is based on i) first defining these patterns and then ii) detecting deviations from these.

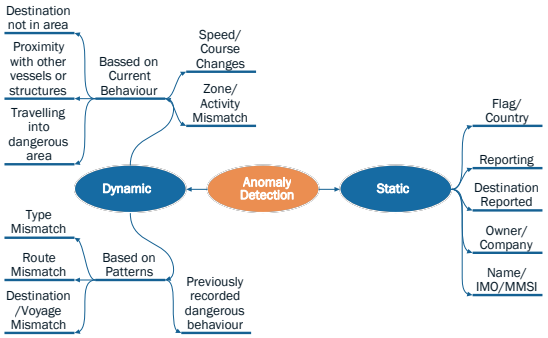


Figure 1. Maritime vessel anomaly as defined in the H2020 BigDataOcean project

To achieve this, we adopt a distributed clustering approach, which is based on the MapReduce paradigm, capable of processing terabytes of maritime spatiotemporal data on a cluster of computers (more details on the approach can be found in [3], [4]). This approach makes it possible to extract common behavioural patterns at scale which are then used to identify deviations from these and can assist in predicting the future location of a vessel (Figure 2). The main datasets used for this part of the work, include data collected from the Automatic Identification System (AIS) [5], combined with the data from the World Port Index[2].

[2] https://msi.nga.mil/NGAPortal/MSI.portal?_nfpb=true&_pageLabel=msi_portal_page_62&pubCode=0015

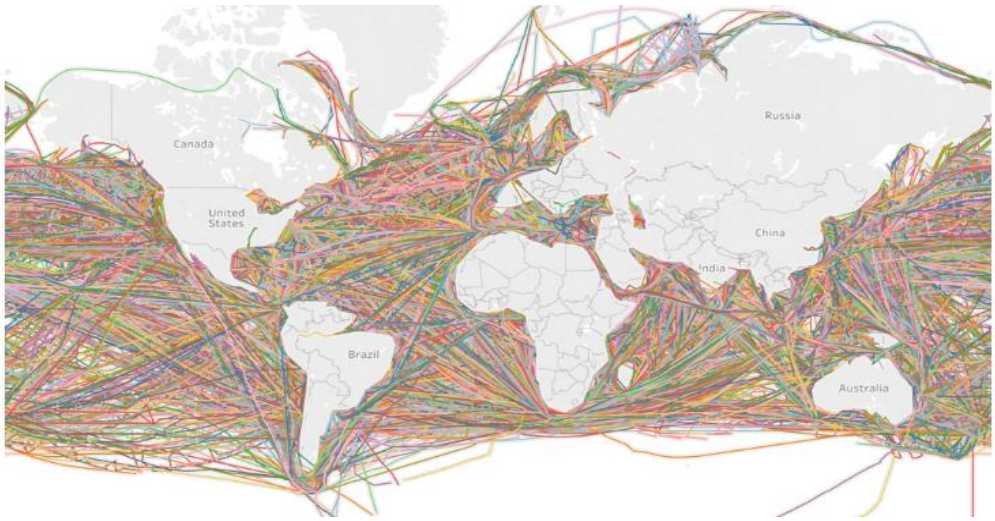


Figure 2. Global port to port connections extracted from global AIS coverage for tankers and cargo vessels as collected by MarineTraffic.com in 2016.

We are able to create a vast network of port to port connections, with calculated lower and upper confidence (or prediction) bounds. For example, with 95% calculated interval, there is a 95% chance that unseen or new observations fall within the lower and upper prediction bounds (Figure 3). These extracted "roads of the seas" define a network across which normal operating vessel positions are distributed.

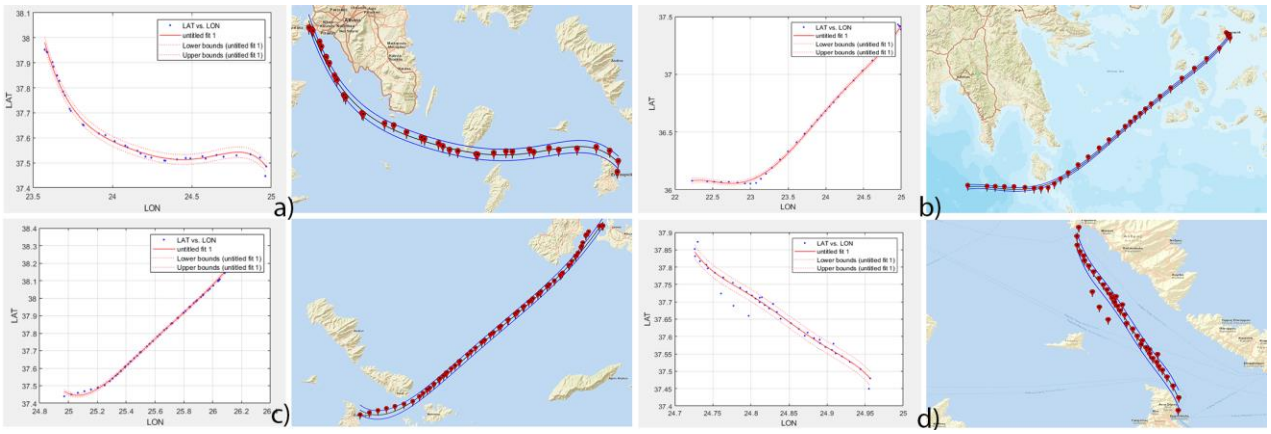


Figure 3. Example Set of extracted route patterns (curve plots and map visualisations) in the Aegean Sea with calculated confidence intervals based on popular ferry trips for 2016. a) Island of Syros to Piraeus; b) Island of Syros to Mediterranean; c) Syros to Cesme; and finally, d) Syros to Andros

3. Discussion and Future Work

The scale at which maritime surveillance data is now collected requires automated processing techniques capable of dealing with the growing volume and velocity of the data if it is to be effectively and efficiently utilized. In this work we briefly outline part of the work performed in the H2020 BigDataOcean project, specifically in the anomaly detection pilot use case, where normal patterns of vessel behaviour are constructed so as to be used as an indicator of "normalcy" from which outliers can then be detected. The traffic route extraction process is based on a distributed data processing approach, performed on large quantities of AIS and port index data. Future work includes automated methods for automatically analysing satellite Synthetic Aperture Radar (SAR) images in the case of a detected anomaly at a specific location and time.

4. Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732310, by Microsoft Research through a Microsoft Azure for Research Grant and through an Amazon Web Services (AWS) for Research Grant.

5. References

[1] J. Ferreira *et al.*, "Maritime data technology landscape and value chain exploiting oceans of data for maritime applications," in *2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 2017, pp. 1113–1122.

[2] E. K. Xidias and D. Zissis, "Adaptive neuro fuzzy inference system for vessel position forecasting," in *2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 2017, pp. 1071–1075.

[3] G. Spiliopoulos, D. Zissis, and K. Chatzikokolakis, "A Big Data Driven Approach to Extracting Global Trade Patterns," in *Mobility Analytics for Spatio-Temporal and Social Data*, 2017, pp. 109–121.

[4] G. Spiliopoulos *et al.*, "Knowledge extraction from maritime spatiotemporal data: An evaluation of clustering algorithms on Big Data," in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 1682–1687.

[5] International Telecommunication Union, "M.1371 : Technical characteristics for an automatic identification system using time-division multiple access in the VHF maritime mobile band," M.1371-5.

The European Marine Observation and Data Network (EMODnet) and big data for Blue Growth

Alessandro Pititto[1] and Antonio Novellino[2]

[1] Cogea srl, Rome, Italy
[2] ETT S.p.A, Genoa,Italy

Abstract. The European Marine Observation and Data Network is an initiative funded by the EU Commission to ensure rapid access to reliable and accurate geospatial information through multi-resolution maps of all Europe's seas and oceans, spanning seven disciplinary themes (bathymetry, geology, seabed habitats, chemistry, biology, physics, and human activities). EMODnet Human Activities has been mandated to create vessel density map of EU waters by processing AIS data. EMODnet Physics provides access to near real time and validated in situ collections of data and creates products on physical parameters of the sea.

Keywords: EMODnet, marine environmental data, big data, AIS data, physical oceanography

1 Introduction

The objective of the European Marine Observation and Data Network is to provide EU Member States and marine data users with the data and information required to make judicious decisions that concern environmental management and business. There is an urgent need for more detailed scientific knowledge of the largest biosphere on the planet and the seabed resources of the oceans, as well as the use of coastal and marine areas. These are important to us not alone for reasons of monitory prosperity but more importantly to protect one of the major life systems on Earth. Our ability to make sustained observations over large areas is currently one of the main limitations for major advances in our understanding of the oceans. Continuous, long-term measurements of physical, chemical, geological, and biological variables in the oceans and the seafloor are required to understand trends and cyclic changes. Enhanced capabilities for making sustained measurements of the ocean will open up new research opportunities and lead to improved detection and forecasting of environmental changes and their effects on biodiversity, coastal ecosystems, and climate. These advances will provide the tools for improved management of ocean resources such as fisheries, and better informed decisions on the use of the coastal zone for recreation, development, and commerce.

2 The European Marine Observation and Data Network (EMODnet)

EMODnet[1] is a long-term marine data initiative from the EU Commission Directorate-General for Maritime Affairs and Fisheries (DG MARE) involving more than 150 organisations for assembling marine data, products, and metadata. It has been developed through a step-wise approach and is currently in its third and final development phase. The organisations involved work together to observe the sea, process the data according to international standards and make that information freely available as interoperable data layers and data products. Unfortunately, marine data collection, storage and access in Europe has been carried out in a fragmented way for many years. Most data collection has focused on meeting the needs of a single purpose by a wide range of private and public organisations, often in isolation from each other. EMODnet provides access to European marine data across seven themes: bathymetry[2], geology[3], seabed habitats[4], chemistry[5], biology[6], physics[7], and human activities[8]. For each theme, EMODnet has created a gateway to a range of data archives managed by local, national, regional and international organisations. Users have free access to standardised observations, data quality indicators and processed data products, such as basin-scale maps.

2.1 EMODnet Human Activities and vessel density maps

EMODnet Human Activities has been mandated to create vessel density maps that can show vessel movement patterns across EU waters.

Fig. 1. Polygon representing the Area of Interest for the vessel density map.

To do so, a set of both satellite and terrestrial AIS data for the year 2017 was purchased from a commercial provider. Considering its geographic and time extent, the dataset contains an overwhelming amount of records, given that the default transmit

rate is every few seconds for Class A AIS, and every 30 seconds for Class B AIS. Therefore, it was necessary to clean and pre-process the data, before start working on the maps: the AIS messages used are only the ones relevant to assess shipping activities (1, 2, 3, 18 and 19) and they were down-sampled to 3 minutes, meaning that two consecutive positions from the same ship are at an interval of 3 minutes – albeit useful for several other purposes, shorter intervals may not be necessary when it comes to calculate a monthly average of density. Duplicate signals were removed; wrong MMSI signals were removed; wrong IMO numbers were corrected; special characters and diacritics were removed; signals with erroneous speed over ground (negative values or speed higher than 80 knots) and/or course over ground (negative values or more than 360°) were removed; S-AIS noise was removed by using Kalman filters; footprint filtering was performed for S-AIS data consistency.

At the time of writing, the pre-processing of data is still ongoing. Afterwards, density will be measured by counting ship positions in grid cells (1 sq km) at fixed time intervals. Density here is intended as a measure of the average number of ships one may expect to find in the unit area (the grid cell). The final result will consist of a series of density maps (by month, ship type, and tonnage class) with the average number of vessels per grid cell. Cells with higher density will have darker colours and vice versa.

The maps and the raster files will be made freely available on the EMODnet Human Activities portal, and users will be allowed to view, download, use and re-use them, whatever their purpose. It is believed that users might use this type of information for multiple purposes, commercial and non-commercial alike. In the future, it is envisaged to enrich the AIS data with more information on ships from sources such as the Lloyd's Register. This would make it possible to obtain more information on e.g. vessels' engine and fuel burnt, so as to make it possible to estimate shipping pollution, noise, impact on seabed, etc. With data from other EMODnet portals, it should also be possible to refine emission dispersion modelling and to compare the findings of this exercise with the actual emissions recorded by monitoring stations.

2.2 EMODnet Physics: a horizontal platform serving blue growth

EMODnet Physics is a domain specific portal of portals aggregating data and metadata from several data portals. A combined array of services and functionalities are offered to internal and external users, such as facility for viewing and downloading, dashboard reporting and machine-to-machine communication services, to obtain free-of-charge data, metadata and data products on the physical conditions of the ocean from many different distributed data sets.

The acquisition of physical parameters is largely an automated process that allows the dissemination of near real-time information. In particular, EMODnet Physics is a stock-share portal strongly federated to the Copernicus Marine Environment Monitoring Service In Situ Thematic Assembly Center. Historical validated datasets are organised in collaboration with SeaDataNet and its network of National Oceanographic Data Centres.

The EMODnet Physics portal is currently providing easy access to metadata, data and products of: wave height and period; temperature and salinity of the water column; wind speed and direction; horizontal velocity of the water column; light attenuation; sea ice coverage and sea level trends (relative and absolute). Lately, EMODnet Physics started working on river runoff data, total suspended matter and underwater noise (acoustic pollution).

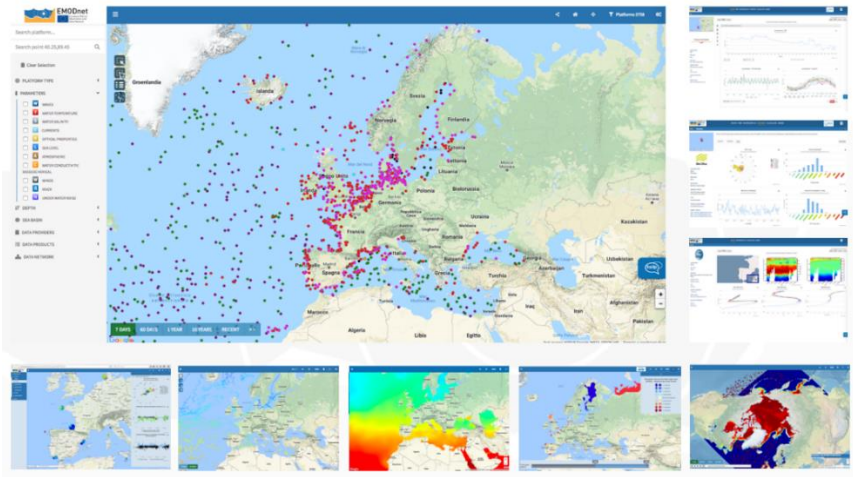


Fig. 2. EMODnet Physics products overview

EMODnet Physics is continuously increasing the number and type of platforms in the system by unlocking and providing high quality data from a growing network. EMODnet Physics was able to make available circa 30.000 platforms and more than 400.000 datasets, and to publish more than 350 map layers[9] derived from the data products.

For each connected platform, a dedicated platform page is available. These pages provide users with metadata, plots, download features, platform products e.g. monthly averages or wind plots, more information and links, as well as statistics on the use of the data from that platform. Data quality information is available in connection to datasets.

EMODnet Physics is developing interoperability services to facilitate machine-to-machine interaction and to provide further systems and services with European seas and ocean physical data and metadata. Interoperability services are provided by a GeoServer infrastructure that is OGC compliant. The WMS and WFS layers offer information about which parameters are available (where and who is the data originator, etc.). EMODnet Physics also provides SOAP – web services which allow linkage to external services with near real-time data stream and facilitate a machine-to-machine data fetching and assimilation. EMODnet Physics is also offering plot widgets[10] to embed a parameters plot/chart into an external portal.

Table 1. EMODnet Physics M2M.

Service	Description	Examples
permaURL	All platforms	http://www.emodnet-physics.eu/map/platinfo/piradar.aspx?platformid=10273 http://www.emodnet-physics.eu/map/platinfo/pidashboard.aspx?platformid=10273
API REST/SOAP	Latest 60 days of data	Service description @ http://www.emodnet-physics.eu/map/spi.aspx www.emodnet-physics.eu/map/Service/WSEmodnet2.aspx www.emodnet-physics.eu/map/service/WSEmodnet2.asmx
OGS WMS, WFS	Postgresql + Geoserver	geoserver.emodnet-physics.eu/geoserver/web examples and service description @ www.emodnet-physics.eu/map/service/GeoServerDefaultWMS www.emodnet-physics.eu/map/service/GeoServerDefaultWFS
THREDDS (OpenDAP, WMS, WCS)	Latest 60 days + HFR data + Ice	thredds.emodnet-physics.eu/thredds/catalog.html
ERDDAP	Latest 60 days	erddap.emodnet-physics.eu
widgets	All plots	www.emodnet-physics.eu/Map/Charts/PlotDataTimeSeries.aspx?paramcode=TEMP&platid=8427&timerange=7

References

1. EMODnet Homepage, http://www.emodnet.eu.
2. EMODnet Bathymetry Homepage, http://www.emodnet-bathymetry.eu.
3. EMODnet Geology Homepage, http://www.emodnet-geology.eu.
4. EMODnet Seabed Habitats Homepage, http://www.emodnet-seabedhabitats.eu.
5. EMODnet Chemistry Homepage, http://www.emodnet-chemistry.eu.
6. EMODnet Biology Homepage, http://www.emodnet-biology.eu.
7. EMODnet Physics Homepage, http://www.emodnet-physics.eu.
8. EMODnet Human Activities Homepage, http://www.emodnet-humanactivities.eu.
9. EMODnet Physics Layer Preview http://geoserver.emodnet-physics.eu/geoserver/web/wicket/bookmarkable/org.geoserver.web.demo.MapPreviewPage?1.
10. EMODnet Physics Widget syntax, www.emodnet-phys-

ics.eu/Map/Charts/PlotDataTimeSeries.aspx?paramcode=PPPP&platid=ZZZZ&timerange=YY; where PPPP is the parameter (e.g. TEMP = sea temperature), ZZZZ is the platform ID (e.g. 8427 is Arkona) and YY is either 7 or 60 (days)

Towards an analytics of optimal ship routes based on meteo-oceanographic datasets

Gianandrea Mannarini

CMCC (Centro Euro-Mediterraneo sui Cambiamenti Climatici), Lecce, Italy

Abstract. Maritime routes optimised for meteo-oceanographic conditions can be represented in terms of two key-metrics: their optimal duration and length. They are here proposed as a basis for an automatised analysis of big route datasets.

Big Data is entering the maritime world mainly fed by two kinds of initiatives: the industrial and the geoscientific ones. In addition to internal and market needs, the industry has been stimulated to go digital by initiatives under the e-Navigation umbrella (IMO[1], IALA[2] , STM-Validation[3]). The geoscientific community has driven the evolution of operational meteo-oceanographic systems (GEOSS[4], GOOS[5], CMEMS[6], EMODnet[7]) towards applications directly aimed at a societal benefit (geoBluePlanet[8], AtlantOS[9]).

The outcome of these efforts is the production of large amounts of data, either observational or model data. While within the geoscientific community open-access data policies have been a standard for a long time (NOAA[10], H-2020[11]), legitimate issues of security and market competition have so far prevented a wider uptake of private data. This barrier could be overcome in a win-win loop that has already been outlined[12].

The H-2020 AtlantOS project aims to achieve a transition from a loosely-coordinated set of existing ocean observing activities to a sustainable, efficient, and fit-for-purpose observing system, engaging stakeholders around the Atlantic. Its WP8 will deliver a suite of products that are targeted at issues of societal concern, such as flooding, maritime safety, harmful algal blooms, and offshore aquaculture. The ship routing contribution to AtlantOS WP8 is delivered by CMCC through the development of VISIR[13].

VISIR computes optimal routes in a dynamic environment, keeping into account the safety of navigation. Its first version was aimed to plan least-time routes in presence

[1] http://www.imo.org/en/OurWork/safety/navigation/pages/enavigation.aspx
[2] http://www.iala-aism.org/products-projects/e-navigation/
[3] http://stmvalidation.eu/
[4] https://www.earthobservations.org/geoss.php
[5] http://www.goosocean.org/
[6] http://marine.copernicus.eu/
[7] http://www.emodnet.eu/
[8] https://geoblueplanet.org/
[9] https://www.atlantos-h2020.eu/
[10] https://repository.library.noaa.gov/view/noaa/10169
[11] https://goo.gl/9gECUo
[12] https://goo.gl/XmKrwf > "Towards a new paradigm for ship routing" (pag.3)
[13] www.visir-model.net

of waves [1] and has been an operational service for the Mediterranean Sea for more than three years [2] by now. The latest VISIR developments include the capability to account, on top of waves, also for ocean currents. Currents impact routes by modifiying the speed over ground (SOG) of the vessel with respect to the speed through water. Thus, additional savings of route duration T^* can be achieved through an optimal use of ocean-currents. However, increases of route length L are also a possible outcome.

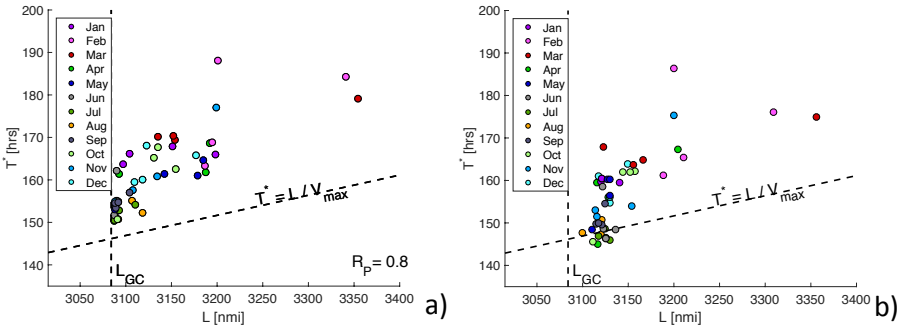


Fig. 1. Scatter plots of L and T^* of optimal routes computed by VISIR in presence of a) waves only and b) both waves and currents. Vessel top seed is $V_{\max} = 21.1$ kts. L_{GC} is the length of the great arc route joining the route endpoints. 4 routes per each month of year 2017 are computed.

Fig.1 shows seasonal route variability in the (L, T^*) plane for a case-study route between the Chesapeake Bay (USA) and Lisbon (Portugal) computed via VISIR using CMEMS ocean currents and waves analyses in input. The half-space below $T^*/L = 1/V_{\max}$ represents the region with route average speed in excess of vessel maximum speed $V_{\max}$. This can be achieved thanks to ocean currents. If the ship engine power is set to be constant, then T^*/L is a proxy of the energy efficiency of the voyage (EEOI), providing an estimation of the route carbon footprint [3].

Furthermore, analysis in the (L, T^*) plane reduces the dimensionality of the original dataset, consisting of the full waypoint-based route information. This enables analysis of larger route ensembles stemming from either observational (voyage reports, AIS data) or model data (optimal paths), from either industrial or academic source.

