Grant Agreement No: 687591

15/07/2018

Big Data Analytics for Time Critical Mobility Forecasting

datAcron

Maritime Data Preparation and Curation (final)

Deliverable Form	
Project Reference No.	H2020-ICT-2015 687591
Deliverable No.	5.4
Relevant Work Package:	WP 5
Nature:	R (report)
Dissemination Level:	Public
Document version:	1.3
Due Date:	15/07/2018
Date of latest revision:	13/07/2018
Completion Date:	13/07/2018
Lead partner:	NARI
Authors:	Cyril Ray, Clément Iphar, Richard Dréo, Waldo Kleynhans, Elena Camossi, Anne-Laure Jousselme, Maximilian Zocholl, Ernie Batty, Quentin Roche, Arnaud Metzger
Reviewers:	Christos Doulkeridis, Manolis Pitsikalis, Alexander Artikis, George Vouros
Document description:	This deliverable provides a description of maritime data and degration techniques prepared and designed for the project.
Document location:	Documents/datAcron/WP5/Deliverables/Final

HISTORY OF CHANGES

Version	Date	Changes	Author	Remarks
1.0	24/05/2018		Cyril Ray	Initial version
1.1	22/06/2018		Clément Iphar, Cyril Ray, richard Dréo, Waldo Kleynhans	Draft version released for comments
1.2	05/07/2018		Cyril Ray, Clément Iphar, Waldo Kleynhans	Draft Final version
1.3	13/07/2018		Elena Camossi, Anne-Laure Jousselme, Maximilian Zocholl, Cyril Ray	Final version

EXECUTIVE SUMMARY

The second task (WP5.2) of the maritime use case work package aims to deliver a definition of the datasets to be used in the research and for the evaluation and validation purposes. Specifically, it describes multiple, heterogeneous datasets that may be used for maritime scenarios and use case validation.

This datAcron deliverable entitled *Maritime Data Preparation and Curation* has been separated in two parts. The interim report was dedicated to the identification of maritime data sources and to the preparation of preliminary data samples [12]. The objective of this first report was to list and describe available (public) data sources relevant to the maritime scenarios and the project objectives, as well as to provide the initial details about their nature and accessibility. It was also mentioned how they support the different maritime scenarios described in deliverable D5.1.

This second deliverable of task 5.2 describes historical data and the live stream prepared for the maritime use case. It is organized in six parts. First, the introduction explains how maritime data preparation has been accomplished to highlight the four Vs of Big Data (Volume, Velocity, Variety and Veracity). Section 2 details historical data prepared to build a reference heteregeneous dataset. Section 3 presents the AIS stream that feed continuously the datAcron architecture. Since assessing results of algorithms under uncertainty is very challenging, controlled data and parameter degradation is used to drive the evaluation. Section 4 presents the approach developed for controlled data degradation. Section 5 presents patterns computed on raw data specifically to enrich the reference database for the evaluation of datAcron algorithms.

TABLE OF CONTENTS

Contents

LIST OF FIGURES

LIST OF TABLES

1 Introduction

Designing and evaluating a big data architecture, related algorithms for processing and analysing data as well as visualisation outputs requires careful data collection, preparation and curation. The objective of the work presented in this deliverable is multiple. First, it provides matters that drive the design of the architecture and the algorithms of the four technical work packages. Secondly, it provides a large variety of real data that include intrinsic veracity issues at the appropriate volume and velocity level in order to run the components of the big data system in real conditions. Working with real data is essential for the credibility of processing and results. However, an accurate assessment of the results can only be experimented on controlled datasets. Lastly, preparation of a reference dataset is also a mandatory work.

During the first year of the project, a comprehensive study of maritime data and an inventory of available public data sources has been realised. This has led to the identification of over forty data sources which have been categorized in three domains (vessel information, contextual information, environmental data) and detailed within sixteen categories. Based on this study, an initial list of representative datasets was delivered to different partners.

Reaching appropriate Maritime Situation Awareness (MSA) for the decision maker requires processing in real-time a high volume of information of different nature (numerical, natural language statements, objective or subjective assessments,...), originating from a variety of sources (sensors and humans - hard and soft), with a lack of veracity (uncertain, imprecise, vague, ambiguous, incomplete, conflicting, incorrect, etc), and coming with high velocity. Based on this postulate, we designed and prepared an adapted maritime dataset. This maritime dataset is composed of two parts:

- A reference (batch) dataset, limited in volume but composed of a large variety of maritime data;
- A real-time stream of AIS messages designed to stress the datAcron architecture under a large volume of surveillance data with high velocity.

Having spatially and temporally aligned maritime dataset including not only ships' positions but also a variety of complementary data sources is of great interest for the understanding of maritime activities and their impact on the environment and towards attaining maritime safety and security. With this objective in mind, different types of data have been prepared and collected within the reference historical dataset. As for core positioning data, work package partners have collected and prepared the data themselves. For all complementary data we favoured interactions with European institutions and projects like SeaDataNet[1], Copernicus[2], EMODnet[3] or Ifremer[4] and the European Commission Joint Research Centre[5]. The section 2 presents the variety of data prepared for the reference dataset.

Data obtained by various types of sensor technologies have to be processed, cleaned up from inconsistencies, transformed, harmonised in format, and aggregated, sometimes simplified. The growing number of sensors (in coastal and satellite networks) makes indeed the sea one of the most challenging environments to be effectively monitored; the need for methods for the data processing of vessel motion data at sea, which are scalable in time and space, is highly critical for maritime security and safety. In particular, the analysis of streaming data from multiple sensors is essential to detect critical events as soon as they occur at sea. The section 3 presents the maritime data stream.

Data measurements have an intrinsic uncertainty, which may be addressed by proper fusion algorithms and clustering in the preparation/preprocessing phase by assessing the quality of

[1] https://www.seadatanet.org
[2] http://marine.copernicus.eu
[3] http://www.emodnet.eu
[4] https://www.ifremer.fr
[5] https://bluehub.jrc.ec.europa.eu/research_areas_maritime

data themselves and by combining measurements from complementary sources. The sources themselves lack the quality; they may be unreliable, incompetent, badly intentioned, imprecise, uncertain, etc. This leads to provide operational experts with information that suffer from equivalent drawbacks, and it can thus be conflicting for the achievement his/her mission. For instance, AIS data are incomplete, intermittent, with errors, and the signal can be spoofed. The assessment of data quality is a very challenging task. Usually, the quality of data is not provided (e.g. meta data) and can only be estimated based on available statistics. The Section 4 will present the work done to assess the quality of the available data and deriving datasets with controlled degradation.

Finally, in order to perform experiments and assess the results of datAcron algorithms, the processing of raw data to extract, create different maritime patterns and clusters is required. Amongst this, maritime routes appear to be one of the most useful features to assess trajectories and movement prediction computed on synopses. The section 5 presents the maritime route and stationary areas processed for this purpose.

2 Reference Dataset

Facing an increasing amount of traffic at sea and its impact on the global ecosystem, many research centers, international organizations, industries have favored and developed sensors and techniques for monitoring, analysis and visualization of sea movements. **Automatic Identification System** (AIS) is an electronic system that enable ships to broadcast their dynamic (position, speed, destination, . . .) and static (name, type, international identifier, . . .) information via radio communications. For the datAcron maritime use case, AIS data have been identified as the main data source for generating the maritime moving objects trajectories (synopses) that will be analysed by the algorithms developed by WP2, WP3 and WP4 and visualised by WP4. Therefore, the preparation of the datAcron reference dataset (as well as the data stream presented in Section 3) has been based on data collected from this positioning system.

Ship positions obtained from AIS are essentials for datAcron algorithms, however they might not be always sufficient to solve maritime surveillance missions as those described in the maritime use case scenarios defined in deliverable D5.1. Only if properly combined and integrated with other data acquired from other information sources, positional data such as AIS can support the datAcron maritime use cases. Table 1 summarises the data prepared for this purpose, which is described in detail in the rest of the section.