References

1. G. Mannarini, N. Pinardi, G. Coppini, P. Oddo, and A. Iafrati. VISIR-I: small vessels – least-time nautical routes using wave forecasts. *Geoscientific Model Development*, 9(4):1597–1625, 2016.
2. G. Mannarini, G. Turrisi, A. D'Anca, M. Scalas, N. Pinardi, G. Coppini, F. Palermo, I. Carluccio, M. Scuro, S. Cretì, R. Lecci, P. Nassisi, and L. Tedesco. VISIR: technological infrastructure of an operational service for safe and efficient navigation in the Mediterranean Sea. *Natural Hazards and Earth System Sciences*, 16(8):1791–1806, 2016.
3. MEPC. Guidelines for voluntary use of the ship energy efficiency operational indicator (EEOI). Technical Report MEPC.1/Circ.684, IMO, 2009.

Big data for Maritime Domain Awareness: An AIS case study

Waldo Kleynhans

IMIS Global Limited, Fareham, UK

1 Introduction

The world's oceans is of critical importance to humanity as it is key to fisheries, shipping as well as the environment. From an economic perspective, it is estimated that 90% of all global goods and energy transportation are done by sea with millions of people being dependent on maritime related activities for their livelihood. As maritime activities increase globally, there exist a greater dependency on technology in the monitoring, control and surveillance of vessel activities. One of the most prominent systems for monitoring vessel activity is Automatic Identification System (AIS). AIS operates in the VHF band and transmits messages from vessels which can be received by other vessels, terrestrial shore stations as well as satellites.

When dealing with AIS data, two pertinent factors to conciser are:

1. **AIS data fidelity**: Due to the fact that AIS is broadcast in a non-secure channel, information could be manipulated / corrupted (such as malicious or inadvertently introduced false GPS positions and errors in vessel parameters). In addition, AIS receivers are not controlled in the same manner as AIS transmitters, which could introduce additional errors at the receiver side [1].
2. **Significant volume increase of AIS messages**: Due to the global increase in vessels fitted with AIS transmitters as well as the proliferation of satellite and terrestrial receiving stations there has been a significant increase in AIS messages received globally (estimated at over a 40% increase over the last four years [1]). While this increase in AIS data volumes is beneficial as this enriches the information available to maritime authorities, processing and storage of these large data volumes can become problematic especially when performing analytics based on historic vessel temporal and positional data.

By using advanced filtering and analytics, IMIS Global Limited has been able to process the AIS data stream to eliminate a large portion of the faulty data (as described in point 1) and is also focusing on the efficient storage and compression of vessel track history derived from AIS positional report data in an effort to deal with the challenges raised in point 2. Storing only fundamental data required to accurately construct vessel historic track (trajectory) data is also one of the key objectives of the datAcron initiative [2].

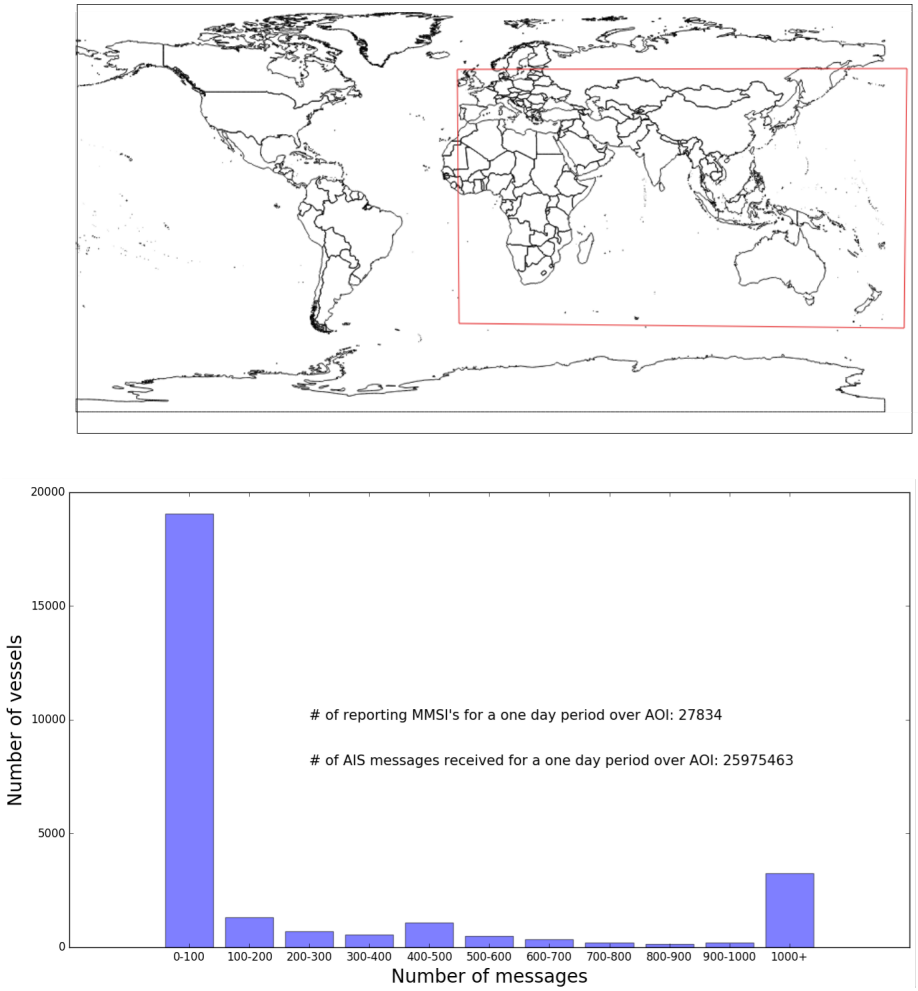


Fig. 1. Area of interest (top), received AIS data from this area is provided by more than a hundred data sources and includes terrestrial receivers, base stations as well as satellites. The distribution of the number of vessels vs. the number of received AIS messages is also provided (bottom)

2 Study Area

The study area as well as the typical data volumes for a 24 hour period for 17 March 2018 is represented in figure 1 (top) where it can be seen that the area of interest (AOI) covers a significant portion of the globe. AIS data from this area is provided by more than a hundred AIS data sources which includes terrestrial receivers, base stations as well as satellites. The total number of vessels that was received during the day was 27834. These vessels transmitted a combined total

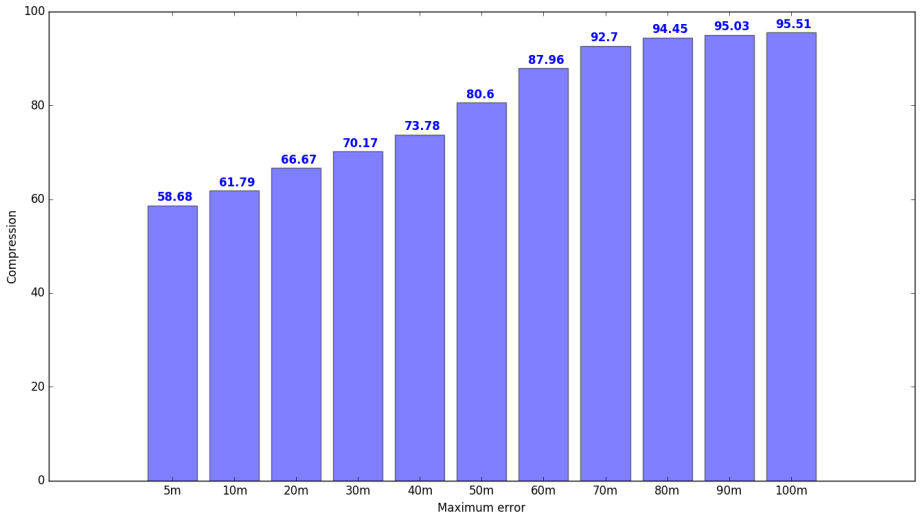


Fig. 2. Data compression as a function of the maximum distance error when when compared to the complete raw AIS positional dataset

of ≈ 26 Million AIS messages during the 24 hour period. The distribution of the typical number of messages received per vessel is also shown in figure 1 where it can be seen that the majority (>70%) of vessel's transmitted AIS messages were received only 0-100 times throughout the AIS network with (≈ 15%) of vessel's total received AIS messages for the day being more than 1000. This can be attributed to the fact that AIS messages transmitted from vessels in the reception range of terrestrial AIS receivers are received far more frequently than from satellite AIS sensors whereas satellite AIS sensors, on the other hand, are able to cover very large areas and consequently can receive AIS message from significantly more vessels but at the much lower revisit rate. It was also found that 10% of the vessels generated >90% of the AIS messages received through the network of receivers in the AOI.

3 Results and discussion

From section 2 it is clear that a diverse AIS reception network generates an enormous amount of data. In many cases, especially when vessels are traveling at very low speeds or are stationary, and vessel track data are of interest, only a fraction of these reported messages are required for reconstructing an accurate vessel track when compared to using the entire compliment of raw AIS messages. An on-line lossy track compression methodology was used in the filtering process of the positional AIS data. In essence the method tracked the AIS message stream originating from each ship in an on-line fashion and compared the spatio-temporal information contained in each newly presented AIS message with that

of the current vessel track history, by calculating a distance metric related to the maximum allowable spatio-temporal distance error from the current track, a decision was made to include the newly presented positional information or discard it. The goal was to be able to then do historic vessel track reconstruction as a function of the maximum allowable track distance error when compared to the complete raw AIS positional dataset. The results are shown in figure 2, it can be seen that, as expected, the compression ratio increases as the maximum distance error metric is relaxed. It was shown that a compression ratio of ($\approx 90\%$) could be achieved when the error metric of between 60m and 70m is selected.

References

1. Batty E. (2018) Data Analytics Enables Advanced AIS Applications. In: Doulkeridis C., Vouros G., Qu Q., Wang S. (eds) "Mobility Analytics for Spatio-Temporal and Social Data" MATES 2017. Lecture Notes in Computer Science, vol 10731
2. Georgios M. Santipantakis et. al.2017. "Specification of Semantic Trajectories Supporting Data Transformations for Analytics: The datAcron Ontology." In Proceedings of Semantics2017, Amsterdam, Netherlands, September 1114, 2017, 8 pages.

A Prototype for Maritime Event Forecasting

Elias Alevizos[1] and Alexander Artikis[2,1]

[1] Institute of Informatics & Telecommunications, NCSR Demokritos
[2] Department of Maritime Studies, University of Piraeus

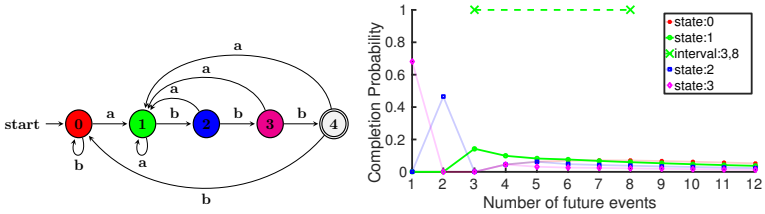
1 Introduction

We have built a prototype for complex event forecasting and applied it to the maritime domain [1]. The problem may be stated as follows: given a stream of input events and a pattern defining relations between such events, in the form of a regular expression, the goal is to estimate at each new event arrival the number of future events that we will need to wait for until the expression is satisfied, and therefore a match be detected.

2 Approach

Event patterns are first converted to deterministic finite automata (DFA) through standard conversion algorithms. As an example, see Fig. 1a, which depicts the DFA for the pattern $R = a \cdot b \cdot b \cdot b$, i.e., an occurrence of a must be followed by three occurrences of b. Next, we derive a Markov chain that will be able to provide a probabilistic description of the DFA's run-time behavior. If the input events are independent and identically distributed (i.i.d.), then there is a direct mapping of the states of the DFA to states of a Markov chain and the transitions of the DFA to transitions of the Markov chain. The transition probabilities of the Markov chain are the occurrence probabilities of the various event types. If the input events are dependent on some of the previous events seen in the stream, i.e., the stream is generated by an m^{th} order Markov process, we perform a more complex transformation. The transition probabilities are then conditional probabilities on the event types. We call such a derived Markov chain a Pattern Markov Chain (PMC) of order m and denote it by PMC_R^m, where R is the initial pattern and m the assumed order of the Markov process. After constructing a PMC, we can use it in order to calculate the so-called *waiting-time* distributions. Given a specific state of the PMC, a *waiting-time* distribution gives us the probability of detecting a full match of the original regular expression in k events from now. Forecasts are in the form of intervals, like $I = (start, end)$. The meaning is that the DFA is expected to reach a final state sometime in the future between $start$ and end with probability at least some constant threshold θ_{fc} (provided by the user). These intervals are estimated by a single-pass algorithm that scans a waiting-time distribution and finds the smallest (in terms of length) interval that exceeds this threshold. See, e.g., Fig. 1b, which shows distributions for the states of the DFA of Fig. 1a when $m = 0$. The dashed green line is the forecast interval produced when the DFA is in state 1, with $\theta_{fc} = 50\%$, i.e., this is the smallest interval whose probability is above 50%.

We implemented a forecasting system, Wayeb, based on Pattern Markov Chains. Algorithm 1 presents in pseudo-code the steps taken for recognition and forecasting.



(a) DFA. (b) Waiting-time distributions.

Fig. 1: DFA and waiting-time distributions for $R = a \cdot b \cdot b \cdot b$, alphabet $\Sigma = \{a, b\}$, $m = 0$.

ALGORITHM 1: Forecasting algorithm

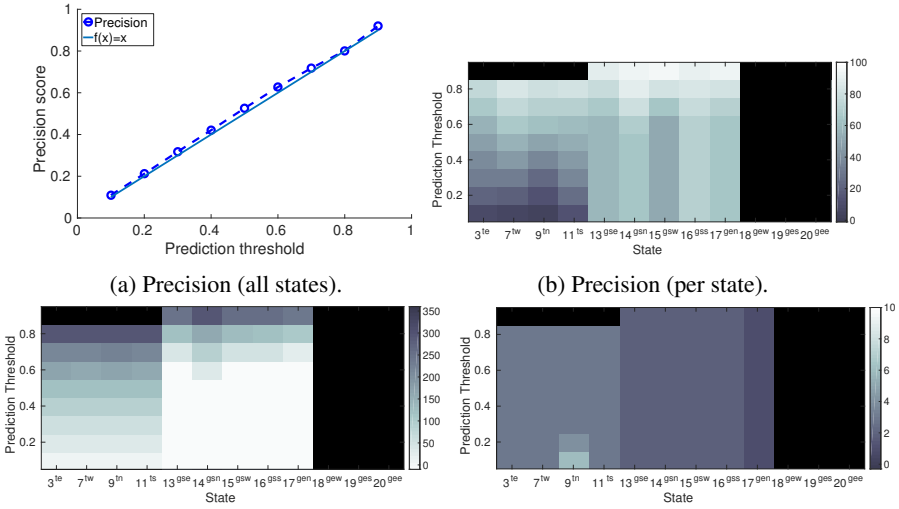
Input: Stream S, pattern R, order m, maximum spread ms, forecasting threshold P_{fc}
Output: For each event $e \in S$, a forecast $I = (start, end)$

1 $DFA_{\Sigma^* \cdot R}$ = BuildDFA(R, m);
2 PMC_R^m = WarmUp(S, $DFA_{\Sigma^* \cdot R}$);
3 F_{table} = BuildForecastsTable(PMC_R^m, P_{fc}, ms);
4 $CurrentState$ = 0;
5 $RunningForecasts$ = $\varnothing$;
6 **repeat**
7 e = RetrieveNextEvent(S);
8 $CurrentState$ = UpdateDFA($DFA_{\Sigma^* \cdot R}$, e);
9 **if** $CurrenState$ *not final* **then**
10 $I = F_{table}(CurrentState)$;
11 $RunningForecasts = I \cup RunningForecasts$
12 **else**
13 UpdateStats($RunningForecasts$);
14 $RunningForecasts$ = $\varnothing$;
15 **end**
16 **until** $true$;

Wayeb reads a given pattern R in the form of a regular expression, transforms this expression into a NFA and subsequently, through standard determinization algorithms, the NFA is transformed into a *m-unambiguous* DFA (line 1 in Algorithm 1). For the task of event recognition, only this DFA is involved. At the arrival of each new event (line 7), the engine consults the transition function of the DFA and updates the current state of the DFA (line 8). Note that this function is simply a look-up-table, providing the next state, given the current state and the type of the new event. Hence, only a memory operation is required.

There are three metrics that we report in order to assess our module's performance and the quality of its forecasts:

– $Precision = \frac{\#\ of\ correct\ forecasts}{\#\ of\ forecasts}$. At every new event arrival, the new state of the DFA is estimated (line 8 of Algorithm 1). If the new state is not a final state, a new forecast is retrieved from the look-up-table of forecasts (line 10). These forecasts are maintained in memory (line 11) until a full match is detected. Once a full match



(a) Precision (all states).

(b) Precision (per state).

(c) Spread (per state).

(d) Distance (per state).

Fig. 2: Results for the pattern $Turn \cdot GapStart \cdot GapEnd \cdot Turn$ with $m = 1$.

is detected, we can estimate which of the previously produced forecasts are satisfied, in the sense that the full match happened within the interval of a forecast (line 13). These are the correct forecasts. All forecasts are cleared from memory after a full match (line 14).

- $Spread = end - start$.
- $Distance = start - now$. This metric captures the distance between the time the forecast is made (now) and the earliest expected completion time of the pattern. Note that two intervals might have the same spread (e.g., $(2, 2)$ and $(5, 5)$ both have $Spread$ equal to 0) but different distances (2 and 5, assuming $now = 0$).

$Precision$ should be as high as possible. With respect to $Spread$, the intuition is that, the smaller it is, the more informative the interval. For example, in the extreme case where the interval is a single point, the engine can pinpoint the exact number of events that it will have to wait until a full match. On the other hand, the greater the $Distance$, the earlier a forecast is produced and therefore a wider margin for action is provided. Thus, "good" forecasts are those with high precision (ideally 1.0), low spread (ideally 0) and a distance that is as high as possible (ideal values depend on the pattern). These metrics may be calculated either as aggregates, gathering results from all states (in which case average values for $Spread$ and $Distance$ over all states are reported), or on a per-state basis, i.e., we can estimate the $Precision$, $Spread$ and $Distance$ of the forecasts produced only by a specific state of the DFA.

3 Demo for Maritime Event Forecasting.

Wayeb was tested against a real-world dataset that came from the field of maritime monitoring. When sailing at sea, (most) vessels emit messages relaying information

about their position, heading, speed, etc.: the so-called AIS (automatic identification system) messages. AIS messages may be processed in order to produce a compressed trajectory, consisting of critical points, i.e., important points that are only a summary of the initial trajectory, but allow for an accurate reconstruction [2]. The critical points of interest for our experiments are the following:

- $Turn$: when a vessel executes a turn.
- $GapStart$: when a vessel turns off its AIS equipment and stops transmitting its position.
- $GapEnd$: when a vessel turns on its AIS equipment back again (a $GapStart$ must have preceded).

We used a dataset consisting of a stream of such critical points from ≈ 6.500 vessels, covering a 3 month period and spanning the Greek seas. Each critical point was enriched with information about whether it is headed towards the northern, eastern, southern or western direction. For example, each $Turn$ event was converted to one of $TurnNorth$, $TurnEast$, $TurnSouth$ or $TurnWest$ events. We show results from a single vessel, with ≈ 50.000 events.

Figure 2 shows results for the pattern

$$Turn \cdot GapStart \cdot GapEnd \cdot Turn \tag{1}$$

where $Turn$ is shorthand notation for

$$(TurnNorth + TurnEast + TurnSouth + TurnWest)$$

with $+$ denoting the OR operator. Similarly for $GapStart$ and $GapEnd$. With this pattern, we would like to detect a sequence of movements in which a vessel first turns (regardless of heading), then turns off its AIS equipment and subsequently re-appears by turning again. Communication gaps are important for maritime analysts because they often indicate an intention of hiding (e.g., in cases of illegal fishing in a protected area).

The aggregate precision score (Figure 2a) is very close to the baseline performance. This precision score is calculated by combining the forecasts produced by all states of the PMC. In order to better understand Wayeb's behavior, a look at the behavior of individual states could be more useful. Figures 2b – 2d depict image plots for various metrics against both the forecast threshold and the state of the PMC. The metrics shown are those of precision (on the recognized matches), spread and distance. In each such image plot the y axis corresponds to the various values of P_{fc}. The x axis corresponds to the states of the PMC. The x axis shows how advanced we are in the recognition process, when moving from one state to the next. The black areas in these plots are "dead zones", meaning that, for the corresponding combinations of P_{fc} and state, Wayeb fails to produce forecasts (i.e., it cannot guarantee, according to the learned transition probabilities, that the forecast intervals will have at least P_{fc} probability of being satisfied). On the contrary, areas with light colors are "optimal", in the sense that they have high precision, low spread (the colorbar is inverted in the spread plots) and high distance in their respective plots. A look at the per-state plots reveals something interesting (Figures 2b, 2c, 2d). Note that, in order to avoid cluttering, we have removed duplicate states

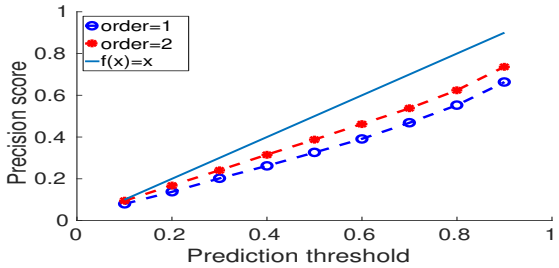


Fig. 3: Results for the pattern $TurnNorth \cdot (TurnNorth + TurnEast)^* \cdot TurnSouth$.

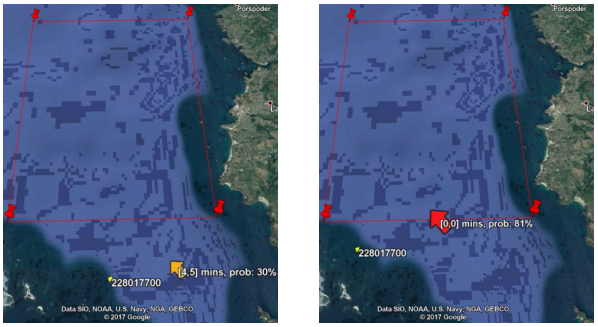
from the per-state plots. In addition, the superscript of each state in the x axis shows the last event seen when in that state. For example, the superscript te corresponds to $TurnEast$, tw to $TurnWest$, tn to $TurnNorth$ and ts to $TurnSouth$ (states 3, 7, 9 and 11 respectively). Similarly for $GapStart$ for which superscripts start with gs (states 13–16) and for $GapEnd$ (ge and states 17–20). These per-state plots show that there is a distinct "cluster" of states (13–17) which exhibit high precision scores for all values of P_{fc} (Figure 2b) and small spread for most values of P_{fc} (Figure 2c). Therefore, these states constitute what might be called "milestones" and a PMC can help in uncovering them. By closer inspection, it is revealed that states 13–16 are visited after the PMC has seen one of the $GapStart$ events (we remind that $GapStart$ is a disjunction of the four directional sub-cases). Moreover, $GapEnd$ events are very likely to appear in the input stream right after a $GapStart$ event, as expected, since during a communication gap (delimited by a $GapStart$ and a $GapEnd$), a vessel does not emit any messages. State 17, which also has a similar behavior, is visited after a $GapEndNorth$ event. Its high precision scores are due to the fact that, after a $GapEnd$ event, a $Turn$ event is very likely to appear. It differs from states 13–16 in its distance, as shown in Figure 2d, which is 1, whereas, for states 13–16, the distance is 2. On the other hand, states 18–20, which correspond to the other 3 $GapEnd$ events, fail to produce any forecasts. The reason is that there are no such $GapEnd$ events in the stream, i.e., whenever this vessel starts transmitting again after a Gap, it is always headed towards the northern direction.

Figure 3 shows results for the pattern

$$TurnNorth \cdot (TurnNorth + TurnEast)^* \cdot TurnSouth$$

This pattern is more complex since it involves a *star closure* operation on a nested *union* operation. It attempts to detect a rightward reverse of heading, in which a vessel is initially heading towards the north and subsequently starts a right turn until it ends up heading towards the south. Such patterns can be useful in detecting maneuvers of fishing vessels.

Figure 3 shows that a model with m=1 is unable to approximate well-enough the correct waiting-time distribution. Increasing the order to m=2 improves the precision score, but it still remains under the baseline performance. One could attempt to further increase the value of m, but this would substantially increase the cost of building the PMC. For $m = 1$, the generated PMC has ≈ 30 states. For $m = 2$, this number rises to



(a) Same vessel, two routes, early snapshot.

(b) Same vessel, two routes, late snapshot.

Fig. 4: *withinArea* event (Google Earth).

≈ 600 and the cost of creating an unambiguous DFA and then its corresponding PMC rises exponentially. When stationarity is assumed (as in our case) and the model does not need to be updated online, an expensive model can be tolerated.

We also demonstrate our method on the so-called *withinArea* event, which reports whether a vessel is located within the boundaries of a designated area, defined as a polygon, e.g., a protected area or a port. Forecasting arrival times at ports can help reduce the operating cost and emissions footprint of ships, as they are typically required to wait outside a port when there is a high traffic volume. Maritime monitoring companies currently rely on manual information in the AIS messages to forecast arrival at ports, which is very often wrong. Thus a forecasting method that does not rely on humans is highly desirable. In order to produce forecasts in terms of time, we sample each trajectory at regular intervals. We finally enrich each message with spatial information about whether the vessel is located within the area, whether it is close to the area and whether its heading points towards a direction that intersects with the area. Fig. 4a and 4b show the behavior of our prototype, where two different routes of the same vessel are shown, at different times, and the red rectangle is the area of interest. The left arrow corresponds to a route that never crosses the area and the right arrow a route that does cross it. The size of the arrow is proportional to the probability of entering the area and its color becomes more red as the $start$ of the forecast interval becomes smaller, i.e., red indicates that the vessel will enter the area very soon. As can be seen, the left route never produces forecasts because the model has learned that this route never leads to a *withinArea* event, whereas, for the right route, as the vessel approaches the area (compare Fig. 4a and 4b) the forecasts become more confident and focused.

References

1. Alevizos, E., Artikis, A., Paliouras, G.: Event forecasting with pattern markov chains. In: Proceedings of DEBS. pp. 146–157. ACM (2017)
2. Patroumpas, K., Alevizos, E., Artikis, A., Vodas, M., Pelekis, N., Theodoridis, Y.: Online event recognition from moving vessel trajectories. GeoInformatica (2016)

Test case development for big data solution evaluation

Maximilian Zocholl[1], Elena Camossi[1] and Anne-Laure Jousselme[1]

[1] NATO STO CMRE, La Spezia 19126, Italy

Exploiting big data for detection tasks requires systematic testing of the big data solution as they are supposed to cope reliably with high dimensional data with large variations. For big data solutions which ingest Automatic Information System (AIS) data [1] typical big data variations are observable in volume, velocity, variety and veracity of the data. The design space, out of which test cases can be selected, is defined by the input factors of the big data solution and the respective big data variations which translate into the dimensions of the design space and the values of the big data variations which are represented the domain of each factor. When combining the typically large number of dimensions with multiple possible values or even continuous value domains for each factor, the number of possible experiments translates seamlessly into combinatorial mayhem and the non-applicability of classical statistical assumptions [2] As the performance of all possible experiments is infeasible, the systematic selection of test cases is crucial. Even though big data dimensions are generally interdependent [3], the reduction of complexity is only feasible by assuming independence between different big data dimensions. To frame the sequential selection process, the described approach serves at narrowing down and defining the design space.

Taking into account domain specific constraints, different big data variations are supposed to be stronger coupled whilst others are only loosely interdependent. In AIS related big data solutions, volume and velocity are functionally stronger interdependent, while veracity and variety are semantically stronger interdependent. For the following, variety and veracity variations are discussed more detailed. While further narrowing down the design space two main challenges arise. Internal and external validity of the test cases need to be guaranteed [4]. Since external and internal validity are necessary conditions for the validity of test cases, they can be used to deduce guidelines.

Firstly, the internal validity of the data needs to be guaranteed. This requires that the difference in the expected results between true and false events is significant. For this purpose, the number of experiments is adapted to the accuracy of the method to be evaluated in the course of the evaluation.

More detailed, measured distances between the information sets that include positive and negative events are reduced stepwise during consecutive tests of a test series. By starting from information sets with a large distance, the initial number of experiments is reduced. After failing one test, there are different ways to proceed.

If the development process was completed before the evaluation started, the test series can either be stopped or the pace of the distance measure can be reduced to refine the evaluation result.

If the development process is not completed, yet, an alternation between evaluation and development phases can lead to an improvement of the evaluation result. For disjoining evaluation and development, a new test has to be created replacing the failed test. This procedure reduces the risk of a development which follows the evaluation criteria but not the concept to be detected. If a purely test-driven development is chosen, the same tests are typically reevaluated.

Secondly, the external validity needs to be guaranteed. This allows for the deduction of two procedures, taking into account dependencies between different big data variations and using different approaches for the test case development.

Firstly, interdependence between different big data variations needs to be taken into account stepwise. For this purpose, the number of test cases evaluating the variations on one big data dimension is adapted to the evaluation result of the test cases evaluating variations of a defining big data dimension. As an example, the variation of the big data veracity requires the existence of a variety of big data to degrade. While unidirectional dependence allows for the complete testing of the independent big data dimension before testing the dependent big data dimension, mutually dependent big data variations require a coordinated extension of the test cases.

For this, big data solution specific constraints on variety and veracity are used successively.
The constraints on the variety are defined by the mapping of the data fields or factors of the big data sources to the input variables of the components of the big data solution. If a data field is not processed, it can be excluded from the following evaluation step.

The constraints on the veracity are defined in dependence to the available big data variety. Starting from a given variety, e.g. variety 1 in Figure 1, the veracity is degraded iteratively until the required performance is testified. If the test is failed, as shown in Figure 1 a), the test is stopped for both veracity and variety variations or continued with similar test cases after the improvement of the big data solution until the required performance is attained, as shown in Figure 1 b). The successful accomplishment of one extent of variation of variety enables the extension of the design space to the next larger test case, Figure 1 c).

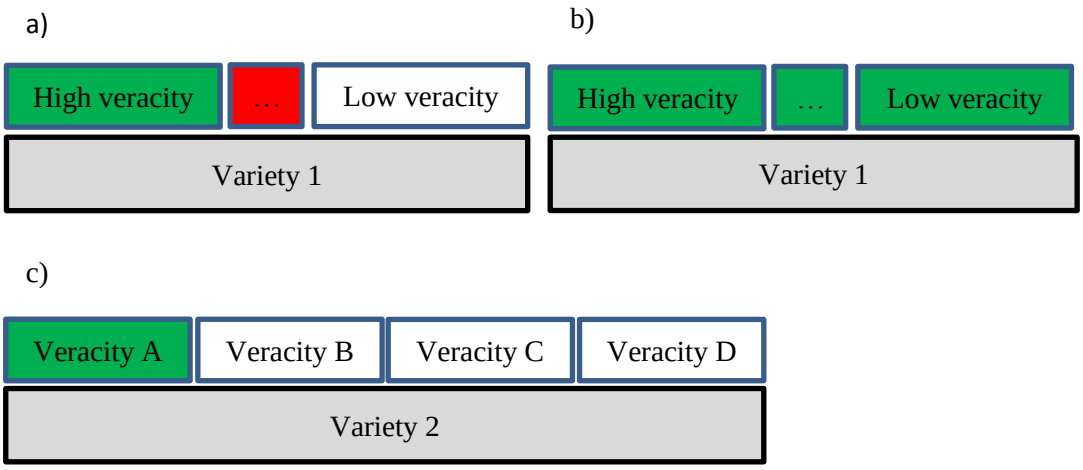


Figure 1: Testing of veracity variation depending on the progress in testing variety variation

For each class of detection task, e.g. tasks relating to “area”, “speed”, “course” and “heading”, “connectivity” and maritime status and each detection task, e.g. “null speed” or “mismatch speed vessel type” can be defined separately by the combination of input data value levels of the data fields, identified to contribute to the detection result. E.g. if a component takes into account longitude and latitude for detecting a speed event, only the degradation of those input variables can be tested.
In case the variables being taken into account for a detection task are unknown, the impact of a veracity variation yields information on it. E.g. varying the “gps_speed” signal and observing the detection of “null speed” events unveils whether the respective AIS field is used for the detection task “null speed”.

With an increasing variety of the information taken into account, detections and testing of these detections becomes feasible for data with very low veracity but only marginal differences to data with high veracity,

e.g. as current eddies can be estimated from AIS signals [5] weather and ocean conditions can yield information about the veracity of AIS signals.

Secondly, for guaranteeing the representative character of the developed test cases the addressed variations need to be set into context by the actual big data variations. By mapping the performed test cases to the design space, gaps can be detected and addressed by future test case developments. The definition and narrowing down of the design space can be supported by the combination of theoretical event descriptions and sample data.

With respect to theoretical concept descriptions, the definition of the design space is improved by two aspects. Firstly, events which are known to exist but which are not included in the sample dataset can be added artificially, e.g. by simulation. Secondly, human interpretable event definitions are specified and instantiated, e.g. collision is specified into collisions between vessels of different types and collisions between one vessel and the mainland.

With respect to the sample data, the definition of the design space benefits from three aspects. Firstly, theoretically impossible factor combinations that are observed in the sample data (e.g. "engaged in fishing" and "at anchor" while changes in position data indicates a moving vessel). Secondly, unknown variations, i.e. variations which are not described, yet, increase the array of testable variations, e.g. the fast and consecutive sending of two different AIS positions from two transceivers from bow and stern of a vessel. Thirdly, Depending on the frequency of factor combination observations, the design space can be characterized in a probabilistic way.

References

1. ITU-R: Technical Characteristics for an Automatic Identification System Using Time-Division Multiple Access in the VHF Maritime Mobile Band, International Telecommunications Union, itu-r m.1371-4, (2010).
2. Fan, Jianqing and Han, Fang and Liu, Han: Challenges of Big Data analysis, National Science Review, vol. 1, no. 2, pp. 293-314, (2014).
3. Gandomi, Amir and Haider, Murtaza: Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, vol. 35, no. 2, pp. 137-144, Elsevier, (2015).
4. Cook, Thomas D and Campbell, Donald Thomas and Shadish, William: Experimental and quasi-experimental designs for generalized causal inference, Houghton Mifflin Boston, (2002).
5. Jakub, Thomas D: Estimating Ocean Currents from Automatic Identification System Based Ship Drift Measurements, University of Colorado at Boulder, (2013).

Data degradation variations for maritime situational indicator detection assessment

Clément Iphar[1], Anne-Laure Jousselme[1] and Cyril Ray[2]

[1] NATO STO CMRE, La Spezia, Italy
[2] Naval Academy Research Institute, Brest, France
{clement.iphar;anne-laure.jousselme}@cmre.nato.int
cyril.ray@ecole-navale.fr

Abstract. The identification and the prediction of maritime situational indicators (MSI), being of foremost importance in maritime scenarios assessment, must be assessed in the scope of degradation variation. In this respect, this paper presents a set of methods for maritime data controlled degradation enabling the analysis of a dataset with various veracity levels. A twofold methodology is proposed: for data controlled degradation and for MSI prediction assessment.

Keywords: Data degradation, Data veracity, Maritime Situational Indicator.

1 The V's of Big Data

Traditionally, four properties are associated with Big Data, which are called the four V's: Volume, Velocity, Variety and Veracity. The Volume is in relation with the amount of data to be handled, whereas the Velocity is effectiveness of gathering and processing. Being more applicative, enabling effectiveness and researches to be done more quickly, the Velocity of data exploitation is more important than the Volume.

The Variety property covers the fact that data in Big Data takes several forms [1], and most of those data sources are recent. The Veracity is a challenge as it is not linked to the quality of the data but with its relation to the world. It represents the fact for a datum to be truthful, *i.e.* to correctly depict the world in an expected way. Our prospective study concentrates on the application of variety and variety variations in the scope of AIS (Automatic Identification System) maritime positioning messages.

2 Degradation methods for veracity variations

The variations in veracity, in the case of AIS messages will be measured by the input data quality. AIS data is imperfect, as errors, falsifications and spoofing cases have been shown [2]. We are interested in means to degrade (enrich) data so that the dimension of veracity varies, and the data quality dimensions defined in [3] (completeness, accuracy, clarity, continuity and timeliness) can be evaluated. The purpose of degradation is to provide reference data based on a qualitative and quantitative data

analysis, with known quality levels so that various scenarios can be set to assess and validate MSI algorithms. The methods we propose in this work can be classified in four families: noise adjunction, data modification, data removal or data addition (Figure 1).

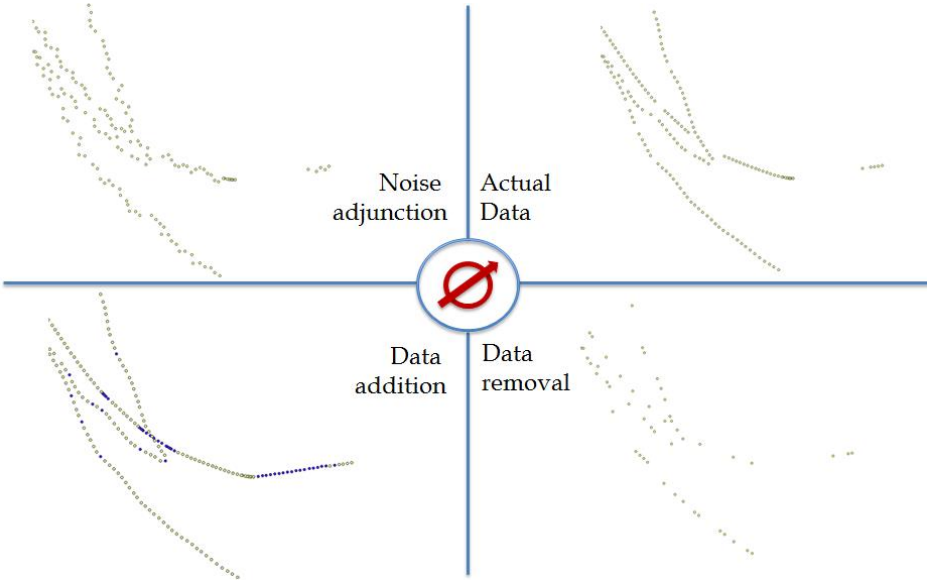


Figure 1: Controlled degradation of data quality

The application of those methods to datasets is intended to modify the inner quality of those datasets, and therefore to modify the outputs of analyses performed on the basis of their data. As the data fields within AIS messages are various, those selected methods will apply to selected data fields, and for instance, the noise adjunction, consisting in the blurring of the value by applying a Gaussian shift to it, can only be applied to physical values such as speed, course or the position.

Data modification can be targeted on the identity or on the coordinates of the vessel. Data removal can affect targeted data fields or whole messages, the frequency of the removed data fields or messages being a variable of the study. In this case, mechanisms for the handling of modified or removed data must be set. In this respect, three main missing data mechanisms are distinguished: missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR) [4].

3 An application to maritime scenarios

As defined in [5], the applicative domain for our study is the information that can be extracted from the study of maritime traffic, for which we have parsed AIS messages from an antenna located in Brest, France, as well as antenna, geographic, meteorological, sea state, vessel and port data, all aligned in both time and space with AIS

data [6]. In this end, a set of MSIs have been set, describing a state of a vessel at a given time or interval. Those MSIs encompass speed, course and vessel location features, as well as the state of the AIS transmission and the navigational status. A set of scenarios along the topics of fishing security, sustainable development protection and maritime security [7] have been developed. In this work, the veracity of data varies following the application of methods presented in Section 2, and the consequences on the scenario assessment are evaluated.

Future work will consist in the development and implementation of a methodology for the assessment of the prediction of MSIs and subsequently for the determination of the scenarios, involving *ad hoc* tools for the controlled degradation of datasets (and therefore the measurement of the veracity of a dataset) and the comparison of MSI results that will be the consequence of such dataset variation.

References

1. McAfee, A. and E. Brynjolfsson (2012). *Big data: the management revolution*. In: Harvard Business Review 90(10), pp. 60-66.
2. Iphar, C. (2017). *Formalisation of a data analysis environment based on anomaly detection for risk assessment – Application to Maritime Domain Awareness*. PhD Thesis, MINES ParisTech, PSL Research University
3. Camossi E., A.-L. Jousselme, C. Ray, M. Hadzagic, R. Dréo and C. Claramunt (2017) H2020 EU datAcron project WP-5.3: *Maritime Experiments Specification*. 67p.
4. Schafer, J. and J. Graham (2002). *Missing Data: Our View of the State of the Art*. In: Psychological Methods 7(2), pp. 147-177.
5. Jousselme A.-L., C. Ray, E. Camossi, M. Hadzagic, C. Claramunt, K. Bryan, E. Reardon and M. Ilteris (2016) H2020 EU datAcron project WP-5.1: *Maritime Use Case Description*. 39p.
6. Ray, C., R. Dréo, E. Camossi and A.-L. Jousselme (2018) *Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance* 10.5281/zenodo.1167595
7. Ray C., E. Camossi, A.-L. Jousselme, M. Hadzagic, C. Claramunt and E. Batty (2016) H2020 EU datAcron project WP-5.2: *Maritime Data preparation and curation*. 37p.

Modelling Dynamic Bayesian Networks to Identify Behaviour of Interest

Mathias Anneken[1], Francesca de Rosa[2], Anne-Laure Jousselme[2], and Sebastian Robert[3]
mathias.anneken@kit.edu

[1] Karlsruhe Institute of Technology (KIT), Vision and Fusion Laboratory (IES), Karlsruhe, Germany
[2] NATO STO - Centre for Maritime Research and Experimentation (STO - CMRE), La Spezia, Italy
[3] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), Karlsruhe, Germany

Abstract. Situation assessment in today's surveillance tasks is still done mostly by human experts, while the amount of data to process steadily increases. This may result in wrong assessments and missing of critical situations. Therefore, a support system to assess incoming data and indicate situations of interest is desirable. This work describes an approach to model expert and domain knowledge in order to reason about situations of interest. The model is based on a Dynamic Bayesian Network (DBN). The DBN will encode different situations (of different levels of abstraction) and their relationship. By using inference algorithms, the existence probability for a situation can be estimated. As an illustration, a smuggling situation is modelled and solved through AIS incoming information.

Keywords: Dynamic Bayesian Network, situation analysis, decision support, smuggling

1 Introduction

Situational awareness is of utter importance for operators in surveillance tasks. This is especially true for the maritime domain, due to the large amount of data gathered by different heterogeneous sensor sources. Therefore, algorithms to identify suspicious behaviour and reducing the flood of data an operator encounters would greatly help the tedious task.

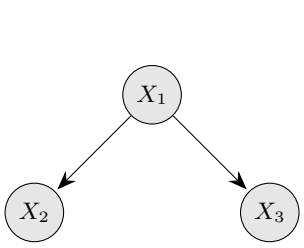
Thus, we present an algorithm which is able to represent activities and specific situations by encompassing the whole context around the object of interest. This algorithm encodes as a Dynamic Bayesian Network (DBN). The network is a graphical representation of Bayesian reasoning in which the conditional probabilities between different random variables are possibly elicited by experts.

2 Bayesian Networks

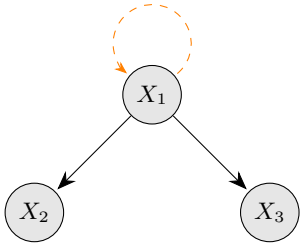
A Bayesian network is a graph. It represents random variables by using nodes and their interdependencies by using edges between the nodes. Altogether they form a directed acyclic graph, e.g. Fig. 1(a). For the parents and children of the nodes, a generalized first-order Markov property holds, which means a node only depends on its immediate parents. For instance, for the random variables $X_1, \ldots, X_n$ the joint probability distribution is given by

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)), \tag{1}$$

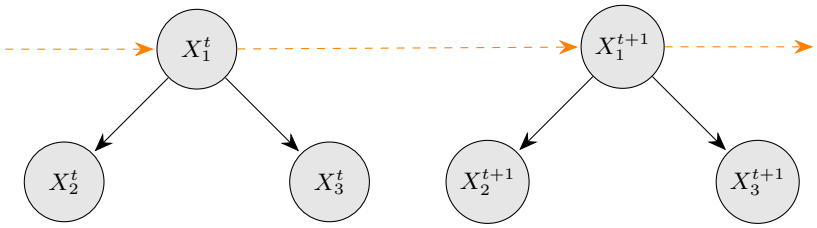
with $Pa(X_i)$ describing the parent nodes of X_i. In the example $Pa(X_1)$ would yield $\{X_2, X_3\}$. Information about the value of X_i is called evidence. The estimation of the probability density distribution for X_i given all the evidences is called inference. There are different kinds of algorithms for exact or approximate inference available.



(a) Bayesian network.



(b) DBN; orange dashed arrows represent temporal dependencies.



(c) Unrolled example; orange dashed arrows represent temporal dependencies.

Fig. 1. Example for a Bayesian network, a DBN, and the unrolled DBN.

A DBN extends the Bayesian network by adding time slices. The slices will be used to represent the development of random variables over the time by including the conditional probability between time slices (either between the same random variables or between different variables). The only limitation is the usage in a

feed-forward manner. An example is give in Fig 1(b). The unrolled network for two time slices is given in Fig 1(c).

For more in-depth information on (dynamic) Bayesian networks see [1, 2].