Type	Source	Provenance	Format	Spatial extent	Temporal extent	Volume	Velocity
Surveillance	Automatic Identification System	Naval Academy	Flat files (comma-separated values)	Western part of the Channel and France	Oct. 2015 to March 2016 (6 months)	19.680.743 messages (1.05 GB)	∼76 messages per min (in average)
	Automatic Identification System	IMISG	Flat files (text)	Europe	January 2016 (1 month)	80.169.806 messages (8.5 GB)	∼1.830 messages per min (in average)
	Automatic Identification System	IMISG	Stream of decoded messages in JSON	Europe	-	∼400 KB / min (in average)	∼1.830 messages per min (in average)
Weather	Sea state	SHOM, IFREMER	Flat files (comma-separated values)	Western part of the channel and France	Oct. 2015 to March 2016 (6 months)	79.652.684 forecasts (3.02 GB)	1463 forecast files, 1 file / 3 hours
	Weather forecast	Met Office (United Kingdom) based on data provided by NOAA (United States)	Flat files (comma-separated values)	Western part of the channel and France	Oct. 2015 to March 2016 (6 months)	71.516 observations (5 MB)	1 observation / hour, from 16 weather stations
Contextual	Geographical	Various, including nautical charts	ESRI shapefiles	Europe	-	22 different features (1.4 GB)	-
	Port Registers	World Port Index, SeaDataNet	ESRI shapefiles	World	-	5754 different ports (70 MB)	-
	Vessel Registers	European Commission, Agence National des Frequences	Flat files (comma-separated values)	-	-	166.683 distinct ships	-

Table 1: Main maritime data prepared

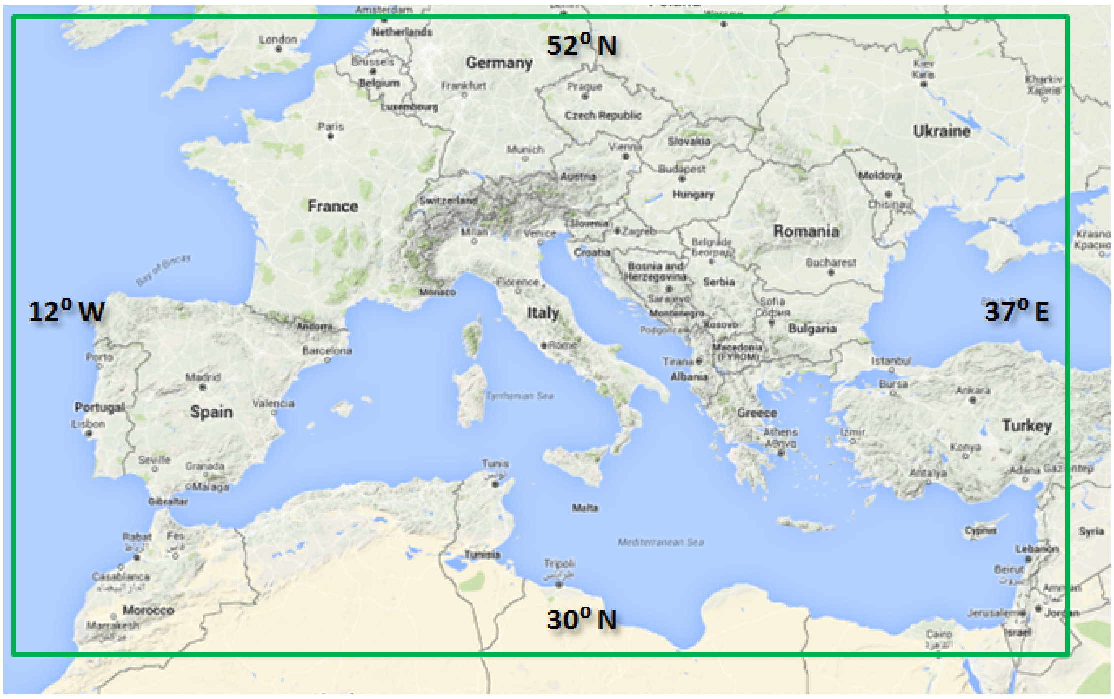

Figure 1: Area of Interest

2.1 Volume & Velocity - based dataset

IMIS has provided one month of AIS messages for the region depicted in Figure 1, which covers most of the European coasts and has been identified as the area of interest for the datAcron maritime use case. It is used to compute synopses and events at a large scale with a reference volume-based dataset.

The AIS data provided by IMIS for the datAcron project has been sourced from a range of terrestrial and satellite AIS sources. The terrestrial sources are collated from various sources in Europe and decimated to limit the amount of data. The satellite data is obtained from a range of ORBCOMM satellites of various generations and include an AIS receiver on the International Space Station (ISS), a range of older generation satellites and 11 new generation satellites that go to make up 19 sources of satellite data.

2.1.1 AIS data format

The AIS data is provided in its raw and unparsed format as received from the AIS data sources, and can contain any of the 27 different message types as described in the ITU-R.M 1371-4 or NMEA 4.0 specification. The data includes a comment or TAG block which provides additional information for the IEC 61162-1 message. An example of this data is found below:

Single-line Message:

```
\s:SIMULATOR,c:1424419673,T:2015-02-20 08.07.53,
e:10000010010000000000i:|X=0|D=1|T=42055.3388029977|P=192.168.2.145:10000|R=IN|*hh
\!AIVDM,1,1,,,18157Rh00=0pPJ1svJ46T5Hf0L08,0*47<CR><LF>
```

The comment or TAG block includes the following parameters:

- **s:** The source of the message.

- **c:** The unix timestamp of the message when received (seconds since midnight, January 1st, 1970)
- **T:** The human readable timestamp of the message when received in yyyy-mm-dd hh.nn.ss
- **e:** The message error flag. Bits in this identifier are set to "1" if any of the 15 errors described in Annex A are true. If no error is present in the data, is this field excluded from the message output
- **i:** Proprietary data and contains the following fields, separated by a "|" character:

  X=Data source RX / TX capability = always set to "0"
  D=Data source "delayed data flag" = always set to "1"
  T=Proprietary timestamp of the message
  P=The IP address and port where the message was received by the MSA
  R=The direction of the message

The IEC 61162-1 sentence is described below:

- **!AIVDM:** VDM Message identifier
- **TotalSentences:** Total number of sentences.
- **SentenceNumber:** Sentence number of this sentence
- **SeqMsgNum:** Sequential message ID. Always empty for a single line message
- **AISChannel:** Channel ID (A, B, C or D or empty if unknown)
- **EncapsulatedMsg:** Bits of the data portion of the AIS message type 1, 2,3, 9,18, 19, 21, 24 or 27 (Each character represent 6 bits - encoding is per NMEA0183). The number of bits (and characters) depends upon the message type.
- **FillBitsNumber:** Number of Fill bits appended
- **Chksum:** Checksum ("12" is the checksum)
- **<CR><LF>** Carriage Return and Line Feed

Multi-line Message:

Line 1:

\g:1-2-1234,s:SIMULATOR,c:1424419673,T:2015-02-20 08.07.53,e:10000010010000000000
i:|X=0|D=1|T=42055.3388029977|P=192.168.2.145:10000|R=IN|*hh
!AIVDM,2,1,8,,5P000Oh1IT0svTP2r:43grwb05q41P000Oh1IT0svTP2r:43grwb05q41P00,0*15<CR><

Line 2:

\g:2-2-1234*2hh\!AIVDM,2,2,8,,0Oh1IT0svT,0*7b<CR><LF>

The comment or TAG block includes the following parameters:

- **g:** Identify line 1 out of 2 lines of group 1234 (for example: g:1-2-1234)
- **s:** The source of the message

- **c:** The unix timestamp of the message when received (seconds since midnight, January 1st, 1970)
- **T** :The human readable timestamp of the message when received in yyyy-mm-dd hh.nn.ss
- **e:** The message error flag. Bits in this identifier are set to "1" if any of the 15 errors described in Annex A are true. If no error is present in the data, is this field excluded from the message output
- **i** :Proprietary data and contains the following fields, separated by a "|" character:

  X= Data source RX / TX capability = always set to "0"
  D= Data source "delayed data flag" = always set to "1"
  T= Proprietary timestamp of the message
  P= The IP address and port where the message was received by the MSA
  R= The direction of the message

The IEC 61162-1 sentences are described below:

line 1

- **AIVDM:** VDM Message Identifier
- **TotalSentences:** Number of sentences.
- **SentenceNumber:** Sentence number of this sentence
- **SeqMsgNum:** Sequential Message ID (0-9)
- **AISChannel:** AIS Channel (A, B, C, D or empty)
- **EncapsulatedData:** First part of the "data" section of the AIS Message
- **FillBits:** Number of fill bits
- **Chksum:** checksum

line 2

- **AIVDM:** VDM Message identifier
- **TotalSentences:** Number of sentences
- **SentenceNumber:** Sentence number of this sentence
- **SeqMsgNum:** Sequential Message ID (0-9)
- **AISChannel:** AIS Channel (A, B, C, D or empty)
- **EncapsulatedData:** Last part "data" section of the AIS Message (per ITU M1371-3)
- **FillBits:** Number of fill bits
- **ChkSum:** Checksum

2.1.2 Dataset parameters

The dataset were provided as a flat file with each AIS message formatted according to the specification given in section 2.1.1 and is defined as:

- **Area of interest:** North West Coordinates are (52 Degrees North, 12 Degrees West) and South East Coordinates are (30 Degrees North, 37 Degrees East). Shown in Figure 1.
- **Start time:** 2016-01-01 00:00:00 UTC
- **End time:** 2016-01-31 23:59:59 UTC
- **ITU Message types selected:** ITU123, ITU5, ITU9, ITU18, ITU19, ITU21, ITU24
- **Total number of AIS messages:** 80169806
- **Total File size:** 8.5G

2.2 Variety & Veracity -based dataset

For experimentations and validation of algorithms, it is essential to define testing areas before applying them at European scale. A testing area is a region of limited coverage where an accurate knowledge of ships' movement and ground truth support is possible for evaluating algorithms accuracy. For such an area, both terrestrial and satellite AIS data are required. It should contain active fishing areas, together with additional contextual information (cartography, regulated areas, known fishing fleet...). Knowledge on local fisheries and connections with operational entities (e.g., control center, local committee of fishery, port authorities, navy) are also necessary to establish ground truth. The western part of France around Brest city has been identified as a test area that fulfills most of the aforementioned requirements. Moreover, the Brest bay itself has local regulations which enforce fishing vessels to use AIS permanently providing a rich-set of high-quality positional information for algorithms' testing and validation.

This *Variety-based* dataset is based on ships' information collected, parsed and prepared by NARI. It contains messages received by a local AIS receiver, integrated with a set of complementary data having spatial and temporal dimensions aligned. The dataset contains four categories of data:

- Navigation data
- Vessel-oriented data
- Geographic data
- Environmental data

It covers a time span of six months, from October 1st, 2015 to March 31st, 2016 and provides ships positions within Celtic sea, the Channel and Bay of Biscay (France). The bounding box has the following coordinates: Longitude between -10.00 and 0.00 and Latitude between 45.00 and 51.00. Figure 2 illustrates that bounding box and position reports.

As the *Volume-based* dataset covers European coasts during January 2016, it is also a source of redundancy for navigation data. The large-scale dataset, presented in the previous section, has been therefore parsed, cropped to the aforementioned region and added as a secondary source to this small-scale dataset.

The dataset has been processed and queried using a relational database management system. This relies on the widespread and open source relational database management system PostgreSQL, with the adjunction of the PostGIS extension, for the treatment of all spatial features

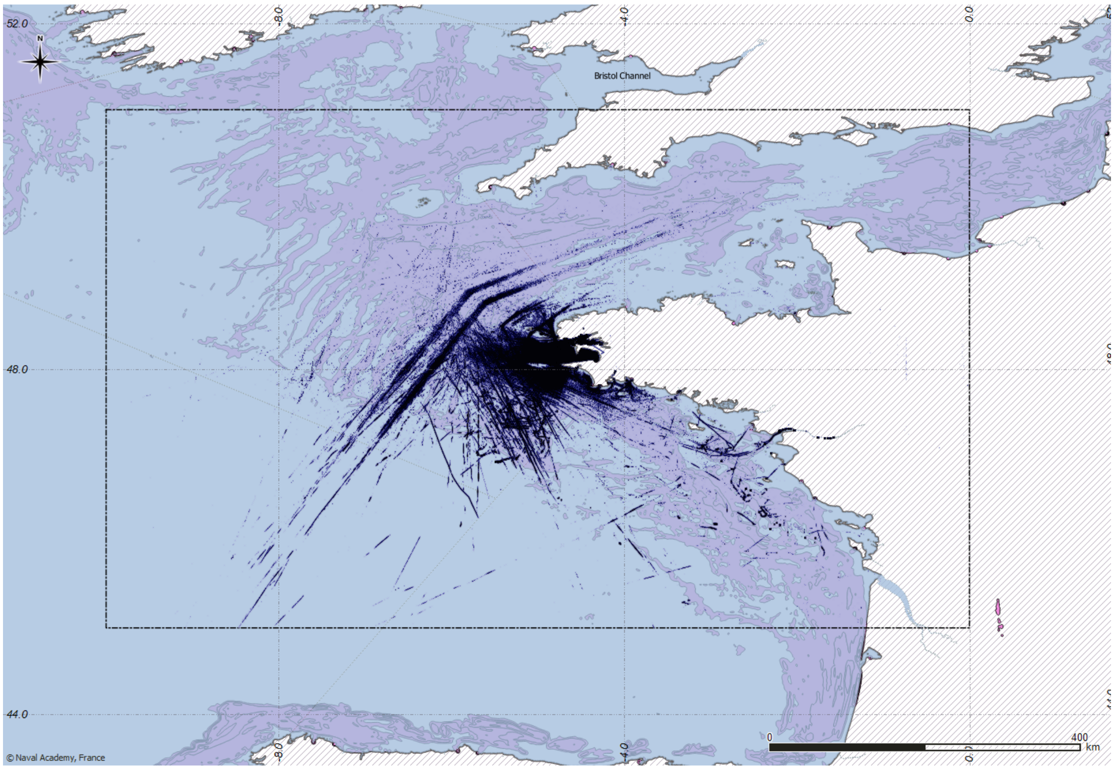

Figure 2: Variety-based dataset, area of interest

proposed in the dataset. All data have been exported and shared in the format of comma-separated value files (CSV, with .csv file extension) or ESRI Shapefile (with .shp file extension) for geographic features. Each data is provided with detailed meta information in the datAcron filestore giving: source, licence, spatial reference identifier (SRID), coverage, temporal range, volume, and a clear description of each data field including respective data type.

2.2.1 Navigation Data

Raw AIS messages are differentiated in 27 types that have been subdivided into two main classes: dynamic and static messages.

- **AIS dynamic messages:** Dynamic messages provide information on position, speed, heading, course over ground, rate of turn. Here, the following International Telecommunication Union (ITU) Message types were selected: ITU 1, ITU 2, ITU 3, ITU 18, ITU 19. The information extracted from these messages contains: the ship identifier (MMSI), the coordinates of the vessel and the associated speed, its heading, the course over ground. Dynamic data also include locations of aids to navigation (ATON) provided by message ITU 21 and search and rescue (SAR) provided by message ITU 9. [*Prepared by NARI*]

- **AIS static messages:** These messages provide ship meta-information such as ship identifiers (MMSI and IMO number), name, type, and dimension of the vessel, its destination, its estimated time of arrival (ETA), its draught (i.e., the vertical distance between the waterline and the bottom of the hull, namely the keel, with the thickness of the hull included). The following ITU message types were selected: ITU 5, ITU 19, ITU 24. [*Prepared by NARI*]

- **AIS status, codes and types:** Some data fields of AIS messages are encoded, usually with integer codes. For a clear understanding and analysis of AIS data, the encoding enumerations with associated information are required. Several files explaining this equivalences have been integrated in the dataset:

- Status: The navigational status (e.g., moored, under way) provided in the dynamic messages is coded by an integer. The corresponding types have been detailed in a status file (in CSV format). [*Prepared by NARI*]
- Country Codes: The list of the country codes corresponding to the first three digits of each MMSI number (e.g., 227 is France) is detailed in a CSV file. [*Prepared by NARI*]
- Types: The ship type is encoded in the static messages by an integer (e.g., 30 for fishing vessels). The corresponding list of types is provided in CSV format. Additionally, a list of 233 ship types has been provided for advanced classification. [*Prepared by NARI*, the extended list is based on [6]]
- ATON: The type of aid to navigation is encoded in the dynamic messages by an integer. The corresponding list of types is provided in a CSV file. [*Prepared by NARI*]

- **Receptor location:** The position of the terrestrial receiver used for NARI AIS data recording, in shapefile format. [*Prepared by NARI*]