3 How to build a situation

In order to build a DBN which corresponds to a situation of interest, a knowledge base has to be established. The term situation as used in this work is defined by [3] as follows:

> »A situation is defined as an external semantic interpretation of sensor data. Interpretation means that situations assign meanings to sensor data. External means that the interpretation is from the perspective of applications, rather than from sensors. Semantic means that the interpretation assigns meaning on sensor data based on structures and relationships within the same type of sensor data and between different types of sensor data.«

Following [4], the external semantic statement is either »true« or »false« and is assigned to a situation S_t at time point t. This statement results due to the temporal sequence of the interaction between modelled objects and their specific attributes. The objects $O^1, \ldots, O^n$ are relevant for the semantic statement. An object $O^i, \forall i \in \{1, \ldots, n\}$ has the relevant attributes $A_1^i, A_2^i, \ldots, A_{m_i}^i$ for $m_i \in \mathbb{N}$. The configuration space $\mathcal{O}_S$ is then given as

$$\mathcal{O}_S = \bigtimes_{i=1}^{n} \bigtimes_{k=1}^{m_i} A_k^i. \tag{2}$$

As a situation is always given in a time context, we define $\Omega = \mathcal{O}_S \times T$, with T as representation of the time dimension. Thus a situation can be interpreted as the mapping

$$S_t : \Omega \to \{0, 1\}. \tag{3}$$

This mapping will be interpreted as binary random variable. For a given sigma algebra $\mathcal{A}$ of subsets of Ω and a probability measure P defined on $\mathcal{A}$ the probability space $(\Omega, \mathcal{A}, P)$ results. P is then a probability distribution of S_t. Thus, the recognition of a situation means to estimate the probability $P(S_t = 1) \in [0, 1]$.

Situations can be divided in different types, which vary in the level of abstraction. In this work, there are elementary situations and abstract situations:

- For an *elementary situation* the support is modelled directly. Thus, the existence probability is mapped deterministically.
- For an *abstract situation* the support cannot be modelled directly. Thus, the existence probability depends on the existence of other situations (both elementary or abstract).

Situations can depend on each other. Either the relationship is a necessary or a sufficient condition. Following [5]:

– A situation A is necessary for another situation B, if the existence of B implies the existence of A, i.e., $B \rightarrow A$.
– A situation A is sufficient for another situation B, if the existence of A implies the existence of B, i.e., $A \rightarrow B$.

The situations and their dependencies can then be translated into a DBN. An example is given in Fig. 2. This example is built by following the algorithms described in [4, 5]. Therefore, each node which is not representing an elementary situation has a temporal arrow.

The next step would involve experts to identify the conditional probability tables (CPT) for each of the nodes. As this would result in a huge amount of parameters (given that each situation is only a binary random variable for a situation with k child situations, the CPT would consist of 2^{k+1} entries). In [4] an algorithm is introduced to identify these parameters by using three parameters, one to tune the sensitivity of the network for positive evidences, one for the sensitivity for negative evidences and one for the asymptotic behaviour.

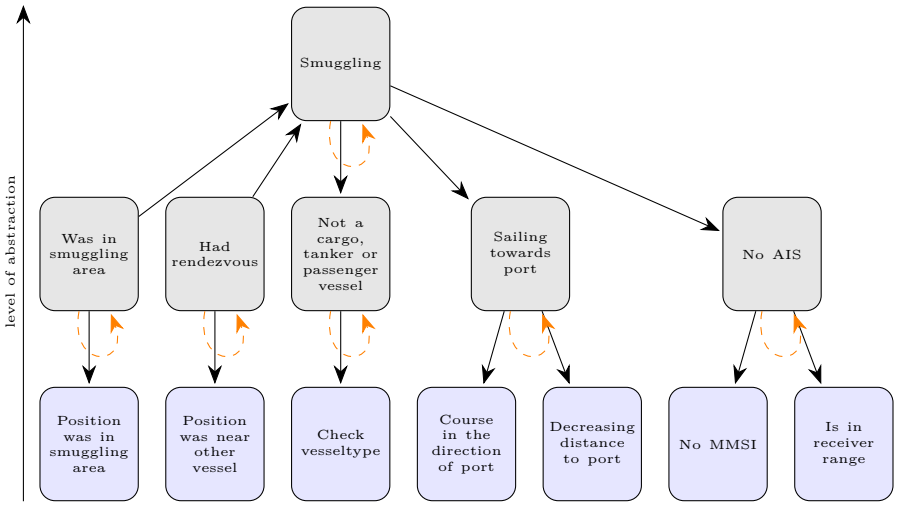


Fig. 2. An example DBN to identify a smuggling activity [4]. The grey nodes are abstract situations, the blue ones elementary. Solid black arrows are conditional dependencies on the same time slice, orange dashed ones are dependencies to the slice before.

4 Data sources

As evidences the data from different kind of sensors is used. This includes AIS, RADAR, electronic navigational charts, reports and many more. In order to take

the reliability of the different sources into account, a Reporting Layer is added to the DBN. This is part of the MARISA project and is further described in [6].

Acknowledgements

The underlying projects to this article are funded by the WTD 81 of the German Federal Ministry of Defense. The authors are responsible for the content of this article.
The MARISA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 740698.

References

1. Barber, D.: Bayesian Reasoning and Machine Learning. Cambridge University Press (2014)
2. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT Press, Cambridge, MA, Cambridge, MA (2012)
3. Ye, J., Dobson, S., McKeever, S.: Situation identification techniques in pervasive computing: a review. Pervasive and Mobile Computing **8**(1) (2012) 36–66
4. Fischer, Y.: Wissensbasierte probabilistische Modellierung für die Situationsanalyse am Beispiel der maritimen Überwachung. PhD thesis, Karlsruhe Institute of Technology (2016)
5. Fischer, Y., Reiswich, A., Beyerer, J.: Modeling and recognizing situations of interest in surveillance applications. In: Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2014 IEEE International Inter-Disciplinary Conference on. (March 2014) 209–215
6. de Rosa, F., Ben Abdallah, N., Jousselme, A.L., Anneken, M.: Source quality handling in fusion systems: a bayesian perspective. Maritime Big Data Workshop (2018)

Source quality handling in fusion systems: a Bayesian perspective

Francesca de Rosa[1], Nadia Ben Abdallah[1], Anne-Laure Jousselme[1] and Mathias Anneken[2]

[1] NATO STO – Centre for Maritime Research and Experimentation, Viale San Bartolomeo 400, La Spezia, Italy

2 Fraunhofer IOSB, Fraunhoferstraße 1, 76131, Karlsruhe, Germany

Information available in the maritime domain can come from a variety of sources (e.g. AIS, LRIT, radar, VTS, VMS, operators, eye witnesses and social media). Those sources differ in nature and potential in *quality*, therefore to correctly fuse and reason with information in a multi-source context, source factors have to be taken into account. While research is still seeking to characterize those source factors, more specifically to define which are the source factors to account for and how to account for them, the mathematical instruments to consider information source *quality* in the fusion process exist. For example, different models have been proposed to account for source *reliability* in the Bayesian framework [1, 2, 3, 4, 5, 6, 7]. The term source *reliability*, is not clearly defined and is open to different interpretations [8]. But it is important to underline that although there is a relationship between source *reliability* and *quality*, the two concepts in general do not coincide. In fact, source *quality* represents a characteristic of the source of information itself, while the concept of *reliability* refers to the ability to rely or depend on the information provided by a source [9]. Therefore, it is not a characteristic of the source per se, rather a mediating element between the source of information and the receiver, which makes an estimate on if and how to rely on a specific source as a function of many factors. Those factors include the capacity and/or willingness of the source of providing good information but possibly also others, such as accountability (related to legal aspects).

The models to handle source quality previously mentioned can be categorised in two groups based on the assumption that reliability is either an exogenous or endogenous variable [10]. In the former family of models (e.g. [1, 2, 3]) the characteristics of information and of the corresponding source are modeled through an overall likelihood ratio. In the later the source reliability is captured explicitly through specific model variables (e.g. [4, 5, 6, 7]). More specifically, Bovens and Hartmann [4] proposed a hierarchical model, in which both the evidence report (*Rep*) and the source reliability (*Rel*) are captured explicitly, while the distinction between the real evidence and the report about the evidence is embedded in the relation between the hypothesis (*H*) and the evidence report (*Rep*). The behavior of the source as a function of the *reliability* is modeled through the *Rep* node. Different works propose models of partially reliable sources behaviours [11, 12], interpreting *reliability* as truthfulness[1].

[1] Truthful refers to someone or something "telling the truth, especially habitually" (http://www.dictionary.com).

The explicit representation of the *reliability* allows easy accounting for specific evidence on it, such as in the case of an intelligence report which might be communicated with additional meta-information on the reliability rating. However, in many real life applications specific meta-information on the source reliability is not provided. Therefore, this important component has to be estimated. This leads to the need to include explicitly the different underpinning factors that build up the *reliability* [13, 14], which will be modeled as ancestors of the *Rel* node.

In this work, we describe how this mechanism for handling source quality in Bayesian reasoning could support vessel behavioural analysis. In fact, this extended model has been implemented in the MARISA Dynamic Bayesian Network (DBN) Behavioural Analysis Service [15], which presents a hierarchical layered structure, proposing an easy yet powerful mechanism to define a multi-source Bayesian Network accounting for source reliability. More specifically, in the DBN a *Reporting Layer* is added to the *Situation Layer*, in which a situation of interest is represented at different levels of abstraction. While the higher hierarchical levels represent the situation under analysis, the lower levels represent the single variables that influence the assessment regarding the situation. Those are the variables on which we might receive information. However, this information is not a direct evidence on the state of the variable, rather a report on the state. Therefore, evidence is entered through the *Reporting Layer*, which accounts for the reliability of the source of information. In the DBN the reliable source behaviour is modelled as a *truthful* source, while the unreliable source behavior is modelled as a *randomizer* [4], which corresponds to a source that provides reports that are equally likely and uncorrelated with the true state of the world.

Acknowledgements

This research has been conducted as part of the MARISA project. The MARISA project has received funding from the European Union's Horizon H2020 research and innovation programme under grant agreement No 740698.

References

1. Birnbaum, M. H., Mellers, B., "Bayesian inference: Combing base rates with opinions of sources who vary in credibility", Journal of Personality and Social Psychology, 45, 792-804, 1983.
2. Birnbaum, M. H., Stegner, S. E., "Source credibility in social judgment: Bias, expertise and the judge's point of view", Journal of Personality and Social Psychology, 37, 48-74, 1979.
3. Corner, A., Hahn, U., "Evaluating Science arguments: Evidence, uncertainty and argument strength", Journal of Experimental Psychology: Applied, 15, 199-212.
4. Bovens, L., Hartmann, S., "Bayesian epistemology", Oxford University Press, 2003.
5. Friedman, R., "Route analysis of credibility and hearsay", Yale Law journal, 96, 667-742, 1987.
6. Goldman, A. I., "Knowledge in a social world", Oxford University press, 1999.

7. Hahn, U., Oaksfor, M., Harris, A. J. L., “Testimony and Argumetation: A Bayesian Perspective”, Bayesian Argumentation, ed. Zenker, F., Springer Library, 2013.
8. de Rosa, F., Jousselme, A.-L., “Critical review of uncertainty communication standards in support to Maritime Situational Awareness”, NATO STO Centre for Maritime Research and Experimentation (in press)
9. http://www.dictionary.com/browse/reliability
10. Hahn, U., Oaksford, M., and Harris, A.J.L. (2012), ‘Testimony and Argument: A Bayesian Perspective’, in Bayesian Argumentation, ed. Zenker, F., Springer Library
11. Claveau, F., “The independence condition in the variety-of-evidence thesis”, Philosophy of Science, 80(1):94-118, 2013
12. Haenni, R., Hartmann, S., “Modelling partially reliable information sources: a general approach based on Dempster-Shafer theory”, Information Fusion, 7(4):361-379, 2006.
13. Fenton, N., Neil, M., Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive Science*, 37, 61–102
14. Lagnado, D.A., Fenton, N., Neil, M., “Legal idioms: a framework for evidential reasoning”, Argument and Computation, 4, 46-53, 2013
15. Anneken, M., de Rosa, F., Jousselme, A.-L., Robert, S., “Modelling Dynamic Bayesian Networks to Identify Suspicious Behaviour”, Proceedings of the Maritime Big Data Workshop, NATO STO - CMRE, 9-10 May 2018

Multivariate temporal data analysis for abnormal vessels behavior detection: exploring different approaches

Rui Maia and Cláudia Antunes

Instituto Superior Técnico, Av. Rovisco Pais 1, Lisboa, 1049-001
rui.maia, claudia.antunes@tecnico.ulisboa.pt

Abstract. Anomaly detection in temporal data is a specific field of data analysis considered as crucial, since abnormal behavior typically represent critical situations that should be addressed. Different vessels behavior might help detecting smuggling or drug trafficking. In this work we explore different anomaly detection approaches that can natively include the multiple dimensions of temporal data without loss of information. We aim at identify fitted approaches to specific maritime big data use cases regarding vessels abnormal or irregular behavior.

1 Introduction

The growing number of systems collecting large quantities of data is leveraging the need for efficient big data analysis approaches. This research work explores machine learning methods that can process heterogeneous datasets of multivariate temporal data aiming behavior analysis and anomaly detection. Experimental datasets include geographic (satellite) vessel positions and characteristics, weather and sea conditions. Anomaly detection methods are tested for their effectiveness in maritime monitoring scenarios. Our research challenges include: complex networks of different sensors in maritime context; categorical and real valued parameters that might have been manipulated by emitting entities; complex relations between different dataset domains and the existence of hidden semantic relations between different temporal series.

2 Experimentation

Different multivariate anomaly detection approaches are included in our experimentation, specifically methods proposed by Ahmed et al. [1], focused on network anomalies and Zheng [2], exploring trajectory data for anomaly detection. Kane et al. [3] based their approach on dimensionality reduction and correlation analysis using Singular Values Decomposition (SVD). Zor et al. [4] proposed a framework for anomalous ferry (vessels) tracks detection and Malhotra et al. [5] applied a neural network Long Short Term Memory (LSTM) to incorporate long term tendencies of data series that are difficult to capture using other techniques.

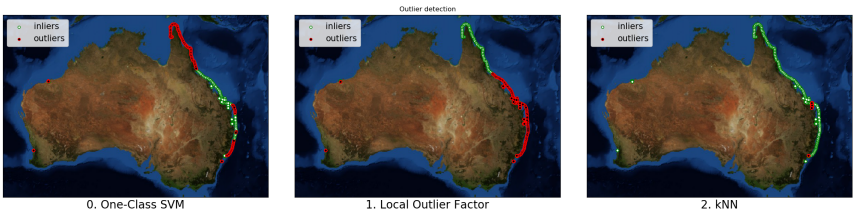


Fig. 1. One-Class Support Vector Machine, Local Outlier Factor and k-Nearest Neighbors methods applied to vessel positions in an unsupervised learning setup.

Figure 1 illustrates the result of the application of three different methods using unsupervised learning methods observing vessels position information.

For the experimentation we are using datasets made available by the Australian Maritime Safety Authority (AMSA), the Australian Ocean Data Network (AODN) and the National Oceanic and Atmospheric Administration (NOAA). We gathered information for two experimental scenarios, namely Australia and USA. Both scenarios include Automatic Identification System (AIS) data and climate datasets, including atmosphere and ocean related information.

3 Future Work

Most anomaly detection research is dominated by univariate time-series anomaly detection methods which only include one dimension varying over time. Univariate approaches are adapted by many authors to the multivariate case. Although, univariate approaches cannot model multivariate anomalies and the hidden relations between series, rather they help to identify anomalies in one or more features of a data series. Multivariate anomaly detection problem raise distinct and complex challenges due to the hidden data structure and semantics between time-series.

References

1. Ahmed, M., Mahmood, A.N., Hu, J.: A survey of network anomaly detection techniques. Journal of Network and Computer Applications **60** (2016) 19–31
2. Zheng, Y.: Trajectory data mining: an overview. ACM Transactions on Intelligent Systems and Technology (TIST) **6**(3) (2015) 29
3. Kane, A., Shiri, N.: Multivariate time series representation and similarity search using pca. In: Industrial Conference on Data Mining, Springer (2017) 122–136
4. Zor, C., Kittler, J.: Maritime anomaly detection in ferry tracks. In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, IEEE (2017) 2647–2651
5. Malhotra, P., Vig, L., Shroff, G., Agarwal, P.: Long short term memory networks for anomaly detection in time series. In: Proceedings, Presses universitaires de Louvain (2015) 89

Remote Sensing Analysis Framework for Maritime Surveillance Application

Egbert Schwarz[1], Sergey Voinov[1], Detmar Krause[1], Olaf Frauenberger[1] and Björn Tings[2]

[1] German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Kalkhorstweg 53, 17235 Neustrelitz, Germany
[2] German Aerospace Center (DLR), Remote Sensing Technology Institute, Henrich-Focke-Str. 4, 28199 Bremen, Germany
Egbert.Schwarz@dlr.de

Abstract. Synthetic Aperture Radar (SAR) and high (HR) and very high (VHR) resolution optical satellite images are valuable sources of information for maritime situational awareness. The objective of the Maritime Security Lab Analysis Framework is to develop and integrate applications to support operational services based on those SAR and optical satellite images in near real time (NRT).
In the frame of maritime surveillance services based on satellite images, the German Remote Sensing Data Center (DFD), part of the German Aerospace Center (DLR), established a framework to support automated NRT processing of huge amounts of image data from different satellite missions provided by a network of ground stations and service providers. Developed at DLR's Maritime Safety and Security Lab Neustrelitz, the Processing Framework is intended to support operational maritime surveillance value adding services. Main components are the Processing System Management (PSM), embedded thematic processors and the Graphical User Interface (GUI).
The presentation will describe the overall workflow of data handling, the interfaces and the operator GUI, which was implemented for operational use at DLR's Ground Station Neustrelitz.

Keywords: Synthetic Aperture Radar, VHR, Near-Real Time processing, Ship detection, Oil detection

1 Introduction

Remote Sensing technologies are getting more and more used as main contribution in maritime situational awareness systems. Information required can be e.g. on marine environment, border surveillance, fishery control or for ice services. For maritime surveillance large areas need to be monitored, while the detectable items remain comparable small, thus requiring observations with large coverage and a possibly high geometrical and temporal resolution. Single satellite data can fulfill these require-

ments in parts, where a tradeoff has been made between coverage, geometrical and temporal resolution and as well specific characteristics, e. g. the spectral band width.

A classical solution for targeted object detection involves Spaceborne Synthetic Aperture Radar (SAR) systems, which are weather independent and able to cover large areas with good spectral resolution. They provide information to different topics, e. g. manmade structures like vessels, buoys, on oil pollution, on icebergs or wave height. Optical images can offer valuable contribution in maritime surveillance domain as well, but due to possible clouds some limitations still remain. Nevertheless depending on their geometrical resolution, optical images can be used to detect objects and to recognize them (at least their types) in order to analyze the behavior of such objects.

In combination with data from other sources it is possible to evaluate and to verify information from satellite data but also vice versa to support the verification of such data. E. g. integrating additional attributes acquired from Automatic Identification System (AIS) extends these capabilities significantly and allows performing such tasks as AIS anomaly investigations or detection of malicious actions. For near real time applications it is important to shorten the time from data collection to dissemination as much as possible to ensure that the extracted information could be provided with respect to the user requirement and in the required time frame. All data required have to be collected from different sources simultaneously according to the time of the satellite acquisition. The workflow control of the processing system must ensure to perform this tasks in parallel, taking priorities into account and to be robust enough in case of error or missing data.

2 Processing System

2.1 Processing Management

Currently the framework [1] is deployed in a cluster of virtual machines, linked with a shared file system GFS-2. The Processing System Management (PSM) [2] is one of the main components of the Data Information Management System (DIMS) which is used in a multi mission context and composed of systems for ordering and production control together with a central archive. The framework of the entire processing system is managed by the PSM, which automatically schedules different processes (processors) according to the processing request in place or by the availability of the data required. All processors are triggered by thematic rules, added to the PSM control system which is calling all processing steps in sequence or in parallel. The general workflow is illustrated in figure 1.

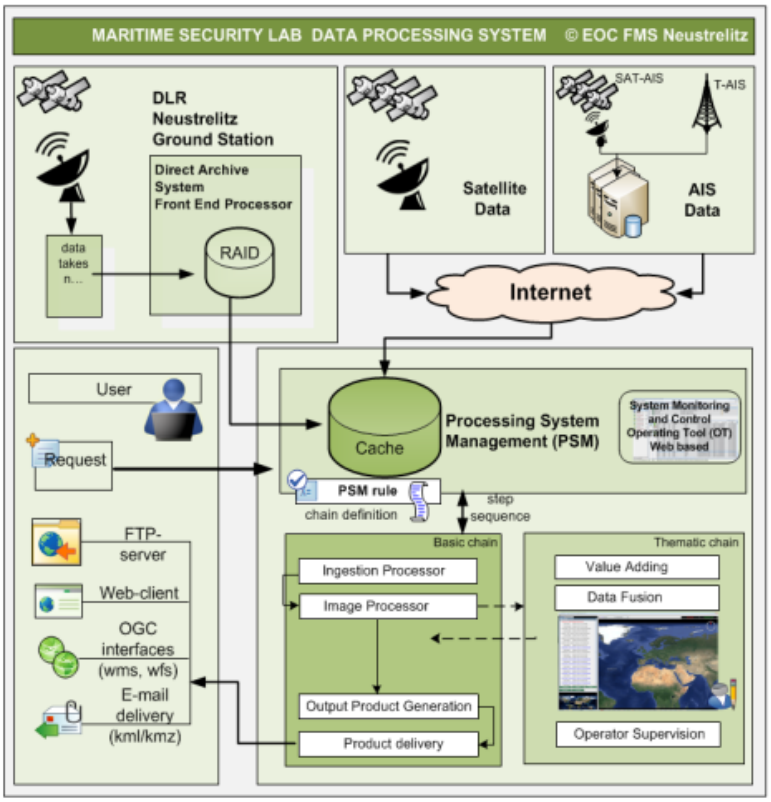


Fig. 1. Maritime Security Lab Optical Data Processing - framework architecture

Since images could be acquired at the same time by different satellites multiple PSM instances are implemented and configured to support different missions in parallel. Each single PSM and the individual processing requests are monitored and controlled with the Operating Tool (OT), a GUI which provides a set of views and allows operator interactions. Among others main processing modules are the “Image Processing”, the “Thematic Chain” and the “Product Dissemination”.

The image processing is the first main step and requires, with respect to the increase of image resolution and size, a high level of parallelization (hardware and software). For the thematic part every processing request contains at least one of the possible value adding types: e.g. Vessel Detection (incl. Wind and Wave), or Oil Spill Detection. Automatic Identification System (AIS) data are merged in accordance to the image extent and collection time. These data are acquired for the requested imaging time and coverage via real-time web interface based on HTTP protocol. Additionally an AIS plausibility check processor [4] is implemented.

For some of these scenarios the framework requires human interaction during the processing. A second GUI in the frame of value adding based on the NASA WorldWind API has been developed and embedded to support these operator interactions.

The PSM communicates with this GUI via a Simple Object Access Protocol (SOAP) connection. In this way the results could serve automatically as an additional input before generation of the final products.

3 Data and Methods

3.1 Image Processing

For near real time (NRT) response satellite payload data, consisting on instrument raw data, satellite attitude data and status information will be transmitted to a ground station within the visibility of the satellite. The first steps after data reception are quality checks, sorting, and assembly to create consolidated level 0 products. Depending on priorities and time constraints, instrument raw data might be pre-selected according to the timeliness applicable from the user request, since the following processing steps are very time consuming. In combination with additional calibration data sets and orbit information, this data will be processed to so-called Level 1 products, which for SAR products consist nominally of gray scale images for different instrument modes with annotated geo-information or images mapped to a geographic grid. Processing steps performed in case of SAR images are transformations, radiometric calibration, quality checks and geometric processing. The products generated are the basic input for thematic processing, so-called level 2 (L2) processing.