- **Theoretical coverage of the receptor:** The theoretical coverage of the AIS receiver, represented as a polygon in shapefile format. Computation has been done with the definition of sea areas from IMO resolution A801(19). It computes a circle of radius R nautical miles where R is equal to the transmission distance between a ship's VHF antenna at a height of 4 meters above sea level and the antenna of the VHF coast station (at a height of H meters) which lies at the centre of the circle. [*Prepared by NARI*]

2.2.2 Vessel-oriented data

- **Fishing Vessels Fleet Register:** The European Union provides a list of fishing vessels flying Member State flags. Many administrative and technical information are included (in CSV). [*Provided by European Commission - Fisheries and Maritime Affairs*]

- **French Fleet Register:** This is the fleet register of the French Frequencies Agency, which includes a great number of French-registered vessels (in CSV). [Provided by French Frequencies Agency (ANFR) sourced by data.gouv.fr]

- **Local Shellfish Vessels:** A list of 63 known shellfish vessels with license in Brest area (in CSV). [*Provided by departmental fishery committee of Finist'ere*]

- **ICCAT Blacklist vessels:** A list of vessels presumed to have carried out illegal, unreported, and unregulated (IUU) fishing activities in the ICCAT convention area and other areas (in CSV). [*Provided by the International Commission for the Conservation of Atlantic Tunas*]

2.2.3 Geographic data

- **Brest port**: A polygon representing the Brest port (in shapefile). It had been prepared fusing S57[7] objects, specifically CTNARE (An area where the mariner has to be aware of circumstances influencing the safety of navigation) and DRYDOC (artificial basin fitted with a gate or caisson). [*Extracted from IHO S57 Nautical Chart*]

[6] https://help.marinetraffic.com/hc/en-us/articles/205579997-What-is-the-significance-of-the-AIS-Shiptype-number-

[7] IHO S-57 (and its revision S-100) is the current IHO standard for digital hydrographic data. It defines a data model and a data structure and format used to implement it. It includes 159 geo-object classes.

- **Ports of Brittany**: Names and coordinates of 222 ports of Brittany (in shapefile). [*Provided by Brittany region, sourced by data.gouv.fr*]
- **SeaDataNet port index:** Names and coordinates of about 5000 halieutic ports throughout the world (shapefile). [*Provided by Seadatanet infrastructure for the management marine data*]
- **World port index:** Names and coordinates of about 3700 major ports and terminal throughout the world (in shapefile). [*Provided by the National Geospatial-Intelligence Agency*]
- **Aircraft stations:** 14 aircraft stations extracted from S57 AIRARE (airport/airfield) objects (as shapefile). An aircraft station is an area containing at least one runway, used for landing, take-off, and movement of aircrafts; This can be the origin of Search And Rescue (SAR) operations. [*Extracted from IHO S57 Nautical Chart*]
- **Anchorage area:** These are areas in which vessels are anchored or may anchor in Brest bay. These regions have been extracted from S57 ACHARE objects and provided as shapefiles. [*Extracted from IHO S57 Nautical Chart*]
- **Brest bay:** Two polygons defining the Brest Bay as shapefile (the extended bay polygon also includes entrance area). [*Provided by Institut universitaire europÃl'en de la mer*]
- **Dumping Ground:** The shapefile contains S57 DMPGRD objects (dumping ground). A dumping ground is a sea area where dredged material or other potentially more harmful material (*e.g.* explosives, chemical waste) is deliberately deposited (Derived from IHO Chart Specifications, M-4). [*Extracted from IHO S57 Nautical Chart*]
- **Europe coastline:** High resolution European coastline (polylines and polygons, as shapefile). The criteria for defining the coastline is the line separating water from land. [*Created by the European Environmental Agency (EEA) for highly detailed analysis*[8]*. The EEA coastline is a product derived from two sources: EU-Hydro and GSHHG*]
- **European Maritime boundaries**: The shapefile contains maritime boundaries that include territorial waters, bi- or multi-lateral boundaries as well as contiguous and exclusive economic zones. [*Provided by European Environment Agency*]
- **Fairway:** The shapefile contains S57 FAIRWY objects (Fairway). A fairway is a part of a river, harbour and so on, where the main navigable channel for vessels of larger size lies. It is also the usual course followed by vessels entering or leaving harbours, called ship channel (International Maritime Dictionary, 2nd Ed.). [*Extracted from IHO S57 Nautical Chart*]
- **IHO World Seas:** The shapefile contains names and polygons of world seas. [*Provided by VLIZ, Flanders Marine Institute sourced by MarineRegions.org*]
- **Inshore Traffic Zone:** The shapefile contains S57 ISTZNE objects (Inshore Traffic Zone). It is a routing measure comprising a designated area between the landward boundary of a traffic separation scheme and the adjacent coast, to be used in accordance with the provisions of the International Regulations for Preventing Collisions at Sea (IHO Dictionary, S-32, 5th Edition, 2457). [*Extracted from IHO S57 Nautical Chart*]
- **Recommended Track:** The shapefile contains S57 RECTRC objects (Recommended Track). A track recommended to all or only certain vessels. [*Extracted from IHO S57 Nautical Chart*]
- **Surveillance Centers:** The shapefile contains 59 coastguard stations (extracted from S57 CGUSTA objects). Surveillance Centers are stations at which a watch is kept either continuously, or at certain times only (IHO Chart Specifications, M-4). [*Extracted from IHO S57 Nautical Chart*]

[8] https://www.eea.europa.eu/data-and-maps/data/eea-coastline-for-analysis-1

- **Ushant TSS:** The shapefile contains the Ushant traffic separation zone (obtained from S57 TSEZNE object). A traffic separation zone is a zone separating the lanes in which ships are proceeding in opposite or nearly opposite directions; or separating traffic lanes designated for particular classes of ships proceeding in the same direction (IMO Ships Routing, 6th Edition). [*Extracted from IHO S57 Nautical Chart*]

- **Western Rescue Stations:** The shapefile contains S57 RSCSTA objects (rescue stations). A rescue station is a place where lifesaving equipment is held. [*Extracted from IHO S57 Nautical Chart*]

- **Exclusive Economic Zones:** Two shapefiles contain Exclusive Economic Zones Boundaries (polygons and polylines). Areas beyond this boundary can be classified as *high seas.* [*Provided by VLIZ, Flanders Marine Institute sourced by MarineRegions.org*]

2.2.4 Environment-related data

- **FAO fishing areas:** The shapefile contains fishing areas estimation provided by the Food and Agriculture Organization of the United Nation. [*Provided by Food and Agriculture Organization*]

- **Fishing constraints:** The shapefile contains two geographic areas where shellfish fishing activity is forbidden in the time window of the dataset. [*Prepared by NARI*]

- **Natura 2000 areas:** Contains a collection of maritime Natura 2000 areas[9] for all the countries of the European Union. Those areas are protected from an environmental point of view and human activities inside the area or in its neighbourhood must be assessed. [*Provided by European Commission, Unit Nature and Biodiversity, DG Environment*]

- **Fishing locations:** The shapefile contains usual fishing grounds at European scale computed from data collected between September 2014 and September 2015. It can be used to assess if a vessel displaying a fishing behaviour is doing it inside an usual fishing ground or not [9, 18]. [*Provided by Joint Research Centre, European Commission*]

- **Ocean conditions:** Several csv files (one per month) containing sea state forecast based on WAVEWATCH III model [1]. [*Provided by SHOM and IFREMER*]

- **Weather conditions:** The csv file contains coastal weather observations from 16 stations located around Brittany over the 6 months period [*Provided by Met Office (United Kingdom) based on data provided by NOAA (United States) sourced by rp5.ua*]

Some of the data collected and prepared comes with utilisation constraints. Therefore, they can be exploited only for research purpose, with a non-profit usage and are limited to the datacron project (including derivative data). This concerns IMISG navigation data (Section 2.1), S57 Brest port, S57 aircraft stations, S57 anchorage areas, Brest bay, S57 dumping grounds, S57 fairway, S57 inshore traffic zone, S57 recommended track, S57 surveillance centers, S57 Ushant TSS, S57 western rescue stations, local shellfish vessels. They are not part of the public dataset described in the following section.

2.3 Public Release

During this work it became apparent that many maritime data are now available publicly [6]. However, while efforts have been initiated to centralize the access to maritime information, most of the data have heterogeneous types and formats and are still independently sourced and

[9] https://www.eea.europa.eu/data-and-maps/data/natura-9

maintained. These data however can provide useful information and knowledge in support to maritime situational awareness as far as they are properly combined, possibly cleaned up from inconsistencies, converted into standard formats to be harmonized, summarized and integrated.

Having spatially and temporally aligned maritime datasets relying not only on ships' positions but also on a variety of complementary data sources is therefore of great interest for the understanding of maritime activities and their impact on the environment. The selection, gathering and preparation of maritime datasets, in respect to this objective, while relatively straightforward is a time-demanding task which is quite challenging.

Considering licenses of each piece of data in the dataset, we derived a public version of the reference dataset where some geographic features[10] has been removed and published online [13]. This dataset contains ships' information collected by NARI through the Automatic Identification System, prepared together with a subset of related data having spatial and temporal dimensions aligned. The public dataset also contains four categories of data: Navigation data, vessel-oriented data, geographic data, and environmental data. It covers the same time span of six months, from October 1st, 2015 to March 31st, 2016 and provides ships position within Celtic sea, the Channel and Bay of Biscay (France).

For this dataset, an additional effort has been made to revise the reference dataset with updated and more accurate version of existing data. Few data have been pre-processed in a different manner (e.g. files merged) in order to facilitate their usage by non experts and metadata describing each piece of data has been created. These metadata contain, for each data subset, a data description, the originating source, the date the data have been exported, their version, their spatial coverage, the spatial reference system used, their size (or volume), the temporal period they cover, and the associated license. Finally, in order to support users in fast integration and exploitation of the dataset, a database model and all SQL queries and scripts for the integration of the data in a database has been included in the the dataset. This relies on the widespread opensource relational database management system PostgreSQL, with the adjunction of the PostGIS extension, enabling the treatment of all spatial features proposed in the dataset.

This public version of the dataset is already in use for teaching at University of Piraeus, NCSR Demokritos and Naval Academy. It has been shared with several researchers, national and European projects (including the MARISA EC project). In July 2018, the dataset will be used in *Mobility Data Analytics* course at the ACM summer school in data science[11]. It will also be the core dataset to be used in a textbook on *Maritime Informatics* driven by Dr Dimitris Zissis (University of the Aegean) and Dr Alexander Artikis (NCSR Demokritos) and to be published in 2019.

[10] Mainly, S57 extracts from S57 nautical charts. A technical note describing which nautical charts and objects are useful for the dataset and the necessary scripts to process them has been written and proposed: https://zenodo.org/record/1182539

[11] https://summerschool.acm.org/2018/learning-from-our-movements-mobility-data-analytics

3 AIS stream

IMIS is delivering an AIS data steam for the area of interest shown in figure 1 via a TCP/IP data stream. The data format is similar to that presented in section 2.1 but is presented as a data stream as opposed to a static file. The characteristics of the datastream are given in the following section.

3.1 Datastream characteristics

- **Area of interest:** North West Coordinates are (52 Degrees North, 12 Degrees West) and South East Coordinates are (30 Degrees North, 37 Degrees East). Shown in Figure 1.
- **ITU Message types provided:** ITU1-2-3, ITU5, ITU9, ITU18, ITU19, ITU21, ITU24
- **Average AIS message velocity:** A total of ≈ 25K-30K over 15 minutes. Figure 3 shows the message velocity over 15min intervals for one day.
- **Average number of vessels transmitting AIS:** ≈ 35 000. Figure 4 shows the number of vessels transmitting AIS messages over 15min intervals for one day.

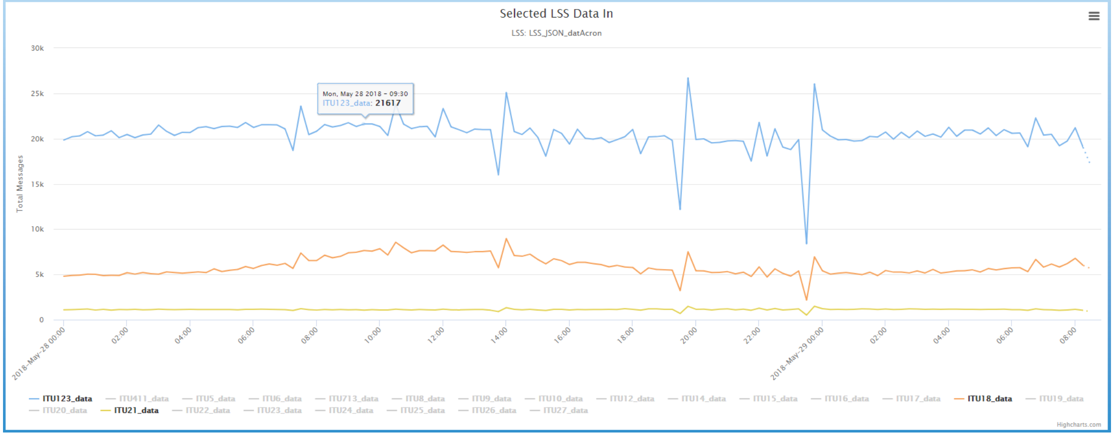

Figure 3: Data stream velocity for one day over the area of interest

3.2 In-stream error detection

The IEC 61162-1 messages can originate from a variety of different sources including AIS receivers, base stations as well as satellite AIS. The messages from these sources are well defined but often produce faulty messages. These faulty messages may result from system faults, data errors and deliberate actions to inject faulty data which significantly influence the veracity of the data which is an important consideration within the datAcron project. For this reason each of the IEC 61162-1 messages coming in are analysed and suspicious or faulty messages are appropriately flagged and stored for later analysis. The attached flag contained in the Comment

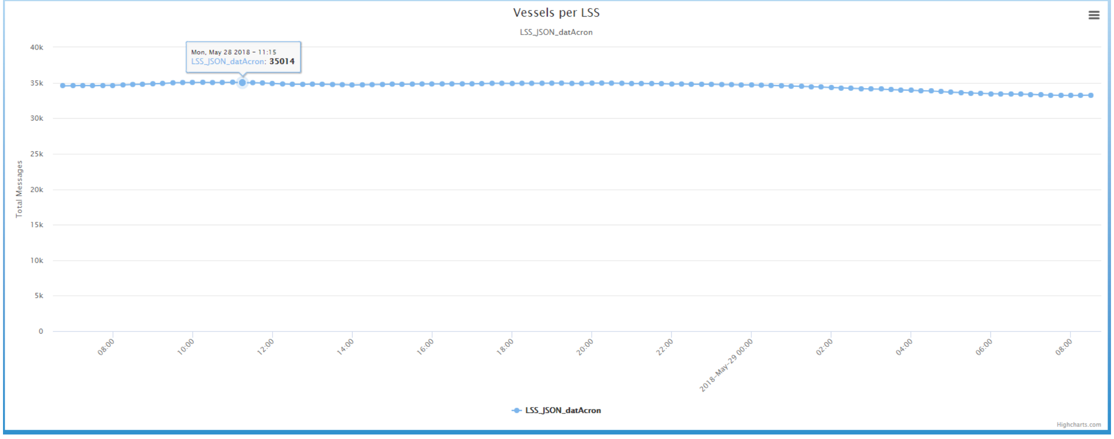

Figure 4: Number of vessels transmitting AIS for one day over the area of interest

Block / TAG Block will identify the message as requiring additional processing and / or attention. Where a source of data is identified as being suspect, this data is flagged along with the reason for the flag being attached.

The following errors are detected and set separate flags in the comment / TAG block, if detected in the message:

1. **Message timing validation:** Time of arrival (TOA) is used to augment and validate the received position.

2. **Slot boundary detection:** Where the slot number from multiple data sources do not coincide, the message is tagged as being suspect.

3. **Message reporting rate validation:** Validating minimum and maximum number of messages per time period.

4. **Improperly formatted MMSI:** Detects messages with a MMSI which are not in the correct range as per the MMSI specifications

5. **Message channel Validation:** Checks that the type of the ITU-R.M 1371-4 message is an authorised type for this specific RF channel.

6. **NMEA 0183 version 4 and / or IEC 61162-1 message type validation:** Validates checksum on both the comment and the NMEA section.

7. **Area definitions:** Checks if AIS positional report is in a valid maritime environment.