In case of optical satellite image processing among others cloud processing is required to consider the cloud cover threshold from the user request. Basic imagery products like panchromatic or multispectral images could be handled and fit as input for the value adding L2 processing.

3.2 Thematic value adding processing

Based on the L1B image product thematic processing can start, which might require additional input from external sources as well. In case of vessel detection the focus is to detect the objects of interest, determine their position, and if possible to determine the type and to estimate parameters.

In case of SAR vessel detection the SAR AIS Integrated NRT Toolbox (SAINT) [3] has been developed at the Maritime Safety and Security Lab Bremen, part of DLR's Remote Sensing Technology Institute. Beside vessel detection the software is used to perform iceberg detection, extraction of wind fields, containing information about wind speed and direction, as well as the detection of the sea state information e.g. the significant wave height (Hs). For monitoring areas like the German Exclusive Economic Zone (EEZ) generally known stationary objects are stored in a filter database and will be masked out. Other steps are filtering out ambiguities resulting from ghost images of large structures on sea or land. Information on detected objects will be merged with AIS information. The AIS information needed is retrieved in parallel to the image processing step, considering the imagery coverage and acquisition time. The AIS data is pre-processed and interpolated according to the image acquisition

time if necessary. Both sources of information will be compared and matching information will be combined.

The derivation of wind and wave information (SAR only) is based on geophysical model functions (GMF). For the wind field extraction in addition meteorological information on the main wind direction is needed, since the information based on SAR image solely is ambiguous. This input is generated by the Weather Research and Forecasting (WRF) Model [5], called as an additional pre-processing step. Running the different processors in parallel ensures that all extracted information is instantly available to the other information extraction processes.

Processing optical images differs with respect to SAR image processing chain. One difficulty is dealing with clouds, which obscure the view in parts, but alter also the signal from ground due to shadows. For vessel detection there is also a need to obtain good geometrical knowledge on the object in order to recognize the shape. Depending on the geometrical resolution and the spectral characteristics of the vessel this information might vary significantly. Using deep learning algorithms it is possible to detect whether an object is a ship, and it might be possible to determine also the type of the vessel.

3.3 Product formats and dissemination

After running the L2 product generation different product formats are available in addition to the NRT L1b product. Among others the system supports kml/kmz, ESRI shape file, json and netCDF. In order to fulfill the variety of user requirements different dissemination options have been developed. This includes sftp/GridFTP or e-mail delivery, access over OGC interfaces (wms, wfs) enabling users to connect the data directly in GIS applications without having a local copy.

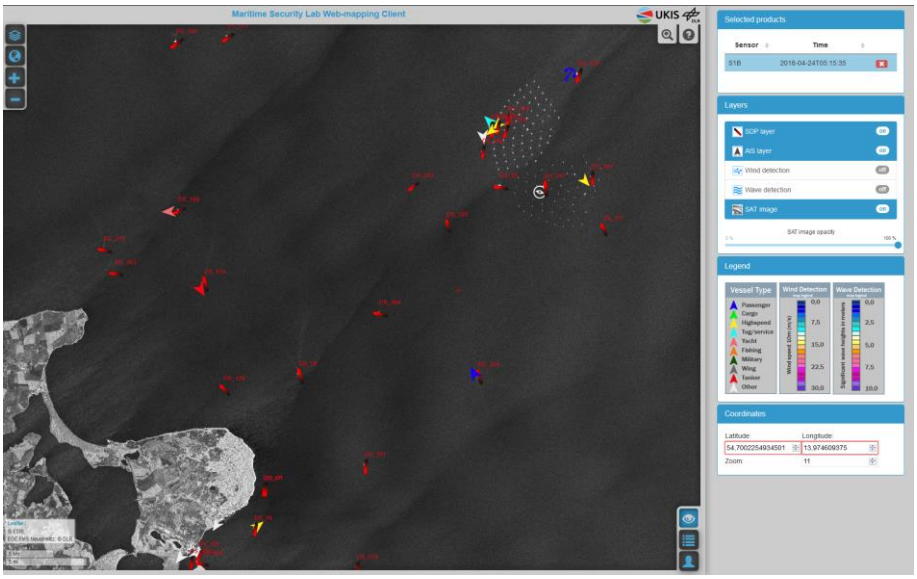


Fig. 2. Vessel detection product view using DLR's Web-mapping Client

Formats compliant to the EMSA CSN Standards are developed as well. Another alternative is a maritime web-mapping client developed, which requires the availability of modern internet browser only.

4 Conclusions

NRT applications for the maritime domain based on remote sensing satellite imagery have been successful implemented. Important for fast responses is the direct combination of data reception and processing. The combination of data from different sources is essential for value adding processing e. g. to refine the information, to avoid ambiguities, to get temporal and spatial context information.

With upcoming satellite missions more satellite data with higher resolution will become available. This will increase the amount of data, but also the amount of smart data like the context information extracted. For automated data processing methods needs to be established, for selecting the correct data and considering interdependencies in order to control the distribution between sequential and parallel processing.

Semi-automated processes need to be improved using deep learning algorithms as an aid for the operator.

References

1. Schwarz, Egbert und Krause, Detmar und Daedelow, Holger und Voinov, Sergey (2015) Near Real Time Applications for Maritime Situational Awareness. Deutscher Luft- und Raumfahrtkongress 2015, 22. Sep. - 24. Sep 2015, Rostock, Deutschland. urn:nbn:de:101:1-20151109505
2. Wolfmüller, M., Dietrich, D., Sireteanu, E., Kiemle, S., Mikusch, E., and Böttcher, M., 2008. Dataflow and Workflow Organization - The Data Management for the TerraSAR-X Payload Ground Segment. In: *IEEE Transactions on Geoscience and Remote Sensing*, 47(1), pp. 44-50
3. Lehner, Susanne und Tings, Björn (2015) Maritime NRT products using TerraSAR-X imagery. In: Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XL-7 (W3), Seiten 967-973. Copernicus Publications. 36th International Symposium on Remote Sensing of Environment (ISRSE), 11.-15. Mai 2015, Berlin, Germany. DOI: 10.5194/isprsarchives-XL-7-W3-967-2015
4. F. Heymann, T. Noack and P. Banyś, Plausibility analysis of navigation related AIS parameter based on time series. In: European Navigation Conference, Vienna, 2013.
5. Michalakes, J., Dudhia, J., Gill, D., Henderson, T., Klemp, J. Skamarock, W., and Wang, W., 2004. The Weather Research and Forecast Model: Software Architecture and Performance. In: Proceedings of the 11th ECMWF Workshop on the Use of High Performance Computing In Meteorology, 25-29 October 2004, Reading, UK

Maritime Analytics System: Operational Platform and Research Cooperation

Rodolphe VADAINE[1], Nadia MAAREF[1], Izaskun BOYER[1], Elodie DA SILVA[1], Ronan FABLET[2], Romain TAVENARD[3], Cédric TEDESCHI[4]

[1] Collecte Localisation Satellites (CLS), Ramonville-Saint-Agne, France
[2] IMT Atlantique, LabSTICC, Brest, France
[3] Univ. Rennes, CNRS, LETG, Rennes, France
[4] Univ. Rennes, Inria, CNRS, IRISA, Rennes, France

Keywords: AIS, maritime awareness, satellite imagery, big data, data analytics

1 Introduction

The surveillance of the maritime traffic is a major issue for security and monitoring of activities. Spaceborne technologies, especially satellite AIS ship tracking and high-resolution imaging, open new avenues to address these issues. Current operational systems cannot fully benefit from the available and upcoming multi-source data streams. In this context, CLS operates the MAS[1] (Maritime Analytics System) solution as a multi-source data fusion platform and analytics system for helping operational users to assess rapidly the maritime situation for the surveillance of the maritime traffic and the detection of abnormal behavior. Going beyond the MAS implementation, CLS is also involved in the SESAME[2] initiative with several academic partners to develop innovative big-data-oriented and machine learning approaches to deliver novel solutions for the management, analysis and visualization of multi-source satellite data.

2 Objectives

The operational users are currently facing a huge volume of data on maritime traffic (either arriving as a stream, or as archive data) consisting of coastal and satellite AIS information, observation from earth observation satellites (SAR and optical) and other tracking systems (VMS, LRIT). The volume is such that it becomes impossible to perform an efficient manual analysis exploiting all available data.
In order to address these new challenges, a data fusion platform coupled with an analytics system is needed for:

- multi source information analysis to automatically assess the normal and abnormal maritime picture
- tracking and analyzing the behavior of specific vessels of interest (either single ones or by fleet) on long periods of time
- automatic detection of specific events/anomalies (rendezvous, transshipment, etc.).

[1] MAS is supported by BPI France through the ALMACEN project funded by the "Investissements d'Avenir" French Government program

[2] SESAME is financed by DGA (French Military Procurement Agency) through the ANR/Astrid Program and with the support of the GIS Bretel (French CPER with Conseil Régional de Bretagne, and FEDER fund)

3 Architectures

We are currently using two different platforms facing different objectives:

- the CLS platform is addressing operational needs of the MAS system for the processing of both real time data received as a stream (using kafka / akka / redis / kudu / MongoDB / ... technologies) and archive data (using HDFS / parquet / Hadoop / java / spark / ... technologies)
- the Grid'5000[3] platform, the nation-wide experimental testbed (spanning 8 different locations in France), hosted and maintained by Inria, is used to support the development and testing of a set of similar technologies, but having in mind the ease of deployment (over Grid'5000) of algorithms and methods developed during the Sesame project.

Having those two platforms in parallel ensures the availability of both an operational solution (MAS on the CLS platform) and a testbed enabling the evaluation of different configurations in a more reproducible environment (Grid'5000).

4 Algorithms

State-of-the-art algorithms are available on the MAS system to detect abnormal behavior and trigger alerts, based on real time and historical data. Going beyond the MAS implementation, one of the goals of the SESAME initiative is to design novel models and algorithms for synergies between AIS and satellite observations and for the automatic detection of abnormal behaviors.

5 Operational Implementation

The MAS system has been thought as a data fusion platform and as an analytics system. The overall purpose of the system is to merge satellite data (such as AIS, VMS, LRIT, SAR Imagery) and analyze this data to detect suspicious and unusual behaviors.
The web user interface is a unique graphical interface which gives instant access to the following features:

- Dynamic display of all relevant information for a fast visualization of the maritime picture
- Fleet management: Identify and manage vessels of interest lists
- Surveillance management: Configuration of algorithms to detect abnormal behavior and trigger alerts. The algorithms are based on real time (MongoDB) and on historical (Hadoop) data.
- Alerts management: To manage the triggered alerts in a "mailbox like" menu.
- Data Export: generation of reports on ships, alerts and areas, and production of statistics.

6 Prospects

The improved capabilities developed in the course of the SESAME project will be used for the design of the MAS system future versions, including better integration of multi-source data and detection of abnormal behaviors.

[3] https://www.grid5000.fr/

Maritime Big Data Analytics: yes, but what for?

Bernard GARNIER [1] and Bruno BENDER [2]

[1] BlueSolutions Consulting, Valbonne, France
[2] Ventura Associates France, Paris, France
http://ventura-expertise.com

Abstract:
In terms of Maritime Safety and Security, there is an ample difference between open source intelligence and actionable information: intervening efficiently means having the time to bring the right asset at the right place, with a full mandate to act. This User knowledge shall be the central driver of any development meant to improve the Maritime Situation Awareness thanks to the "Maritime Big Data" opportunity enabled by the dematerialization of all the shipping and fishing documents and the "dare to share" policy among Maritime Administrations. A workable process is to revert the usual "technology race" paradigm: develop first innovative Concepts of operations (CONOPS) then be opportunistic to find the most suited technological innovations, not the usual vice-versa!

Keywords: Big Data Analytics, Maritime Security, Weak Signals Analysis, Concepts of Operations (CONOPS), Maritime Surveillance Awareness (MSA), CISE, Course of Action, Decision Aids.

1 Current State-of-Play

When discussing with the Maritime Surveillance communities, the current challenge is to cope with MSA screens already over-populated with more or less reliable ship tracks thanks to the progress of worldwide AIS data sharing: most systems aggregate AIS collected by States (e.g. the SafeSeaNet service from EMSA), by satellites (S-AIS collect), by MPAs/drones; LRIT and VMS reports can be also integrated, allowing to track permanently an ever-increasing number of vessels worldwide including those joining voluntarily (fishing, military vessels and yachts). The analysis tools currently implemented for "abnormal behavior detection" are generally limited to obvious deviations (sudden changes of course, speed etc.) and don't really help much detecting likely illegal activities.

Moreover the simplistic algorithms of these tools can be easily by-passed by unlawful actors by producing "clean" reports, including elaborating fictive positions and journeys, un-reporting dangerous goods, hiding structural or mechanical problems etc. For example, without human intelligence (HUMINT) combined with expertise in trafficking schemes, the likelihood to detect the maritime segment of a "professional" drug traffic is today close to zero.

In the same time, the broad Maritime Safety and Security Community cannot ignore the rapid growth of the powerful marketing tools fed by our daily private activity through the web – and we only see in the targeted advertising of our navigator the emerged fraction of the iceberg… The conjunction with automated profiling supported by machine learning and artificial intelligence (generically referred to as AI hereafter) is rapidly climbing the classical scale of information added value, as illustrated by fig.1.

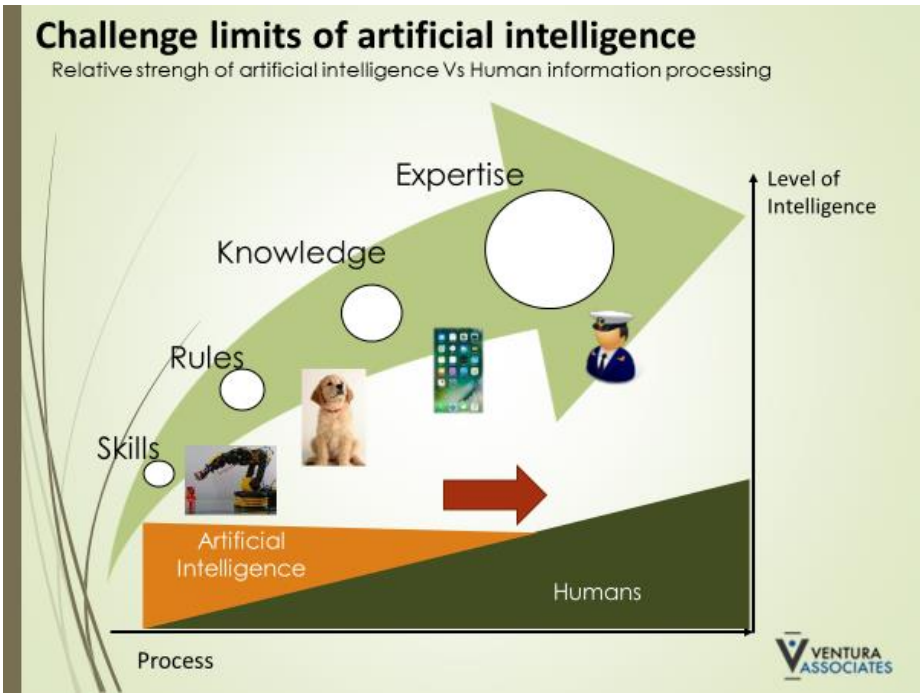


Fig. 1. The progressive penetration by AI of the Information Added Value Chain.

2 The Challenges of Maritime Big Data Analytics

2.1 How would we defeat the smart cheater?

As the development of massive intelligence gathering tools screening all the modern ICT systems is now a public evidence, ad hoc attitudes have soon developed for keeping undetected deviant behaviors by avoiding leaving evident triggers for the current tools used for Big Data Analytics of homeland security agencies: this combines using "clean" wordings, swapping equipment frequently, mastering the risk of leaving compromising metadata etc. Because none of the communication and information exchange systems would appear "safe" for undertaking illegal activities or terrorism, the safe way for them is to hide into the innocent crowd by controlling entirely the ab-

sence of alerting and tracking triggers in its own "numeric track". The same applies for the "professional" IUU fishing or the drug trafficking.

To implement a real AI capability, a MSA system should be fed not only by the multiple ship declarations (AIS, but also voyage data, cargo data, crew lists, catch reports, log books etc.) but also by data and information that **cannot be possibly spoofed or tampered**. One of the main challenge for that will be the access to the original information as captured by the primary sensors, before being translated and streamlined to be exchanged across systems.

For example, this would be the case for gathering in the S-AIS payload characteristics of the VHF signal burst carrying the AIS message, then possibly associated as a metadata to the message itself. Properly processed with the knowledge of the precise satellite course, these additional features of the physical signal might allow detecting an inconsistency with the declared position.

A Maritime Big Data primarily fed by declarative information would soon face educated cheaters able to provide perfectly clean profiling results, as it is already the case for the organized crime community in our societies. Our prime focus shall be the Big Fish, not just the inadvertent and unsophisticated rascal trying to round-up his meager fish catch with an occasional but far more visible trafficking...

2.2 The developing Maritime Information Sharing Environments

In the same time, a massive step forward in terms of easing data access is underway with the steady progress of the Common Information Sharing Environments (CISE [1, 2] in EU, MNMIS [3] in NATO, MISE [4] in the US). Conjugated with the full dematerialization of commercial data (ship voyage data, cargo lists, fishing logs, crew and passengers lists etc.), the generalization of common ontologies and data exchange formats is almost achieved; the access nodes are under testing including the "translators" to interface legacy data repositories and operational MAS systems.

But will they be really used?

3 The pivotal role of proper CONOPS

The response resides in investigating as early as possible the potential Concepts of Operations (CONOPS)/Concepts of Use (CONUSE) as a guidance for developing appropriate big data analytics: how far detecting a suspicious aggregate of "weak signals" could really help to prevent a criminal undertaking to succeed? What is the operational time line of an actionable information, and its "use-by-date" (when it will become useless as too late to modify the course of events)? What is the sufficient level of certainty of a would-be criminal pattern to mobilize a Maritime Patrol Aircraft? To dispatch a helicopter with a sniper? To board an inspection team on a fast craft? Will the availability of drones modify this threshold?

It is a common issue in anti-terrorism or in fighting against illegal activities that weak signals obtained from big data analytics may help to flag a potential threat alert, but are not ascertained and detailed to the point of justifying immediate counter-action. This means acknowledging that this massive investment in the Big Data Analytics is not allowing preventing all criminal endeavors, as proven in almost all recent terrorists attacks in Western democracies.

It is not only true for intelligence: we all know in our small community that oil spill detection services from space radar imaging are not often resulting into a verification flight because the false alarm rate is too high compared to the burden of triggering a specific reconnaissance flight.

Big Data Analytics supported by AI represent currently such a massive industry development we cannot ignore. We are here to think on what we might achieve by browsing efficiently the incoming massive data access enabled by the cross-sectoral and cross-border maritime information sharing platforms; but we must first of all bring the User into the driving seat. Developing new tools just because we have the technology but with a too naïve view on how it might be used (by whom, for which benefit etc.) would add to the too many heavily subsidized "experimentations" never followed by a true transformation of the User community tools and operational practices.

The main message of this paper is that we have to work-out seriously and as a very first step **the Uses Cases angle**: what sort of ship/cargo/voyage/crew profiling will create a more effective CONOPS, who needs it, how often, what is the "Quality of Service" required etc. are the questions we must answer before starting the first line of software code.

It is worth noting this precedence of CONOPS and the opportunistic approach of non-invented-here technologies is indeed exactly what makes terrorists and organized crime always one step ahead of us today…

4 The "Smart Command" Approach

Ventura has developed an original concept called "SMART COMMAND" to initiate the development of AI /Human-based solutions. Such modules, are responding to the commanders needs to have an updated analysis of enemy courses of actions. Once integrated as modules in a maritime C2 connected system, that kind of solution will allow decision makers to adapt their postures and elaborate accurate own courses of actions. The operational analysis provided will integrate big data week signals (bottom-up) and updated behavior assumptions (top down).

Its corporate strategy is to ally a permanent senior expertise and analysis in maritime operations with ICT and military systems engineering. The operational experience is

provided by people having just left operational commanding roles, not offices. The daily confrontation of this operational view with the system view of our engineers and system architects is the only way to implement the **pivotal role of the User** for the whole loop of advanced Command Concepts developments.

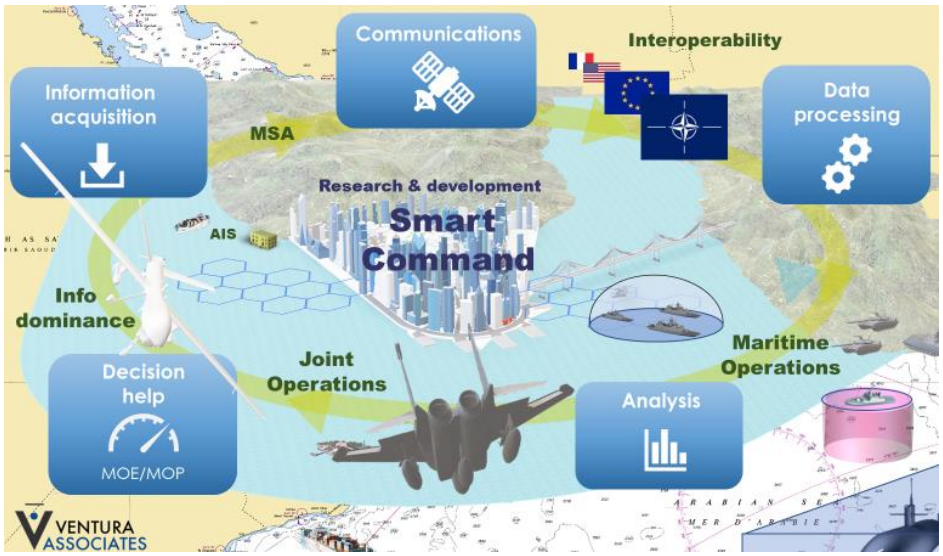


Fig. 2. The Smart Command data cycle from acquisition to decision help.

5 To conclude: be a smart follower

To tackle the Maritime Big Data challenge posed by this Forum, our plan is to revert the precedence of technology: what we all have experienced when involved into Defense Procurement processes is to handle new systems, and only then try to rethink our concepts of operation to take the best of this new technological advantage. We are now convinced that the future facing us as summarized above requires reverting totally the approach: develop the future concepts of operation first, then turn opportunistic in terms of Big Data technology acquisitions! We (Defense) have to accept that in the Big Data and AI emerging technologies, **we can only be "followers"**: at least we have to be smart followers knowing exactly **what effect we need to achieve** before starting benchmarking what we might find for that on the shelves...