8. **Duplicate / MMSI target check:** Checks if two vessels which are geographically separate report with the same MMSI, IMO number, Call Sign and Name

9. **Hydrodynamic validation:** Checks validity of Speed Over Ground and Rate of Turn against the reported beam and length as well as vessel type.

10. **ITU-R M. 1371 message content validation:** Checks if a message has all fields configured for verification and are within the range of the applicable specification.

11. **Talker Identity validation:** Messages from specific talkers, such as AIS, ECS or PC are processed.

12. **MMSI / MID Blacklisted or Wanted vessels:** Checks against a list of wanted or blacklisted MMSI or IMO numbers and can be flagged if required.

4 Data degradation and preparation

The question addressed in this section is how to effectively assess the quality of the results provided by algorithms (e.g. synopses generation, event detection and forecasting) at the Maritime Situational Indicators (MSIs) and scenario levels [2], used in the context of maritime surveillance, as described by the different uses cases of D5.1. In order to properly evaluate algorithms, their output should be compared against a validated (reference) dataset of trajectories labeled with proper indicators. However, reference or ground truth information is generally difficult to obtain in large scale moving object data streams and databases (e.g., European scale). Indeed, the annotation of real moving object datasets with Maritime Situational Indicators is extremely challenging as it is time consuming, the proper annotation tools do not exist yet, and the ground truth highly depends on the human task (contrary to more objective labels such as vessel type).

Therefore, the available reference dataset presented in Section 2 has been enriched with reference information. To this end, additional information describing intrinsic quality of the reference dataset has been extracted and further used to generate small datasets (raw or synopses/MSI level) exhibiting typical vessel behaviour, aligned with the scenario definitions [5] and the experimental plan [2] proposed. The approach followed aims at providing simulated behaviours credible to operational experts.

This methodology, designed for the evaluation of maritime situations with experts has been experimented in March 2018 with WP4 partners. The prototype setup and the results of this experiment based on collision scenarios will be reported in the deliverable D5.5.

Datasets created for experimental validation might be either purely synthetic and automatically generated based on some motion models, or pseudo-synthetic by modifying existing real data with a controlled process (Figure 5). In the first case they may be biased by the model applied, while in the second case they preserve some characteristics of the original observations. However, they cannot be produced massively because the labelling is done manually.

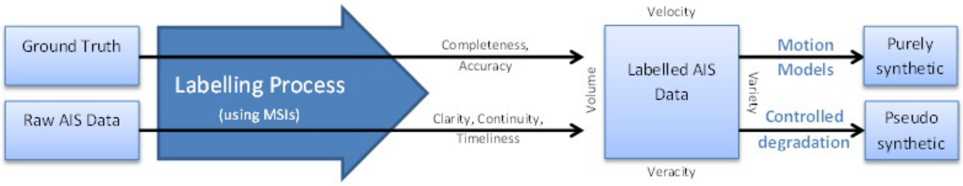

Figure 5: Labeled AIS datasets generating process

The next sections detail the data degradation and enrichment techniques developed in the scope of datAcron algorithms evaluation. The generation of modified data focuses on AIS data, however it can be accompanied by the alteration of real geographical areas, METOC information and other contextual data. This includes for example the modification of regulated or protected areas in space and time (i.e., moving a polygon, reducing the period for fishing, . . .). Similarly, a modification of a real trajectory to instantiate suspicious behaviours in the maritime security scenario can be complemented with information of the ship being on black or wish list.

4.1 Data degradation techniques

In order to vary the veracity of the dataset, several techniques have been developed to reflect the well documented lack of veracity in AIS data. Indeed, AIS data are not perfect as errors, falsifications and spoofing cases [15], or inaccuracy or incompleteness with missing data fields [4] have been demonstrated.

Additionally, we provide methods to modulate the data quality by either removing suspicious data (e.g.,*data cleansing*) or worsening the data quality (*data degradation*). In this work, we do not focus on data cleansing methods and rather define means to degrade AIS data along to the corresponding dimensions of veracity.

The proposed methods can be classified into four families: 1) noise adjunction, 2) data modification, 3) data removal and 4) data addition. The four methods are illustrated in Figure 6, where a dataset of two trajectories is modified according to the four above methods. The upper part of the figure shows the original data, as it is available in our dataset. The bottom part presents the resulting datasets after the application of the four techniques presented in the rest of this section.

4.1.1 Noise adjunction

Noise adjunction consists in the blurring of the actual observed data by applying a random Gaussian shift to the value, centered in the value itself and with a variable standard deviation, to be set accordingly to the degree of noise desired. Noise adjunction may be applied to rate of turn, speed, course, heading, longitude and latitude data fields in the AIS dynamic messages and to the draught as well as to vessel dimensions reported in the AIS static messages.

4.1.2 Data modification

Data modification consists in the change of a data field or of several data fields. Such modification can be targeted on the identity of the vessel (i.e., identity fields of the static messages such as name, IMO number, call sign, or the national vessel identity number fields for both static and dynamic messages) or on the whereabouts. Such a modification can affect either a single point or a whole trajectory.

4.1.3 Data addition

Data addition consists in the creation of new synthetic data in the dataset, leading to the augmentation of the number of entries. Several operations of data addition can be distinguished: the copy of a trajectory (with possibly several instances), the change of the temporal or the spatial values of it, and the creation of specific events such as a collision. The later case corresponds to the creation of trajectories of vessels with normal behaviour but which trajectory crosses the trajectory of in the original dataset.

4.1.4 Data removal

Data removal consists of the deletion of some data referring to some “missing data mechanisms”. Three main missing data mechanisms are distinguished: missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR) [17, 4]. The object of the deletion can be either the data field or the entire message. In the case of data field deletion, the number of entries in the dataset is not reduced, however some data fields are emptied, which means their value is turned to null. The data fields can be selected randomly or targeted and the frequency of the removal is a parameter of the veracity variation. In the case of message deletion, the number of entries decreases as the entire entries in the dataset are removed. The deletion can then affect either a single entry, either a single trajectory, either a succession of several consecutive points or randomly selected points in particular to simulate the lack of coverage of the AIS receivers. It is possible to simulate the existence of *black holes*, [16] or simulate the coverage quality of a receiving station, where the probability for the message to be actually received varies with the location of the vessel.

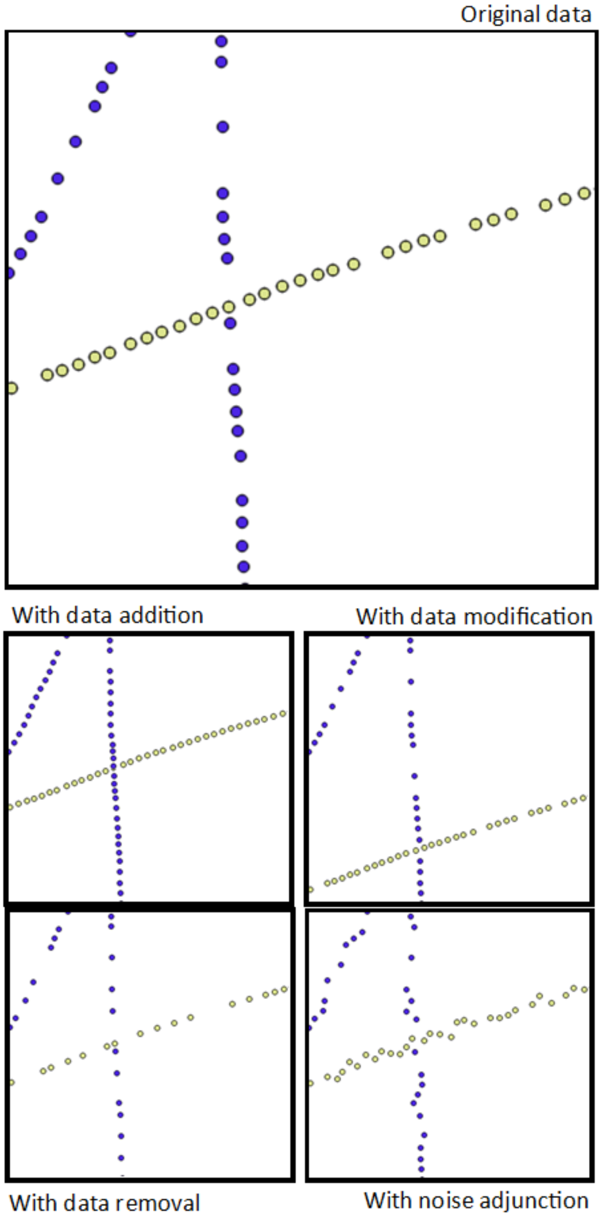

Figure 6: Application of data degradation techniques to a data sample

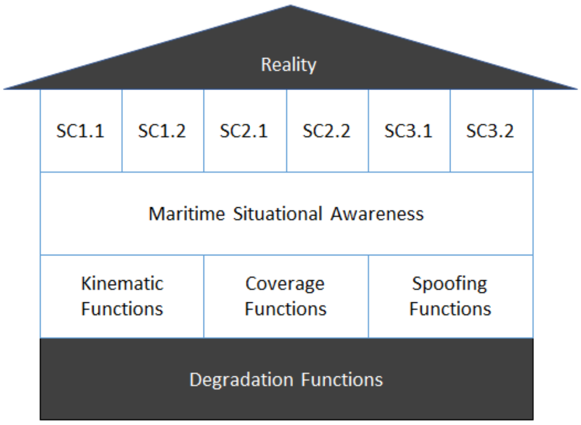

Figure 7: The inductive approach: the code construction

4.2 A library of degradation functions

A toolbox of degradation functions has been designed and implemented, based on the data degradation techniques presented in section 4.1, to be further used to degrade a given dataset automatically.

To design the degradation functions, we first identified the modifications applicable to each AIS field (see Figure 7). Degradation functions can be categorised along:

- **Kinematic degradation**, modifying the dynamic aspect of a trajectory, i.e., its position, speed.
- **Coverage degradation**, for example simulating a poor emission, poor reception, no voluntary emission.
- **Spoofing degradation**, modifying the content of the AIS static messages (cf. Section 2.2.1) such as the destination, name, MMSI fields.

We particularly focused our attention on the kinematic degradations category. Table 2 summarises all degradation methods that have been implemented, further classified into basic functions, superior functions and advanced functions depending on their complexity:

- **Basic functions** can be implemented without using other functions, like the modification of AIS fields;
- **Superior functions** use only basic functions, to apply and allow more realistic and coherent modifications;
- **Advanced functions** use superior functions and manage to create entire navigation patterns.

In the remainder of this section, we detail the implementation of one function per category.

Delete trajectory waypoints (deleteWP(i))

This function deletes the i^{th} waypoint in the current trajectory. In Figure 8, we show an example of the result of this function. The original trajectory is depicted in orange while the modified one is in green. To clearly visualise the result of the deleteWP() function, the modified trajectory has been shifted of 30arcmin to the East.

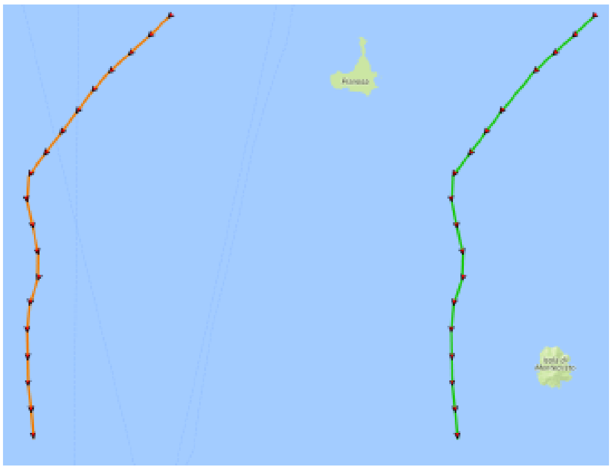

Figure 8: Result of the function deleteWP(4)

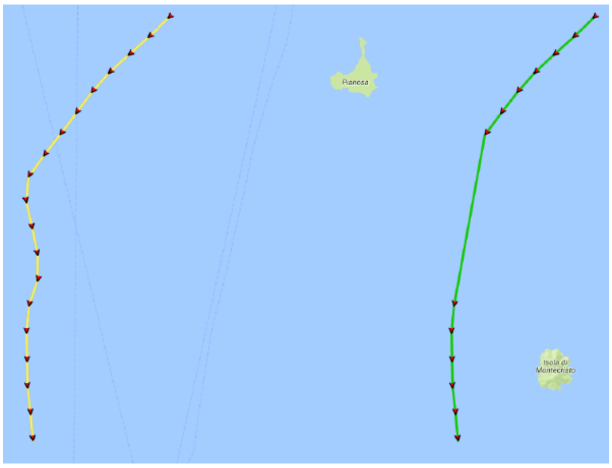

Figure 9: Result of createGap("2015-02-01 00:20:00","2015-02-01 00:50:00")

Basic Functions	Superior Functions	Advanced Functions
modifyMmsi(mmsi)	movePoint(i,lng,lat)	reverseCourse()
modifyPosition(lng,lat)	addPoint(lng,lat,cog=null,trueheading=null,sog=null,rot=null)	routeTo(lng,lat)
translate(decLng,decLat)	rotate(angle)	joinTo(trajectory)
deleteWP(i)	createGap(dateOfBegin,dateOfEnd)	
calDist(i,j)	simulateCovering(conditions)	
calCOG(i,j)		
set_date(idateString		
cut(i)		
copier()		
paste(trajectory,atTheEnd=false)		

Table 2: Library of degradation functions within the toolbox

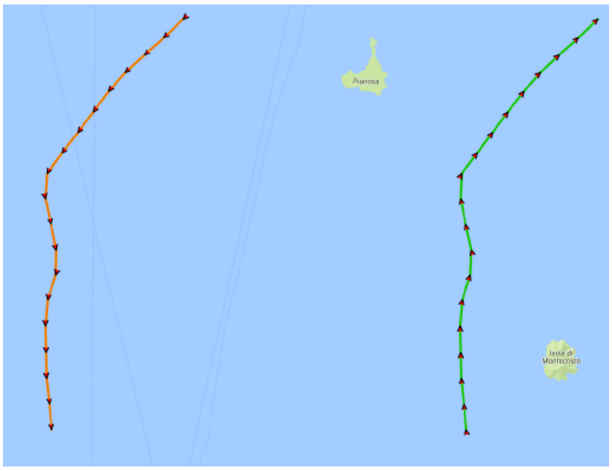

Figure 10: Result of reverseCourse()

Create gaps in trajectories (createGap(dateOfBegin,dateOfEnd))

This function deletes all the waypoints of the current trajectory between the dates "dateOfbegin" and "dateOfEnd". This function enables the effect of an AIS transponder voluntarily turned off (see figure 9) to be simulated. For a better visualisation, the modified trajectory has been shifted of 30arcmin to the East.

Reverse the course of a trajectory (reverseCourse())

Reverses the direction of the current trajectory, using the same path and the same dates for the waypoints but with opposite course over ground (see figure 10). For a better visualisation, the modified trajectory has been shifted of 30arcmin to the East.

4.2.1 Automation of function selection

Given the previously described degradation functions, an extensive set of trajectories can be derived from an original dataset of real vessel trajectories. In order to artificially build the patterns corresponding to the different datAcron scenarios, additional constraints need to be taken into account during the degradation process.

For example, to create a collision between two ships, their respective trajectories must intersect spatially and the locus of the collision must encompass waypoints with identical emission dates. Two approaches can be followed to create such a pattern:

- either we select a pair of trajectories that are already spatially superimposed but which do not overlap temporally and we modify their timestamps so they match at the specific point of collision,

- or we artificially make them intersecting by means of some kinematic modification (in this case applied to the position) functions described above.

The library of degradation functions provides thus a rich set of basic constructs to build different patterns in a non unique manner. Having different ways to produce equivalent patterns offers a good flexibility in the creation of these patterns which is desirable for instance to make them more realistic or provide a wider diversity in similar patterns.

4.2.2 Various use cases for degradation automation

We detail here an example of use of this library to artificially create collisions between vessels. To create a collision, six combinations of degradation functions are possible, as presented in Table 3. All these options must be linked to specific situation. Indeed using reverseCourse() in a TSS creates a possible collision but an anomaly detection algorithm will detect it very easily, so reverseCourse() cannot be applied in the case of a TSS. However, it can be very artful to apply it in a canal or more generally in confined waters, where ships are used to evolve very close to each other, in opposite directions.