References

1. COM(2010) 584 of 20.10.2010 – Communication on a Draft Roadmap towards establishing the Common Information Sharing Environment for the surveillance of the EU maritime domain [accessible at: https://ec.europa.eu/maritimeaffairs/sites/maritimeafairs/files /docs/body/integrating_maritime_surveillance_en.pdf]

2. CISE Impact Assessment Study – Executive Summary SWD(2014) 224 final of 8/7.2014 [accessible at: http://eur-lex.europa.eu/legal-cotent/EN/TXT/PDF/ ?uri=CELEX:52014SC0224&from=EN]
3. MNMIS ; Multinational Maritime information sharing as in AC/224(JCGC2)N(2017)0001 – Smart defense initiative 2.33.
4. Maritime information sharing environment portal accessible at https://www.hsdl.org/?abstract&did=719137.

A Perspective on Applied Human Factors in Support to the Maritime Big Data challenge

Francesca de Rosa[1] and Anne-Laure Jousselme[1]

[1] NATO STO – Centre for Maritime Research and Experimentation, Viale San Bartolomeo 400, La Spezia, Italy

The maritime domain, such as many other domains (e.g. crises management, health, first aid and traffic control), is increasingly moving toward automation, which entails that some of the cognitive tasks that underpin decision-making are progressively performed by Decision Support Systems, which seek to support Maritime Situational Awareness (MSA). One of the reasons for the need for automation can be found in the challenges posed by the increased *volume*, *velocity*, *variety* and possible lack of *veracity* of the information at hand, which would require an information processing capacity often beyond human ability. However, it is important to underline that Decision Support Systems are not surrogate of the human element, as they provide assistance in problem-solving and decision-making, acting as "enabler[s], facilitator[s], accelerator[s] and magnifier[s] of human capability, [but] not [as] its replacement" [1]. Therefore, the operational environment includes both technological elements (hardware and software) and human elements [2]. Those human elements do not only determine the procedural component of the environment, but might also play several concurrent roles, such as "decision maker, monitor, information processor, information encoder and storer, discriminator, pattern recognizer [,] . . . ingenious problem solver" [3] or disseminator. Therefore, Decision Support Systems could greatly benefit from a holistic design approach in which the human component is central [4, 5].

Many Human Factors (HF) methods have been developed with the scope of assessing different aspects of human-machine interaction, such as the ones that focus on usability, physical tasks, cognitive tasks, processes, human errors, mental workload or the assessment of Situational Awareness (SAW) [6]. Although, some of the HF methods for SAW assessment (e.g. SAGAT [7]) have been validated and extensively used [6], those methods are not able to directly respond to the call for systems grounded on intelligible and possibly intuitive reasoning and communication schemes, which ensure algorithms transparency and accountability [8].

This work will describe how the authors developed some innovative HF methods [9, 10], with the final goal of supporting standardization efforts on communication protocols and the design of transparent automated reasoners and information fusion algorithms, mimicking the human information combination process in a multisource context. The above mentioned methods aim at eliciting through a gamified approach the impact of uncertainty factors related to information and source of information on human Situational Assessment processes and final Situational Awareness [11]. While the Risk Game [9] mainly explores the impact of information quality on human SAW, the Reliability Game [12] focuses on the characterization of source factors (e.g. type and quality) impact on human Situational Assessment process. More specifically, during the game session, participants are provided through the use of cards with information and possible additional meta-information on source factors. Through the

cards the participants are requested to assess the current scenario situation. Due to the specific game board and question design, the game allows collecting data on player's *source quality* ratings, single information assessment (expressed through the card position on the board) and *confidence* in the assessment ratings. This data gathering technique has been validated both as an engaging and effective method [12], that allows collecting data that could be used to directly inform the design of fusion systems and the standardization of communication protocols.

References

1. Stikeleather, J.: Big data's human component. Harvard Business Review, 2012, https://hbr.org/2012/09/big-datas-human-component
2. Christensen, J.: The Nature of Systems Development. In: Human Factors Engineering: Engineering Summer Conferences. University of Michigan, Ann Arbor (1985)
3. Pew, R.: Human Skills and Their Utilization. In: Human Factors Engineering: Engineering Summer Conferences. University of Michigan, Ann Arbor (1985)
4. Nemeth, C. P.: Human Factors Methods for Design: Making Systems Human-Centered. CRC Press, Boca Raton (2004)
5. Hall, D. L., Jordan, J. M.: Human-centered Information Fusion. Artech House, Boston (2010)
6. Stanton, N. A., Salmon, P. M., Walker, G. H., Baber, C., Jenkins, D. P.: Human Factors Methods: A Practical Guide for Engineering And Design. Ashgate Publishing Company, Brookfield (2006)
7. Endsley, R. M.: Measurements of situation awareness in dynamic systems. Human Factors, 37 (1), 65—84 (1995)
8. Rainie, L., Anderson, J.: Code-dependent: Pros and cons of the algorithm age. Technical report, Pew Research Center, Washington DC (2017)
9. Jousselme, A.-L., Pallotta, G., Locke, J.: A Risk Game to study the impact of information quality on human threat assessment and decision making. Techical Rep. CMRE-FR-2015-009, NATO STO Centre for Maritime Research and Experimentation, La Spezia (2015)
10. De Rosa, F., Jousselme, A.-L., De Gloria, A.: Gamified Approach in the Context of Situational Assessment: a Comparison of Human Factors Methods. Advances in Human Factors, Software, and Systems Engineering, Proceedings of the 9th International Conference on Applied Human Factors and Ergonomics, Springer (in press)
11. Endsley, R. M.: The application of human factors to the development of expert systems for advanced cockpits. In: Human Factors Society 31st Annual Meeting, pp. 1388—1392. Human Factor Society, Santa Monica (1987)
12. De Rosa, F., Jousselme, A.-L., De Gloria, A.: The Reliability Game for Source Factors and Situational Awareness Experimentation (submitted for publication)

Deep learning-based Classification for Marine Big Data Analysis

Emna Hachicha Belghith[1], François Rioult[1], and Medjber Bouzidi[2]

[1] Laboratoire Greyc, UMR 6072, Université de Caen Normandie, France
{firstname.lastname}@unicaen.fr
[2] Sinay Marine Company, Caen, France
medjber.bouzidi@sinay.fr

1 Introduction

Marine Data management is taking advantage of the growth of the Big Data paradigm in order to achieve good performance level on acoustic data. The combination of big data and Marine data, referred to as marine big data, is recently drawning the attention of the R&D community, specifically in analyzing and classifying acoustic sounds. In such context, several projects have been initiated to manage such data e.g., Argo [1], the Green Marine [2], AIMS [3].
Despite this attention, limited work exists to analyze marine big data and make efficient evaluation of the human offshore activities on marine fauna [4]. Different approaches for classifying marine data have been proposed so far, mainly with a focus on separately treating fish, whale or vessel signals [5, 6]. Even though machine learning, in particular deep learning technique, is highly recommended in such context, there has been hardly any uptake in that area. The existing proposals do not cover various marine data such as invertebrates, wind, rain.
In this work, we propose a deep learning-based approach for classifying acoustic sounds, towards an automated support of marine sound analysis in big data architectures. We define a classifier, based on deep convolutional neural network, that covers three audio data types: (i) Biophonia (i.e., living species), (ii) Geophonia (i.e., natural phenomenons) and (iii) Anthropophonia (i.e., human activities). Our approach have been validated through experiments using a real marine dataset from an industrial partner.

2 Approach Overview

We show an overall overview of our Marine Big Data approach in Fig. 1. The left part presents three type of inputs: (i) Biophonia, (ii) Geophonia, and (iii) Anthropophonia.These data constitute the main components of an acoustic landscape that covers underwater sounds. The left part represents the architecture of the 4 Vs of Big data (i.e., Variety, Velocity, Volume and Veracity) [7]. While the buttom part shows the different phases of our established model, that is based on deep convolutional neural network. We advocate that our approach helps, by correctly classifying Marine Data, in addressing the various challenges related to the 4 Vs of Marine Big data.

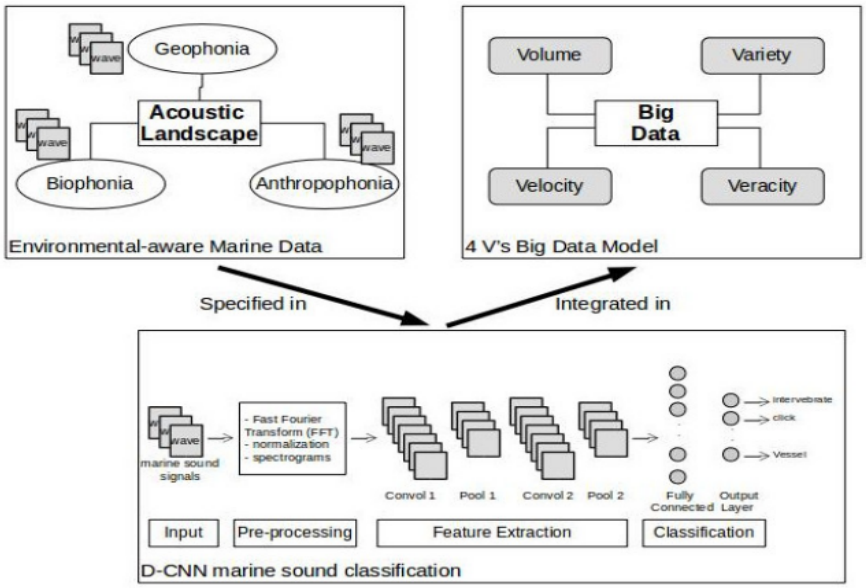


Fig. 1. Approach Overview

2.1 Marine Sound Classification

We briefy describe the steps for the implementation of our model: Our dataset consists of a set of 10,296 wave signals harmoniously divided into ten classes: noise, rain, wind, whistling, click, drum, kwa, intervebrate, vessel, sonar.

We have applied a set of operations: sampling, Fast Fourier Transform (FFT), and normalization. Afterwards, the spectrograms are transformed into images, on which a deep convolutional neural network (CNN) [8] is applied.

We fit our model with our dataset using the proportion of 80/20 of training/testing data. We have computed the accuracy metrics for our 3-layers model and 1-layer shallow model. Table 1 shows the obtained values of the accuracy metrics. We denote that our model outperforms the shallow one with 68,3%, 52.4% compared to 37.4%, 29.8% in terms of training and testing accuracy.

Table 1. Accuracy Metrics

	Training Accuracy	Testing Accuracy
D-CNN	68.3 %	52.4 %
Shallow Model	37.4 %	29.8 %

3 Conclusion

We conclude that we have developed a deep learning-based classification approach that takes into account the variety of underwater sounds, covering a wide acoustic landscape. Our aim is to automatically support marine sound analysis in big data architectures. Further, we conduct experiments, using a real marine dataset, that validate our proposal.

Acknowledgements

This work is supported by the European Project FEDER/FSE 2014-2020 [3].

References

1. : The international argo project. (http://www.argo.net/)
2. : The green marine project,. (https://www.green-marine.org/)
3. : Acoustic integrated monitoring system,. (http://aims.sinay.fr/aims-portal)
4. Huang, D., Zhao, D., Wei, L., Wang, Z.: Modeling and analysis in marine big data: Advances and challenges. Mathematical Problems in Engineering (2015)
5. Das, A., Kumar, A., Bahl, R.: Marine vessel classification based on passive sonar data: the cepstrum-based approach. IET Radar, Sonar Navigation **7** (2013) 87–93
6. Pourhomayoun, M., Dugan, P., Popescu, M., Risch, D., Lewis, H., Clark, C.W.: Classification for big dataset of bioacoustic signals based on human scoring system and artificial neural network. CoRR **abs/1305.3633** (2013)
7. Mauro, A.D., Greco, M., Grimaldi, M.: A formal definition of big data based on its essential features. Library Review **65** (2016-03-01) 125–135
8. Bengio, Y., Lecun, Y. In: Scaling learning algorithms towards AI. MIT Press (2007)

Estimating fishing effort using AIS data: an application to the European fishing fleet

Maurizio Gibin[1], Fabrizio Natale[2], Alfredo Alessandrini[2] Michele Vespe[2], Giacomo Chato Osio[1]

[1] D.2 Fisheries and aquaculture, Directorate D – Sustainable resources, Joint Research Centre, Ispra, Varese, Italy
[2] E.6, Knowledge Centre on Migration and Demography, Joint Research Centre, Ispra, Varese, Italy

Abstract. The introduction and diffusion of Automatic Information Systems for identification and tracking, presents the opportunity to quantify pressure on the marine ecosystem deriving from human activities such as shipping and fishing. However, estimating fishing effort from AIS data needs further investigation to ascertain whether AIS data can be used as a statistically sound proxy of the real fishing effort or it can only be used when coupled with VMS and Log-book data.

Keywords: AIS, Fishing Effort estimation, Modelling.

1 Introduction

Fishing effort is defined as fishing capacity times fishing activity, calculated on the basis of time spent in a well-defined area. Vessel Monitoring System (VMS) and logbooks are used to actively monitor the activity of fishing. Harvest control rules are used in the Common Fisheries Policy (CFP)as a measure to limit the amount of fishing effort in particular area temporary or indefinitely.

Fishing effort recorded by logbooks (and not VMS) is collected by Member States and then submitted to the EU Data Framework Collection (DCF). The effort data is distributed through the DCF data dissemination website [1] at a resolution of 1 by 0.5 degrees using the ICES rectangle geography, which covers most of FAO Area 27. Such resolution however is too coarse when assessing the impact of fishing on the marine ecosystem.

Commission Decision of the 18 December 2009 Appendix XIII [2], the European Commission lays a set of environmental indicators aimed at quantifying the effect of fisheries on the marine ecosystem. Among the nine indicators proposed, indicator 5 (Distribution of fishing activities), 6 (Aggregation of fishing activities) and 7 (Area not impacted by mobile bottom gears) are specific to fishing effort. Appendix XIII indicates that the data source used for the calculation of the indicators should be VMS data of more than 15 meters long vessels with a time resolution of thirty minutes and with a DCF level 6 métier classification [3].

The main goal of the Marine Strategy Framework Directive (MSFD) is to achieve a Good Environmental Status (GES) for marine waters that will “provide ecologically

diverse and dynamic oceans and seas which are clean, healthy and productive" by 2020 [4]. The MSFD sets out a series of descriptors aimed at measure and quantity the status of the marine ecosystem. Descriptor 6 [5] in particular verifies that "Sea-floor integrity is at a level that ensures that the structure and functions of the ecosystems are safeguarded and benthic ecosystems, in particular, are not adversely affected". The activity of fishing represents a pressure on the marine ecosystem with impacts varying according the fishing gear used. Mobile towed gears and in particular trawled gears have the highest impact on the sea-floor.

An EU wide accessible and statistically sound high resolution fishing effort layer, provided in a familiar data format for spatial analysis, would facilitate the calculation of DCF 5, 6 and 7, MSFD D6 and the monitoring of fine scale fishing activities.

The introduction and diffusion of Automatic Information Systems for identification and tracking and Search and Rescue (SAR) in maritime safety, presents the opportunity to quantify pressure on the marine ecosystem deriving from human activities such as shipping and fishing. AIS data is not subjected to confidentiality and can be requested for research purposes or acquired from commercial vendors. We briefly present the creation of the first high resolution fishing effort layer for EU waters, estimated using AIS data.

2 Data

The data used in the analysis are historical AIS data from the 1st October 2014 to the 30th of September 2015 and covers EU waters and the EU fishing fleet of more than 15 meters long with information on speed, direction, time and geographical coordinates of each fishing vessel [6]. After the regular cleaning routines (unacceptable geographical coordinates values, unrealistic speed and direction values) the data have been reduced to a five minutes' resolution and linked to the European Fleet Register. The clean dataset was then formatted into a comma separated value file and used in the statistical software R for analysis and QGIS for mapping.

3 Methodology

Fishing effort estimation using AIS data follows the same modeling approach used for VMS data. Most of published research [7] on estimating fishing effort using VMS and logbook data focus on mobile towed gear and much work is needed in modeling static gears (pots and traps). The underlying principle for the estimation of fishing effort for mobile towed gears is that fishermen, while at sea, will try to maximize their time spent fishing (rational behavior). The activity of fishing can be roughly divided in at least three main components: in-port, fishing and steaming [8]. For demersal (bottom contacting) trawled gears the fishing component will have speed values lower than steaming but higher in density because of the rational behavior of fishermen. The in-port component instead will have lower or close to zero speed values with a concentration of AIS messages in the proximity of the ports.

The fishing component represents that part of the vessel track where the trawled gear has been used. Several approaches have investigated the identification of the fishing component and the corresponding fishing speed values [7]: from simple arbitrary thresholds, to more sophisticated approaches where fishing is a stochastic process following a statistical distribution. Our methodology is based on the assumption that the components follow a normal distribution and the resulting joint distribution can be intended as a Gaussian Mixture Model (GMM) [8] (Fig.1).

Fig. 1. The histogram of the speed values for a Pair Trawler Bottom in the period 1st October 2014 - 30th September 2015 from Vespe et al. [6]. The in-port component was excluded by filtering out messages at a specified distance from the coast and below 0.5 knots.

Through the Expectation-Maximization (EM) algorithm we identified the fishing component for every fishing vessel's track in the period considered. The fishing speed values were selected in the range $\mu \pm k * \sigma, with\ k = 1$, of the fishing component distribution. The remaining messages have been aggregated in a grid of 1 km by 1 km.

A land mask was applied to the raster grid, excluding all those messages at less than 0.02 degrees from land. All cell that had a count of one and a value of one were filtered out. Finally, we assess the statistical distribution of the maximum values and

we select the ones with the highest values with the lowest count reclassified using the focal mean of its first neighboring cells in all directions.

4 Results and Discussion

Figure 2 shows the rendering of the final output of the model. The effort layer is distributed, as raster file since January 2018, through the Joint Research Open Data Catalog [9].

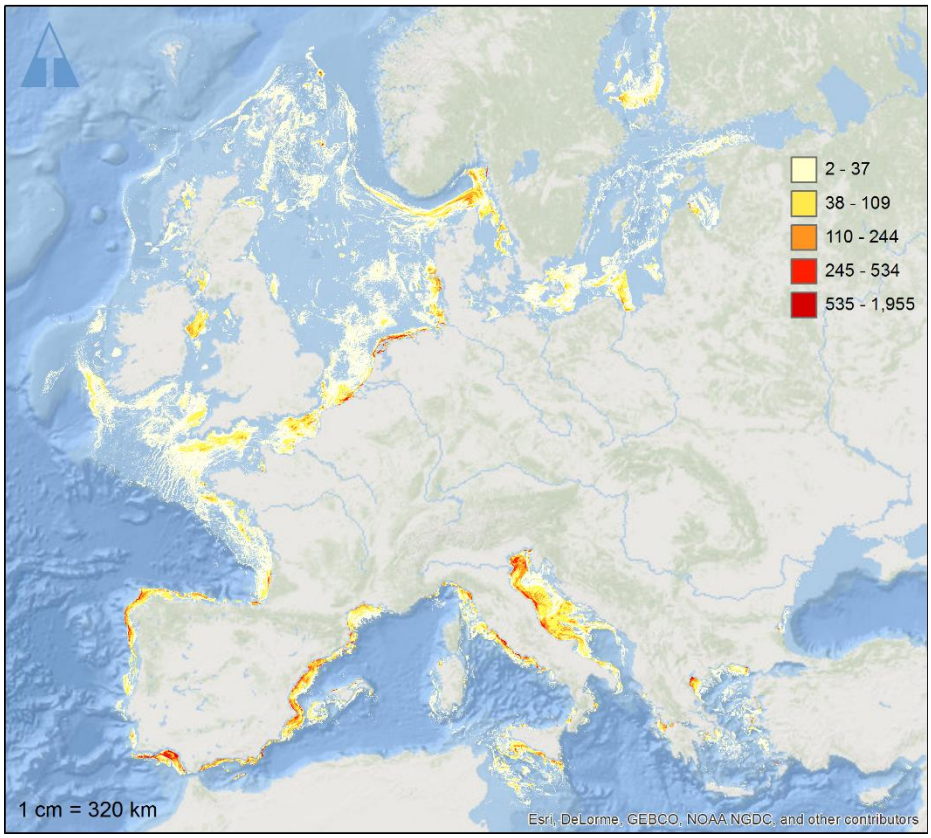


Fig. 2. A map showing the final output of the model: the number of points classified as fishing per square kilometer. Fishing effort estimation is aggregated and distributed as a raster dataset through the Joint Research Centre (JRC) Open Data Catalogue: https://data.jrc.ec.europa.eu/

The distribution of fishing effort is concentrated around the coast and in the continental platform, confirming that most of the fishing activity of the large fleet does not occur in deep waters. The area around the coast of Norway follows the shape of the Exclusive Economic Zone (EEZ) and in the Baltic Sea the Russian EEZ is visible, showing the lack of data for non-EU countries.

The fishing effort layer should not be used as a precise measure of fishing effort but rather like an estimation of where most of the fishing is concentrated. AIS data are affected by noise and coverage issues. AIS coverage can be assessed using ancillary sources [6] and improved by acquiring terrestrial and satellite data.

The initial settings of the fishing effort estimation model are not robust. The distance from the coast used to identify the in-port component is a sensitive parameter that can filter out a large number of AIS messages. The range of the fishing speed values varies if we consider different values for the scalar *k.*

A brief validation of the results was carried out during the workings of the International Council for the Exploration of the Sea (ICES) working group on spatial fisheries data (WGSFD) in May 2016 [10]. The analysis, proved to be difficult due to the lack of a proper comparable dataset and showed that AIS effort tends in total to underestimate VMS effort and that gear attribution with the fleet register tends to overestimate the share of Otter Trawler Bottom and underestimate the other gears. The working group report advises further investigation to establish if AIS should only be used to improve the time resolution of VMS or AIS data derived effort can be used to quantify the impact of fishing on the marine ecosystem. The WGSFD recommendation is now more pressing especially in light of recent published research and data [11]. Global Fishing Watch [12] a Google founded initiative, created a global layer of fishing footprint estimated from AIS data and is providing free online public access to fishing vessels positions.

Before fishing effort layers derived only from AIS data can be used in EU policy applications two main changes are needed: the creation of a common modeling framework to accommodate all type of gears and the related initial sensitive settings. Although AIS data can be used for monitor and control purposes by EU Member States; VMS data remains the official source of fishing effort used by fishing authorities in the management of the Common Fisheries Policy. If AIS data is used as a fisheries control tool at EU level for the large fleet, Member States will more likely to expect the same level of confidentiality and data access limitations that VMS data have.

Acknowledgements.
The authors would like to thank MSSIS, courtesy of the Volpe Center of the U.S. Department of Transportation and the U.S. Navy, and MarineTraffic for providing the AIS data used in this study.

References

1. The DCF Data Dissemination website: https://stecf.jrc.ec.europa.eu/data-dissemination
2. Appendix XIII ,https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32010D0093
3. The DCF métier classification, https://datacollection.jrc.ec.europa.eu/c/document_library/get_file?uuid=296dffd3-9c81-4759-b691-9b1654ea66b9&groupId=10213#page=16

4. The MSFD webpage: http://ec.europa.eu/environment/marine/good-environmental-status/index_en.htm
5. MSFD Descriptor 6: http://ec.europa.eu/environment/marine/good-environmental-status/descriptor-6/index_en.htm
6. Vespe, M., Gibin, M., Alessandrini, A., Natale, F., Mazzarella, F., Osio, G.C.: Mapping EU fishing activities using ship tracking data, Journal of Maps (12), 520-525, (2016) DOI: 10.1080/17445647.2016.1195299
7. Lee, J., South, A. B., and Jennings, S.: Developing reliable, repeatable, and accessible methods to provide high-resolution estimates of fishing-effort distributions from vessel monitoring system (VMS) data. – ICES Journal of Marine Science, 67: 1260 – 1271, (2010)
8. Natale F, Gibin M, Alessandrini A, Vespe M, Paulrud A: Mapping Fishing Effort through AIS Data. PLoS ONE 10(6): e0130746, (2015) . https://doi.org/10.1371/journal.pone.0130746
9. The file is available at the following URL: https://data.jrc.ec.europa.eu/dataset/jrc-fad-ais1415
10. ICES: Interim Report of the Working Group on Spatial Fisheries Data (WGSFD), 17–20 May 2016, Brest, France. ICES CM 2016/SSGEPI:18. 244 pp, (2016).
11. Kroodsma, D.A., Mayorga, J., Hochberg, T., Miller, N.A., Boerder, K., Ferretti, F., Wilson, A., Bergman, B., White, T.D., Block, B.A., Woods P., Sullivan, B., Costello, C.,Worm, B.: Tracking the global footprint of fisheries, Science, Vol. 359, Issue 6378, pp. 904-908 (2018) DOI: 10.1126/science.aao5646.
12. Global Fishing Watch: http://globalfishingwatch.org/

MERIDIAN is Listening to the Sounds of the Deep Ocean with Deep Learning

Fabio Frazao, Renata Dividino, Ryan Gosse, Ines Hessler, Oliver Kirsebom, Johna Lautof, Kim Mortimer, Erico de Souza, Gary Blades, and Stan Matwin

Dalhousie University, Halifax NS, B3H 4R2, Canada
{firstname.lastname}@dal.ca
www.meridian.cs.dal.ca

Abstract. Ocean development must be done sustainably, which includes controlling and/or mitigating the impact of noise. In order to enable ocean scientists to fully exploit Canada's acoustics ocean data, to monitor trends in the state of the ocean acoustic environment, and to allow more timely, effective and efficient protection of valued marine species and protected areas, the MERIDIAN consortium is developing a research data infrastructure to consolidate and support ocean acoustic data management and analytics. This infrastructure will be a widely used resource for the Canadian and international academic community to drive discovery and innovation while supporting the development of the Ocean Soundscape Atlas for Canada.

Keywords: Advanced Data Sciences · Big Data Management · Ocean Noise · Ocean Soundscape Atlas.

1 Introduction

Ocean noise from shipping and other offshore industrial activities is becoming a significant issue due to its potential impacts on protected marine species, especially great whales and other marine mammals who use sound to communicate, sense their environment, navigate, and feed. Future developments in the ocean must be done sustainably, and must advance in the field of analytics and Big Data to improve the ability of ocean scientists to exploit Canada's ocean waters and to monitor trends in the state of the ocean acoustic environment. Under the Canada Foundation of Innovation Cyberinfrastructure program, our Marine Environmental Research Infrastructure for Data Integration and Application Network (MERIDIAN) project focuses on the development of data services based on advanced data science methods and algorithms, visualization tools and techniques, and best practice guides for Big data management to establish a national data resource for noise in the ocean. These tools are valuable for academic community as well as for governmental institutions and the industry, for monitoring trends in the state of the ocean acoustic environment, and will enable more timely, effective and efficient protection of these valued marine species.

Currently, existing Canadian acoustic data and ancillary environmental data are diffused and dispersed. MERIDIAN research infrastructure will federate, collate and integrate existing disparate sources of information on noise in the ocean. MERIDIAN's expertise contribution to **Big data management** for ocean data is based on common formats, standard metadata and semantic technologies. Agreements on common core metadata are necessary to ensure data interoperability and to allow researchers to effectively manage and (re-)use marine acoustic data and related products in a larger ocean context. The development of the MERIDIAN metadata standards will be aligned with the standards already adopted or that are current under development by the Ocean data management community (e.g. the Canadian Integrated Ocean Observatory System [2] (CIOOS), Community of Practice on Ocean Data Management [1] (CoP ODM)). MERIDIAN's main goal is to make high-quality ocean data findable, accessible, interoperable and re-usable to serve the scientific community.

Advanced analytics in form of sophisticated techniques and tools will be developed to advance discovery. In particular, we are going to provide data analytics strategies and methods based on Machine Learning and Deep Learning for the detection, classification and localization (DCL) of whales. These tools will automatically extract and classify these specific vocalization and dialects from large acoustic data sets or streaming data produced by survey equipment. This system is under development and a demonstration can be provided during the workshop. Furthermore, and also under development, MERIDIAN will provide data analytics based on Machine Learning and Deep Learning tools to advance acoustic predictions from shipping tracking data [3](using the global satellite Automatic Identification Systems - AIS - data streams).

Furthermore, MERIDIAN will adopt, adapt and develop **visualization tools** necessary for the benefit of the scientific research, socio-economic and regulatory sectors. MERIDIAN is working on a modern, interactive, and animated multi-dimensional visualization tool of acoustic fields from AIS sources, the Canadian Soundscape Atlas. This 3D visualization of basins (i.e. in relation with water masses structure, topography, bottom properties, and whale exposure to critical levels) will help to better illustrate and understand the interaction of vessel noise with whales [4].

References

1. Wilson, L., Smit, M., Wallace, D. W. R.:Towards a Unified Vision for Ocean Data Management in Canada: Results of an Expert Forum. (2016) http://hdl.handle.net/10222/72192.
2. Lenore B.: The developing Canadian Integrated Ocean Observing System (CIOOS). Pensoft Publishers, vol. 2, (2017) https://doi.org/10.3897/tdwgproceedings.1.20432.
3. de Souza E., Boerder K., Matwin S., Worm B.: Improving Fishing Pattern Detection from Satellite AIS Using Data Mining and Machine Learning. PLoS ONE 11(7), e0158248, (2016) https://doi.org/https://doi.org/10.1371/journal.pone.0158248
4. Aulanier, F., Simard, Y., Roy, N., Gervaise, C., and Bandet, M.: Effects of shipping on marine acoustic habitats in Canadian Arctic estimated via probabilistic modeling and mapping. Mar. Poll. Bull. 125(1): 115-131 (2017).

Scalable Spatio-Temporal Analysis through Open Standards: The European Datacube Engine

Peter Baumann[1,2]

[1] Jacobs University, Bremen, Germany
[2] rasdaman GmbH, Bremen, Germany
baumann@rasdaman.com

Keywords: Datacubes, Big Data Analytics, standards

1 Abstract

Datacubes [6][4][18] form an enabling paradigm for serving massive spatio-temporal Earth data in an analysis-ready way by combining individual files into single, homogenized objects for easy access, extraction, analysis, and fusion - "a cube says more than a million images". In common terms, goal is to allow users to "ask any question, any time, on any size" thereby enabling them to "build their own product on the go".

Originally, the concept of queryable multi-dimensional datacubes dates back to 1992 [5]. Today, large-scale datacubes are becoming reality: For server-side evaluation of datacube requests, a bundle of enabling techniques is known [20] which can massively speed up response times, including adaptive partitioning, parallel and distributed processing, dynamic orchestration of heterogeneous hardware, and even federations of data centers.

Looking at the concrete example of rasdaman ("raster data manager") [16,18,4,1] we find that its Array Database approach of enabling datacube services with a flexible query language allows for flexible, fast, scalable, and standards-based user services; still, rasdaman users do not need to see the query language, but may use their well-known clients [1], ranging from map image navigation (ex: OpenLayers, Leaflet) over Web GIS (ex: QGIS, ArcGIS) and virtual globes (ex: NASA WebWorldWind, Cesium) to analytics environments (ex: R, python) – see Figure 1 left.

Operational datacube services exceed 2.5 PB [1], and datacube analytics queries have been split across 1,000+ cloud nodes [8]. Intercontinental datacube fusion has been accomplished between ECMWF/UK and NCI Australia [1], as well as between ESA and NASA, using efficient fusion (in database lingo: *join*) techniques [2]. In the BigDataCube project, a public/private datacube federation between the German Copernicus hub, CODE-DE, and a commercial geo cloud service provider, cloudeo AG, is being established, with plans to extend this federation with international Earth data providers [7].

Research Data Alliance (RDA) in a deep investigation of functionality and performance [17] has clearly demonstrated rasdaman's leading position; for example, in the

performance comparisons done rasdaman can be orders of magnitude faster than competitors such as Open Data Cube, SciDB, and PostGIS Raster. Not the least this is due to the adaptive storage management given to administrators for service tuning [3] and the automatic multi-parallel CPU/GPU processing [9], see Figure 2. This performance has not only been described, but also demonstrated live at various occasions and conferences [8].

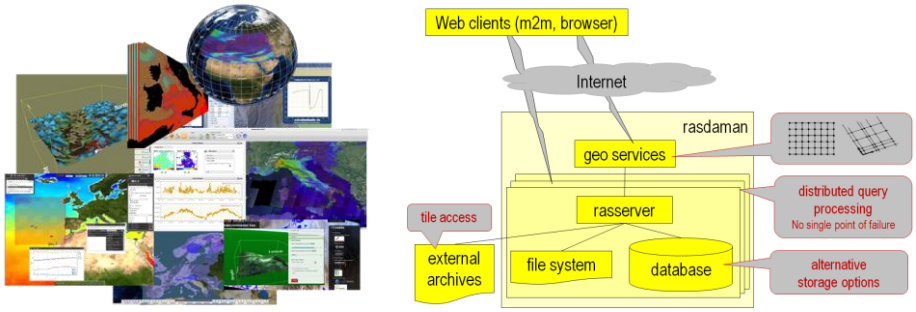


Fig. 1. Datacube portals powered by rasdaman (left) and rasdaman architecture (right).

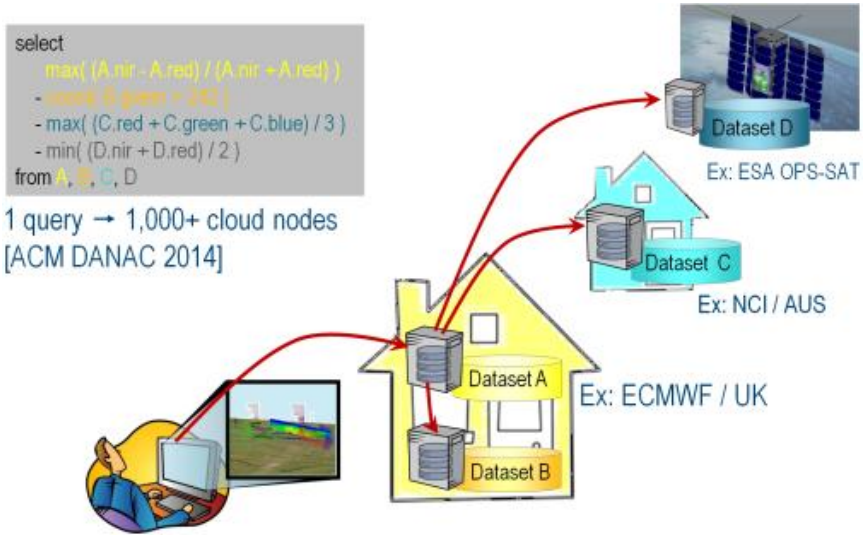


Fig. 2. Parallelization and federated processing in rasdaman.

From a standards perspective, datacubes belong to the family of coverages, as per ISO and OGC [14]; the coverage data model is represented by OGC Coverage Implementation Schema (CIS), the service model by OGC Web Coverage Service (WCS) [14] together with its OGC Web Coverage Processing Service (WCPS) [15], OGC's geo datacube query language. The rasdaman engine is official OGC Reference Implementation [13]. Additionally, ISO is finalizing MDA ("Multi-Dimensional Arrays") as an application-independent extension to SQL [11], based on rasdaman as its blueprint.

The ease of deployment and analytics enables service providers with a tight schedule to concentrate on their business model, rather than doing tedious own software development and maintenance. This has already given rise to startups using rasdaman as their platform, such as the Greek precision farming company, EOfarm [19].

In our talk we present the concept of queryable datacubes, the standards that play a role, as well as interoperability successes and issues existing, based on our work on the European Datacube engine, rasdaman, which is powering today's largest operational datacube services and, in April 2018, has been distinguished with the NATO Defence Innovation Award [12].

References

1. P. Baumann, A.P. Rossi, B. Bell, O. Clements, B. Evans, H. Hoenig, P. Hogan, G. Kakaletris, P. Koltsida, S. Mantovani, R. Marco Figuera, V. Merticariu, D. Misev, B. Pham Huu, S. Siemen, J. Wagemann: Fostering Cross-Disciplinary Earth Science Through Datacube Analytics. In. P.P. Mathieu, C. Aubrecht (eds.): Earth Observation Open Science and Innovation - Changing the World One Pixel at a Time, International Space Science Institute (ISSI), 2017, pp. 91 - 119
2. P. Baumann, V. Merticariu: On the Efficient Evaluation of Array Joins. Proc. Workshop Big Data in the Geo Sciences (co-located with IEEE Big Data), Santa Clara, US, October 29, 2015
3. P. Baumann, S. Feyzabadi, C. Jucovschi: Putting Pixels in Place: A Storage Layout Language for Scientific Data. Proc. IEEE ICDM Workshop on Spatial and Spatiotemporal Data Mining (SSTDM'10), December 14, 2010, Sydney, Australia
4. P. Baumann, D. Misev, V. Merticariu, B.P. Huu: Datacubes: Towards Space/Time Analysis-Ready Data. In: J. Doellner, M. Jobst, P. Schmitz (eds.): Service Oriented Mapping - Changing Paradigm in Map Production and Geoinformation Management, Springer Lecture Notes in Geoinformation and Cartography, 2018
5. P. Baumann: Language Support for Raster Image Manipulation in Databases. Proc. Int. Workshop on Graphics Modeling, Visualization in Science & Technology, Darmstadt/Germany, April 13 - 14, 1992
6. P. Baumann: The Datacube Manifesto. http://earthserver.eu/tech/datacube-manifesto, seen 2018-04-20
7. BigDataCube: www.bigdatacube.org, seen 2018-04-20
8. A. Dumitru, V. Merticariu, P. Baumann: Array Database Scalability: Intercontinental Queries on Petabyte Datasets (system demonstration). Proc. 28th Intl. Conf. on Scientific and Statistical Database Management (SSDBM), Budapest, Hungary, July 18 - 20, 2016
9. A. Dumitru, V. Merticariu, P. Baumann: Exploring Cloud Opportunities from an Array Database Perspective. Proc ACM SIGMOD Workshop on Data analytics in the Cloud (DanaC'2014), June 22 - 27, 2014, Snowbird, USA
10. EarthServer: www.earthserver.eu, seen 2018-04-20
11. ISO: FDIS SQL 9075:2018 Multi-Dimensional Arrays (MDA)
12. NITEC: 2018 Defence Innovation Challenge for SMEs and Academia. www.nitec.nato.int/defence-innovation-challenge, seen 2018-04-20
13. OGC: Compliance Navigation. http://www.opengeospatial.org/resource/products/stats, seen 2018-04-20
14. OGC: Coverages & Datacubes. http://myogc.org/go/coveragesDWG, seen 2018-04-20

15. OGC: Web Coverage Processing Service (WCPS) Language Specification. OGC document 08-068, 2008
16. rasdaman: www.rasdaman.org, seen 2018-04-20
17. RDA: Array Databases: Concepts, Standards, Implementations. Research Data Alliance, 2018, http://dx.doi.org/10.15497/RDA00024, seen 2018-04-20
18. P. Strobl et al: The Six Faces of the Datacube. Proc. ESA Big Data from Space Conference (BiDS), Toulouse, France, November 28 – 30, 2017
19. A. Tzotsos et al: A Datacube Approach to Agro-Geoinformatics. Proc. IEEE Agro-Geoinformatics, Fairfax, USA, August 2017
20. Wikipedia: Array DBMS. http://en.wikipedia.org/wiki/Array_DBMS, seen 2018-04-20

Maritime Anomaly Detection of Stealth Deviations from Standard Routes Applied to a Real-World Scenario

Enrica d'Afflisio, Paolo Braca, Leonardo M. Millefiori

Considering the possibility of coverage gaps (e.g., counterfeit AIS reports, AIS device shutdown, limited radar coverage, etc.), a maritime anomaly detection problem is studied assuming an Ornstein-Uhlenbeck (OU) mean-reverting stochastic motion model for the vessel dynamics.

Specifically, in [1, 2], we consider the anomaly detection problem where a certain vessel switches off its AIS transponder for a certain amount of time, in order to hide its deviation from a planned route, which is characterized by a nominal velocity. The vessel would then try to revert back to the planned route and to the original nominal velocity. The decision to take is whether a deviation happened or not, relying only upon two consecutive AIS contacts, i.e. the last contact before the AIS device shutdown and the first one after the AIS device reactivation. The proposed anomaly detection strategy is based on a hypothesis testing procedure that builds on the changes of the OU process long-run mean velocity parameter. Two hypotheses can be therefore envisioned: the first one that the vessel navigates according to the nominal condition and the alternative one that the vessel moves away from the nominal route once the AIS transponder has been shut down.

The OU process has been shown to be better suited to model the behavior of a significant portion of real-world vessel trajectories than respect to conventional models [3–9]. In this framework, the use of the OU model turns out to be a valuable tool when vessel information is not available, providing a good estimation of a ship's position and velocity, even after several hours. Specifically the OU stochastic process is used to represent the velocity of the vessel, with a long-run mean term representing the nominal velocity of the ship [3]. In other words, the velocity of the vessel is a modified Wiener process so that there is a tendency of the process to move back towards the long-run mean value, with a greater attraction when the process is further away from it.

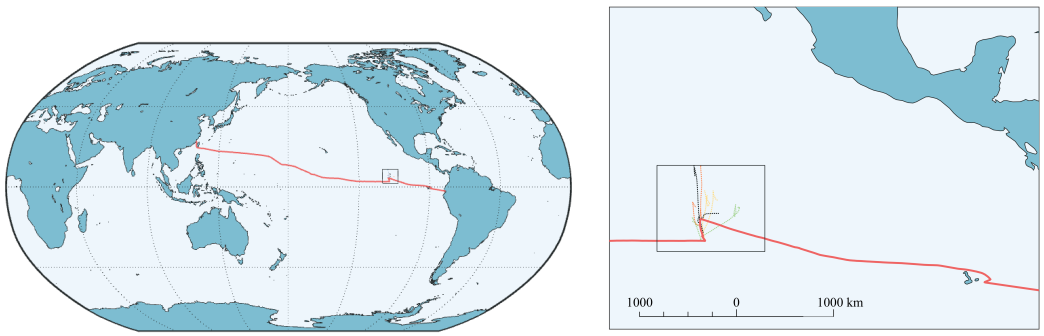


Figure 1: The track of cargo vessel reveals rendezvous with four fishing vessels in the Pacific.

The strategy proposed in [1, 2] is applied to a real world example of anomalous behavior corresponding to the about five-month track of a cargo vessel shown in Fig. 1. The vessel navigates with a nominal speed of about 5 m/s in the waters of the Pacific Ocean [10]. Nearby the Galápagos Exclusive Economic Zone (EEZ), the vessel shuts the engines down and starts drifting, with an apparent deviation from its route. The reason of this deviation of the vessel is to rendezvous with four tuna longliners at about 1700 miles away from the Galápagos. Each fishing vessel spends about 12 hours moving along the cargo vessel at a distance of about 30 m, which indicates the boats were likely tied up. This behavior suggests a substantial transfer of cargo was possible [10]. The AIS track related to the observation window confined to the area of concern, where the apparent anomaly behavior occurs, is

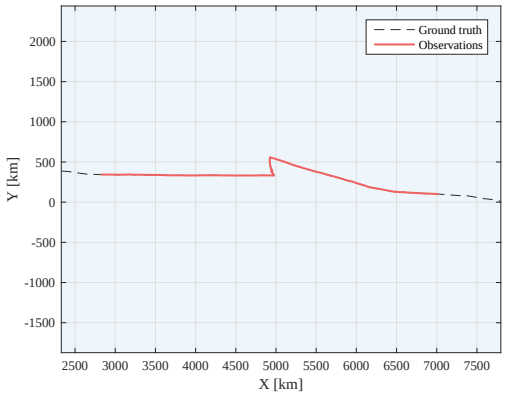


Figure 2: Complete AIS track.

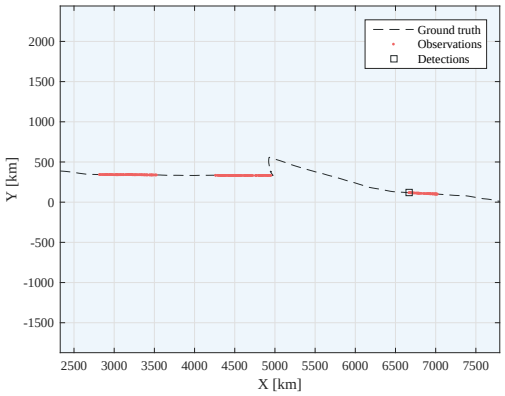


Figure 3: Simulated gaps in the AIS track.

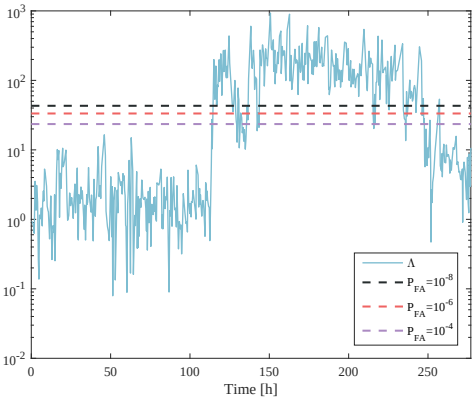


Figure 4: Test statistic related to the complete AIS track.

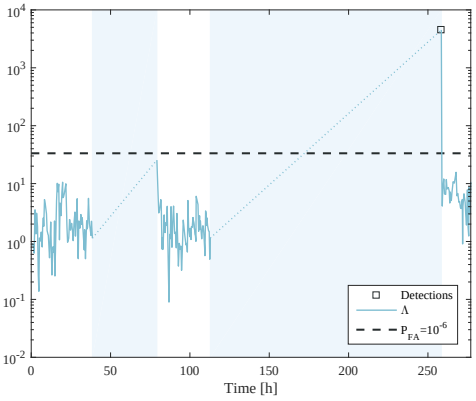


Figure 5: Test statistic related to the incomplete AIS track.

shown in Fig. 2, revealing a deviation from the normal route during a time frame of about 5 days. Fig. 4 displays the test statistic exceeding the threshold (plotted for different values of the false alarm probability) corresponding to the deviation from the nominal condition. The proposed detection strategy is tested with simulated gaps in AIS data, as shown in Fig. 3. In particular the first gap occurs in a section of the trajectory where there is no deviation from the nominal conditions, while the second one occurs where the deviation actually happens. As shown in Fig. 5, the deviation can be properly detected while no detection is correctly declared in the first gap.

Therefore, the anomaly detection algorithm appears to be useful in practical situations, as it could be applied automatically and simultaneously to several trajectories in order to reveal possible deviations, difficult to be identified by the simple visual inspection of a human operator.

References

[1] E. d'Afflisio, P. Braca, L. M. Millefiori, and P. Willett, "Detecting stealth deviations from standard routes using the Ornstein-Uhlenbeck process," *IEEE Transactions on Signal Processing*, submitted.

[2] ——, "Maritime anomaly detection based on mean-reverting stochastic processes applied to a real-world scenario," *Information Fusion (FUSION) 21st International Conference*, 10 - 13 July 2018.

[3] 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.

[4] 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, November 2016.

[5] L. M. Millefiori, P. Braca, and G. Arcieri, "Scalable distributed change detection and its application to maritime traffic," in *2017 IEEE International Conference on Big Data (Big Data)*, Dec 2017, pp. 1650–1657.

[6] P. Coscia, P. Braca, L. M. Millefiori, F. Palmieri, and P. Willett, "Maritime traffic representation based on sea-lanes graph construction criteria using multiple Ornstein-Uhlenbeck processes," *IEEE Transactions on Aerospace and Electronic Systems*, to be published, 2018.

[7] G. Vivone, L. M. Millefiori, P. Braca, and P. Willett, "Model performance assessment for long-term vessel prediction using HFSW radar data," in *2017 IEEE Radar Conference (RadarConf)*, May 2017, pp. 0243–0247.

[8] L. M. Millefiori, P. Braca, K. Bryan, and P. Willett, "Long-term vessel kinematics prediction exploiting mean-reverting processes," in *2016 19th International Conference on Information Fusion (FUSION)*, July 2016, pp. 232–239.

[9] G. Vivone, L. M. Millefiori, P. Braca, and P. Willett, "Performance assessment of vessel dynamic models for long-term prediction using heterogeneous data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6533–6546, Nov 2017.

[10] J. J. Alava, M. J. Barragán-Paladines, J. Denkinger, L. Muñoz-Abril, P. Jiménez, F. Paladines, and et al., "Massive Chinese fleet jeopardizes threatened shark species around the Galápagos marine reserve and waters off Ecuador: Implications for national and international fisheries policy," *International Journal of Fisheries Sci Res.*, 2017.

A multi-task deep learning model for vessel monitoring using AIS streams

D. Nguyen[1], R. Vadaine[2], G. Hajduch[2], R. Garello[1], and R. Fablet[1]

[1] IMT Atlantique, Lab-STICC, UBL, 29238 Brest, France
{van.nguyen1, rene.garrelo, ronan.fablet}@imt-atlantique.fr
[2] CLS - Space and Ground Segments, 29280 Brest, France
{rvadaine, ghajduch}@cls.fr

Abstract. AIS data streams provide new means for the monitoring and surveillance of the maritime traffic. The massive amount of data as well as the irregular time sampling and the noise are the main factors that make it difficult to design automatic tools and models for AIS data analysis. In this work, we propose a multi-task deep learning model for AIS data using a stream-based architecture, which reduces storage redundancies and computational requirements. To deal with noisy irregularly-sampled data, we explore variational recurrent neural networks. We demonstrate the relevance of the proposed deep learning architecture for a three-task setting, referring respectively to vessel trajectory reconstruction, abnormal behaviour detection and vessel type identification on a real AIS dataset.

Keywords: AIS · vessel monitoring · deep learning · abnormal behaviour detection · vessel type identification · recurrent neural networks

1 Context

In the modern world, maritime safety, security and efficiency are vital. For example, about 90% of the world trade is carried by sea, but only 2% of them is physically inspected. Vessel monitoring, therefore, becomes an essential demand. Besides that, the construction of a maritime situation map is also necessary for multiple purposes: security, smuggling detection, EEZ intrusion detection, transshipment detection, fishing activities control, maritime pollution monitoring, etc.

Over the last decades, the development of terrestrial networks and satellite constellations of Automatic Identification System (AIS) has opened a new era in maritime surveillance. Every day, AIS provides tens of millions of messages, which contain ships identification, their Global Positioning System (GPS) coordinates, their speed, etc. This massive amount of data would be very useful if the information contained inside could be extracted, analyzed and exploited effectively.

Several efforts have been conducted in order to create automatic/semi-automatic AIS analysis systems. The aims are to extract useful information from AIS data stream[9] [8], and use it for specific tasks such as maritime routes detection [7] [4],

vessel trajectory prediction [1] [10] or anomaly detection [3] [5]. However, those models depend on strong assumption, and can not capture all the heterogeneous characteristic of noisy, irregularly sampled AIS data.

In this work, we propose a multi-task model which explores deep learning, and more specifically recurrent neural networks to process AIS data stream for multiple purposes: trajectory reconstruction, anomaly detection and vessel type identification.

2 Proposed multi-task RNN model for AIS data

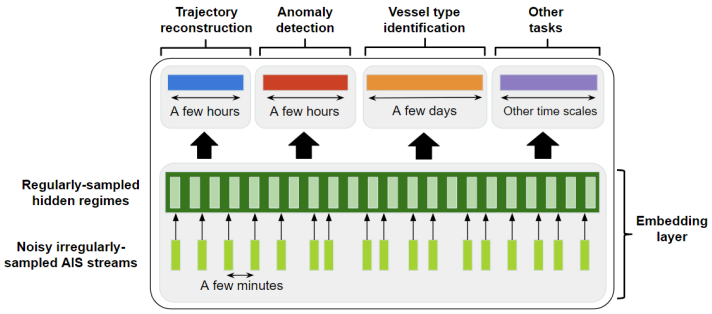


Fig. 1. Proposed RNN architecture.

As sketched in Fig. 1, we propose a multi-task Recurrent Neural Network (RNN) for the analysis of AIS data streams. The key component of this model is the embedding layer, which introduces hidden regimes. These regimes may correspond to specific activities (eg, under way using engine, at anchor, fishing, etc.). The embedding layer relies on a latent variable RNN [2]. It operates at a 10-minute time scale and allows us to deal with noisy and irregularly-sampled AIS data. Higher-level layers are task-specific layers at different time-scales (e.g., daily, monthly,...) to address the detection of abnormal behaviors, the automatic identification of vessel types, the identification of maritime routes,....

3 Results

We implemented the proposed framework for a three-task setting in the Gulf of Mexico to deal with vessel trajectory reconstruction, abnormal behavior detection and vessel type identification. Preliminary results are reported here for AIS data in January 2014, which amount to 10 154 808 AIS messages.

3.1 Vessel trajectory reconstruction

The trajectory reconstruction layer is a particle filter, estimates the position of vessel where data are missing. We follow [6] and take into account maritime

contextual information to build this filter. Instead of using TREAD [8] to extract maritime routes, the contextual information in our case is here learned by the embedding layer.

We test the trajectory reconstruction by deleting 2-hours segments in vessel tracks, then reconstruct these missing segments. The model is able to perform some surprising good results like those shown in Fig. 2.

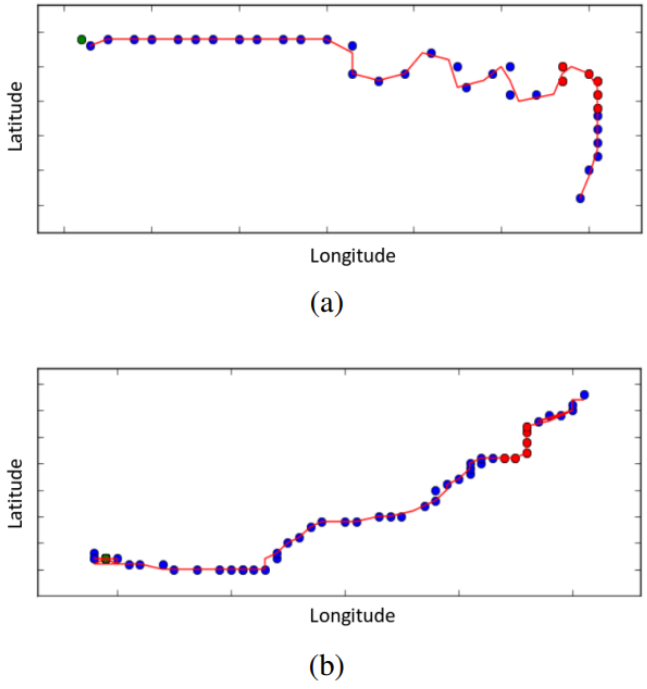


Fig. 2. Two examples of vessel trajectory prediction. Blue dots: received AIS messages; red dots: missing AIS messages; red line: estimated trajectory. The model could predicts these turns because others vessels in this regions did the same.

3.2 Abnormal behaviour detection

This layer addresses the detection of abnormal vessel behaviors at a 2-hour time scale. Our model learns the distribution of vessels' trajectories from the training set, both in terms of geometrical patterns, space-time distribution as well as speed and heading angle features. Any trajectory in the test set that does not suit this distribution will be considered as abnormal. An example of the outcome of the detector is shown in Fig. 3.

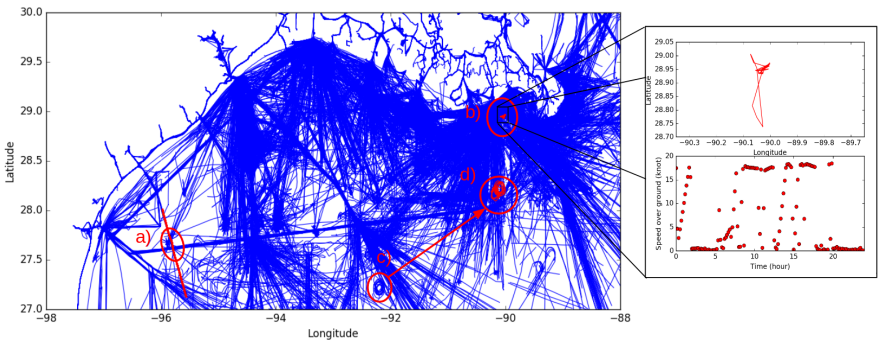


Fig. 3. Three examples of detected abnormal tracks: Tracks in the training set (which itself may contain abnormal tracks) are presented in blue. Abnormal tracks detected in the test set are presented in red; a) this track diverges from the usual maritime route in this area. b) example of abnormal speed pattern, c) example of simulated geometric pattern correctly detected as abnormal (this example was simulated by translating behaviors observed in zone C to zone D.

3.3 Vessel type identification

Using a Convolutional Neural Network (CNN) on top of the RNN, we design a vessel type classifier. This layer operates at a 1-day time scale. The targeted classification task comprises 4 classes of vessel: cargo, passenger, tanker and tug. We reach a relevant f1-score of **88.01%**.

4 Conclusions and perspectives

We introduced a deep learning model that can process the AIS stream on-the-fly for multiple purposes. The use of variational recurrent neural networks provide our model the ability to deal with irregular time sampling and noisy AIS data streams. Three tasks have been tested with successful outcomes. Other tasks (fishing detection, AIS on-off switching detection, etc.) can be added by simply plugging in other task-specific layers on top of the current ones.

Future work could involve benchmarking experiments with current state of the art methods, including the evaluation of the ability of the proposed approaches to scale up to global AIS data streams. The fusion with other sources of information in the maritime domain could be a promising solution.

5 Acknowledgements

This work was supported by public funds (Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche, FEDER, Région Bretagne, Conseil Général du Finistère, Brest Métropole) and by Institut Mines Télécom, received in the framework of the VIGISAT program managed by "Groupement Bretagne Télédétection" (BreTel).

The authors acknowledge the support of DGA (Direction Générale de l'Armement) and ANR (French Agence Nationale de la Recherche) under reference ANR-16-ASTR-0026 (SESAME initiative), the labex Cominlabs, the Brittany Council and the GIS BRETEL (CPER/FEDER framework).

References

1. Ammoun, S., Nashashibi, F.: Real time trajectory prediction for collision risk estimation between vehicles. In: 2009 IEEE 5th International Conference on Intelligent Computer Communication and Processing. pp. 417–422 (Aug 2009). https://doi.org/10.1109/ICCP.2009.5284727
2. Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A., Bengio, Y.: A Recurrent Latent Variable Model for Sequential Data. In: Advances in neural information processing systems. pp. 2980–2988 (Jun 2015)
3. Laxhammar, R.: Anomaly detection for sea surveillance. In: 2008 11th International Conference on Information Fusion. pp. 1–8 (Jun 2008)
4. Lee, J.G., Han, J., Whang, K.Y.: Trajectory Clustering: A Partition-and-group Framework. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. pp. 593–604. SIGMOD '07, ACM, New York, NY, USA (2007). https://doi.org/10.1145/1247480.1247546, http://doi.acm.org/10.1145/1247480.1247546
5. Mascaro, S., Nicholso, A.E., Korb, K.B.: Anomaly detection in vessel tracks using Bayesian networks. International Journal of Approximate Reasoning **55**(1, Part 1), 84–98 (Jan 2014). https://doi.org/10.1016/j.ijar.2013.03.012, http://www.sciencedirect.com/science/article/pii/S0888613X13000728
6. Mazzarella, F., Vespe, M., Damalas, D., Osio, G.: Discovering vessel activities at sea using AIS data: Mapping of fishing footprints. In: 17th International Conference on Information Fusion (FUSION). pp. 1–7 (Jul 2014)
7. Pallotta, G., Horn, S., Braca, P., Bryan, K.: Context-enhanced vessel prediction based on Ornstein-Uhlenbeck processes using historical AIS traffic patterns: Real-world experimental results. In: 17th International Conference on Information Fusion (FUSION). pp. 1–7 (Jul 2014)
8. Pallotta, G., Vespe, M., Bryan, K.: Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. Entropy **15**(6), 2218–2245 (Jun 2013). https://doi.org/10.3390/e15062218
9. Ristic, B., Scala, B.L., Morelande, M., Gordon, N.: Statistical analysis of motion patterns in AIS Data: Anomaly detection and motion prediction. In: 2008 11th International Conference on Information Fusion. pp. 1–7 (Jun 2008)
10. Simsir, U., Ertugrul, S.: Prediction of Position and Course of a Vessel Using Artificial Neural Networks by Utilizing GPS/Radar Data. In: 2007 3rd International Conference on Recent Advances in Space Technologies. pp. 579–584 (Jun 2007). https://doi.org/10.1109/RAST.2007.4284059

Multi-domain assessments in AIS falsification cases

Clément Iphar[1], Cyril Ray[2], Aldo Napoli[3], Pierre-Yves Martin[4] and Alain Bouju[5]

[1] NATO STO CMRE, La Spezia, Italy
[2] French Naval Academy Research Institute, Brest, France
[3] CRC, MINES ParisTech, PSL Research University, Sophia Antipolis, France
[4] CEREMA, Brest, France
[5] L3i, La Rochelle University, La Rochelle, France
clement.iphar@cmre.nato.int
cyril.ray@ecole-navale.fr
aldo.napoli@mines-paristech.fr
pierre-yves.martin@cerema.fr
alain.bouju@univ-lr.fr

Abstract. This paper presents the ins and outs of the DéAIS project, a French national multi-partner project on the topic of maritime data falsification. The objectives and the scientific approach are explained, and the results are presented according to the various fashions that were discriminated: on the risk analysis, the signal, or the data integrity assessment sides.

Keywords: AIS, integrity assessment, signal analysis.

1 Objectives

The maritime environment undergoes an ever-growing activity that favoured the apparition of localisation systems such as the Automatic Identification System (AIS), which enables a real-time surveillance of the maritime traffic and increases safety of the navigation. However, it has been demonstrated that AIS falsification does exist, and could engender new maritime risks and illegal acts [1].

DéAIS project[1] proposes a methodological approach for modelling, analysing and detecting these new maritime risks. The objective is to detect when a ship's AIS system is undergoing an attack. For this purpose, real-time AIS information is analysed and compared to historical, expected or predicted information.

First, an in-depth analysis of AIS was necessary in order to understand its intrinsic characteristics, advantages and drawbacks, as well as its normative and organisational aspects. This analysis highlighted a proven monitoring deficiency on this system as vulnerabilities such as voluntary falsification or disruption cases were demonstrated, as well as inherent drawbacks of the system in its technology and architecture.
Issues have been classified in three families: AIS physical features, radio communication links and information systems using AIS. Those families offer various analysis

[1] http://deais.crc.mines-paristech.fr

possibilities; therefore, several research directions were taken: concentrating on the risk analysis of the system itself, analysing the communication signal and focusing on the data within the messages.

2 Scientific Approach

Based on the fact that the AIS does not carry perfectly genuine data (beyond data errors), that those inaccuracies are not perfect and therefore are detectable, and that impacts on the real-world can be substantial, a set of objectives has been set for detection of falsifications. Relying on the accurate understanding of the way the system is supposed to work, of its vulnerabilities and the errors and falsification that have been highlighted, these objectives include the creation of an attacking platform allowing the creation and the broadcast of falsified data, the modelling of a statistic and algorithm-based falsification detection mechanism, the creation of an information system for the real-time handling of data taking into account archived or forecasted data, and the modelling of risks that are inducted by an inadequate use of AIS, as well as an assessment of the risks linked to AIS errors, anomalies, falsification or spoofing.

3 Results

An EBIOS risk analysis of the AIS has been performed [2], consisting in the analysis of vulnerabilities, failures and risks associated with it, enabling the identification of issues that could actually emanate from the use of AIS. This method has been chosen for its compliance to ISO norms and a list of circa 350 risks has been established.

Real data were collected by antennas set up near the Brest harbour and in the La Rochelle area. Studies on the signal side were performed, in particular involving transceiver signature [3] (physical features characteristic to a type of transceiver) and coverage maps [4] (as it has been demonstrated that the reception coverage is not uniform in the surroundings of the antenna, due for instance to the presence of masks). The presence of unexpected vessels in specified areas has been addressed.

Figure 1. *Two computers, Gnuradio, Balduzzi (improved), SDRsharp, AISMon, openCPN, Matlab, SAAB R4, SAAB R5 secured, HackRF, VHF USB dongle, and additional advanced SDR equipment for lab experiments. Also includes a mobile platform based on a laptop containing all falsifcation and detection features allowing for external experiments.*

In order to process falsification and detection safely, a laboratory emission-reception platform has been built, enabling the realisation of made-up AIS attacks (Figure 1). In addition, the development of an AIS message construction tool allowed us to play a large set of scenarios, without relying on real data.

The detection software relies on a database combining parsed AIS messages from the Brest antenna, geographic, meteorological, sea state, vessel and port data, all aligned in both time and space [5] that has been created in order to enhance the analysis of data. A methodology of analysis of AIS messages based on integrity of data on several layers (within a single message, between different messages, on temporal series, with environmental data) has been established, with the purpose of assessing the truthfulness of the transmission [6]. The analysis system and the associated algorithms have been applied to scenarios on identity and position issues, message saturation, and on report frequency adjustment messages for tests on real and falsified data.

Acknowledgement

This research has been supported by ANR and DGA under grant ANR-14-CE28-0028.

References

1. Balduzzi M., A. Pasta, and K. Wilhoit. A security evaluation of ais automated identification system. In Proceedings of the 30th Annual Computer Security Applications Conference, pages 436–445. ACM, 2014.
2. Iphar C., A. Napoli and C. Ray (2016) Risk Analysis of falsified Automatic Identification System for the improvement of maritime traffic safety. In: Proceedings of the 2016 ESREL conference, Glasgow, UK
3. Alincourt E., C. Ray, P.-M. Ricordel, D. Dare-Emzivat and A. Boudraa (2016) *Methodology for AIS Signature Identification through Magnitude and Temporal Characterization*. In: Proceedings of the 2016 OCEANS conference, Shanghai, China
4. Salmon L., C. Ray and C. Claramunt (2016) *Continuous detection of black holes for moving objects at sea*. In: Proceedings of ACM SIGSPATIAL IWGS, San Francisco, USA
5. Ray, C., R. Dréo, E. Camossi and A.-L. Jousselme (2018) *Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance* 10.5281/zenodo.1167595
6. Iphar, C. (2017). *Formalisation of a data analysis environment based on anomaly detection for risk assessment – Application to Maritime Domain Awareness*. PhD Thesis, MINES ParisTech, PSL Research University

CMRE-CP-2018-002

9 Acknowledgements

The workshop has been sponsored by the Research and Innovation collaborative project datAcron (Big Data Analytics for Time Critical Mobility Forecasting), whose aim is to advance Big Data technologies in order to increase the capacities of systems to analyse large numbers of moving entities in large geographical areas by developing novel methods for real-time entities tracking and event forecasting. The datAcron project is funded by the European Union's Horizon 2020 Programme under grant agreement No. 687591.

Document Data Sheet

Security Classification		Project No.
RELEASABLE TO THE PUBLIC		DKOE
Document Serial No.	**Date of Issue**	**Total Pages**
CMRE-CP-2018-002	January 2019	90 pp.

Author(s)

Elena Camossi, Anne-Laure Jousselme

Title

Proceedings of the Maritime Big Data Workshop

Abstract

The NATO STO Centre for Maritime Research and Experimentation, as part of its mission to put forward the technological maritime research, with the support of the European Union's Horizon 2020 Programme has organized on May 9-10, 2018, the Maritime Big Data Workshop (MBDW). The workshop gathered together researchers, technological providers and members of the operational community to exchange their experience on Big Data innovations for maritime security, safety and security of maritime navigation and transport, sustainable fisheries and exploitation of ocean resources. For two days, 37 researchers and experts from Brazil, Canada, France, Germany, Greece, Italy, Portugal, South Africa, and Vietnam presented their work and main findings on Maritime and Big Data, including the outcomes of 6 ongoing Maritime Big Data projects and initiatives funded by the European Union: datAcron, MARISA, Ranger, EUCISE, AtlantOS, EMODnet.

The workshop's results enable to draw some preliminary conclusions on the current research and developments in Maritime Big Data. There is a general interest towards concrete societal and operational needs, coupled with an emerging tendency to develop methods combining heterogeneous, potentially complementary, information streams (mainly AIS, paired with SAR, Radar, METOC, acoustic), with an increasing attention towards source quality. The approaches adopted come from different areas of research, mainly machine learning and data mining, incorporating also techniques developed in Information and data fusion, but also data warehouse and online analytical processing.

The current trend towards experimenting open source Big Data technologies is challenged by the integration of diversified sources of information, which comes with an increased exigence of enhanced data management capabilities for harmonised data sharing and processing that can overcome the sole exploitation of kinematic data. Meanwhile, there is a prevailing requirement to reduce the uncertainty of detection and prediction results, entailing the development of capabilities to formally handle information and source quality. Analogously, the emergence of novel Artificial Intelligence approaches that, despite showing promising results, challenge results' interpretation, requires an increased involvement of experts in all the phases of the development (the so-called ``Human in the loop"), and the holistic incorporation of approaches addressing human factors' aspects.

Keywords

Maritime Big Data, , Maritime sensors networks, Maritime Intelligent Surveillance and Reconnaissance, Maritime Situational Awareness, Maritime Interoperability, Maritime Information Fusion, Maritime Cyber Security, Human factors, Maritime Open Data, Efficiency of Navigation, Sustainable fisheries

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