Type	Combination	Situation
1	reserveCourse()	confined waters
2	set_date()	last resort or trajectories entering/exiting an harbour
3	routeTo() + setDate()	trajectories in a bottleneck
4	translate() + setDate()	trajectories with a small bounded box
5	rotate() + setDate()	great bodies of water and distant parallel trajectories
6	movePoint() + setDate()	close parallel trajectories

Table 3: The different combinations for the collision scenario and their use for different situations

4.3 Data preparation

In order to perform a given experiment, a scenario dataset has to be prepared, featuring all desired data and events. This data preparation has been organised to fit with experiments involving maritime experts. In that context, the expert visualises detected events on the interactive user interface developed by WP4 with the objective to understand and assess the maritime situation. The maximum period retained for such experiments is 30 minutes that can be organised in different manners depending on the evaluation objectives[12]. The experimental dataset can be divided into:

- About 30 minutes of continuous data (*i.e.* there is only one time frame of about 30 minutes of data)
- Three consecutive time frames of about 10 minutes each, in which each time frame tells a story and is not linked with the other ones. In this case there are two occurrences (about 10 minutes and about 20 minutes after the start of the experiment) when the maritime situational picture changes altogether and where two consecutive pictures do not depict the evolution of the first one towards the second one.
- A succession of inhomogeneous time frames depicting typical maritime situational cases.

[12] Let us note that the assessment of events at the MSI level is not limited by the extent of the dataset prepared for the scenario level assessment. Instead, the whole reference dataset is considered, possibly enriched with this experimental datasets

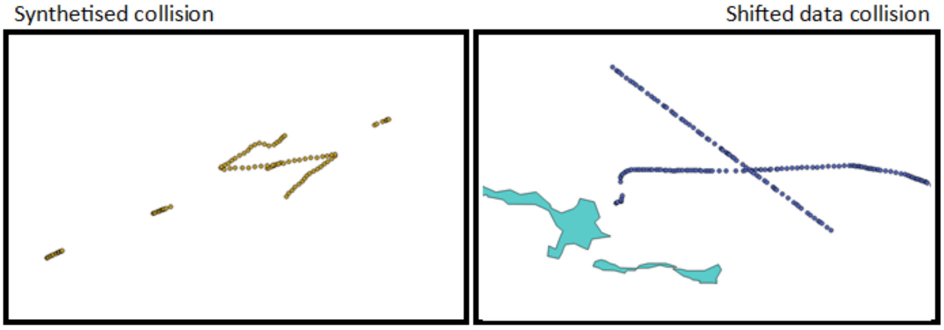

Figure 11: Data preparation of collision cases

Several situations can be created and integrated as part of the different scenarios. The situation presented to the user will then consist of a set of normal AIS tracks enriched with specific events as described below. These events can include:

- A collision between a real vessel trajectory and a synthetic vessel trajectory;
- A collision between a real vessel trajectory and the shifted trajectory in time and space of another real vessel trajectory;
- A synthetic near-collision;
- The shift in time of a real tugging case;
- The shift in time and space of specific trajectories in order to simulate a given behaviour.

A data shift in space is defined by two parameters Δlon and Δlat, signed real numbers standing for the desired shift in degrees of longitude and latitude, respectively. This is useful to spatially align two events, or to place a trajectory in a given location. A data shift in time is defined by the parameter Δt, signed integer standing for the difference wanted in seconds. This temporal shift is useful to temporally align two events, or to force one trajectory at a given time. A spatio-temporal shift combines both in order to align two events or trajectories in space and time.

4.3.1 Collision simulation

A collision may be simulated by either synthetising or shifting a trajectory that would cross the trajectory of an existing vessel in the dataset in a point that can be either pre-defined or randomly selected. The purpose here is to create various situations to stimulate the expert's reaction. By showing collisions with trajectories which are degraded differently, some assessment of the expert's situational awareness can be made (including response time).

Figure 11 presents two cases of collision: on the right-hand side of the image is shown a synthetised collision (created data), whereas on the left-hand side of the image, a shifted trajectory is shown. Both the simulated vessel and the actual vessel types can be chosen to correspond to a specific situation (e.g. two cargo vessels colliding, or one cargo vessel colliding with one fishing vessel).

The simulated trajectory must be realistic and should comply with the vessel kinematics, as well as the speed, course and possible turning behaviour that vessels usually have in that neighbourhood. In addition, it must be ensured that the newly created trajectory complies with the landmasses and possible restricted maritime areas so that the trajectory remains as much as possible realistic.

The modification of data is performed through a program in R taking into account a variety of parameters. A list of the main parameters is presented here:

- *original_data* - The name of the table in the database;

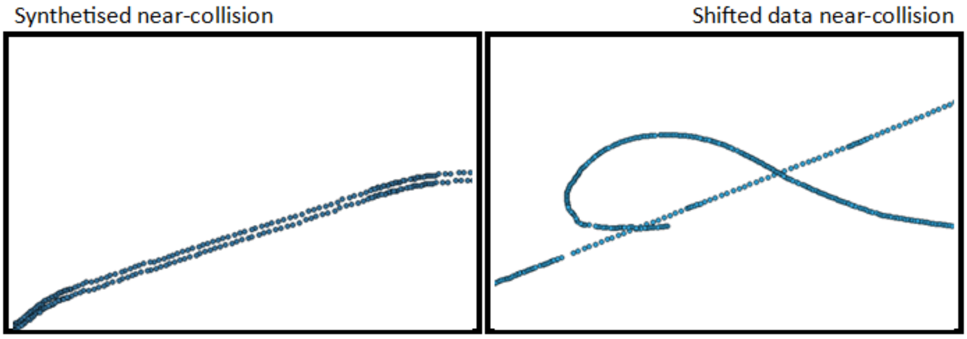

Figure 12: Data preparation of near-collision cases

- *temporal_ constraint* - Specifies if a temporal constraint will be applied to the dataset. if false, all trajectories are selected;
- *spatial_ constraint* - Specifies if a spatial constraint will be applied to the dataset. if false, all trajectories are selected;
- *begin_ sec, end_ sec* - Temporal boundaries of the temporal constraints. Ignored if if temporal_constraint is FALSE;
- *lat_ min, lat_ max, lon_ min, lon_ max* - Spatial boundaries of the spatial constraints. Ignored if spatial_constraint is FALSE;
- *wschema, newtablename* - The names of the working schema and the new table containing the synthetised collisions;
- *speed_ low, speed_ high* - Minimal and maximal speed value for a point;
- *traj_ time, traj_ hole, traj_ minpoints* - Trajectory detection variables, corresponding to the maximal time for a trajectory, the maximal time between two consecutive points for the trajectory to be considered as valid and the minimal acceptable number of points for a valid trajectory;
- *colnumbers* - The number of collisions to synthetise in the dataset;
- *colangles* - Preferred angle of collision for the synthesis;
- *coltraj_ minpoints, coltraj_ maxpoints* - Minimal and maximal number of points for a synthetised trajectory;
- *validity_ landmasses* - Do we check or not if no synthetised point is overlapping a landmass.

4.3.2 Near-collision simulation

The purpose of creating near-collision events is to attract the attention of the expert, so that the expert user has to distinguish between real collisions and near-collisions, and raise an alert properly. A near-collision can be synthetised using the real trajectory to which the near-collision is done. The contacts are copied, the direction is reversed, the time is reversed so that the near-collision occurs at the wished location. An example of synthetised near-collision is presented in the left-hand part of Figure 12. This figure also features on its right-hand side a case of shifted data near-collision creation. Similarly to the case of a collision creation, a trajectory is shifted to cross in space and to create a near-collision in time with another targeted trajectory. The Δt between the passage of the two vessels over the same point can be adapted to the size of the vessels, their alleged manoeuvrability and the degree of near-collision desired.

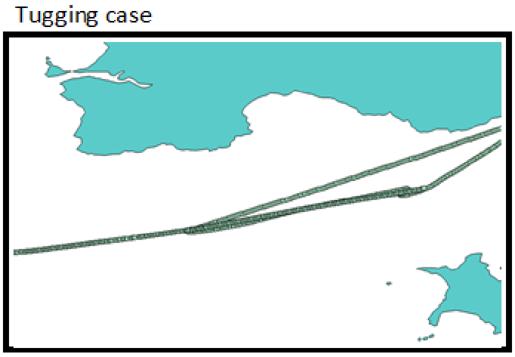

Figure 13: Data preparation of tugging case

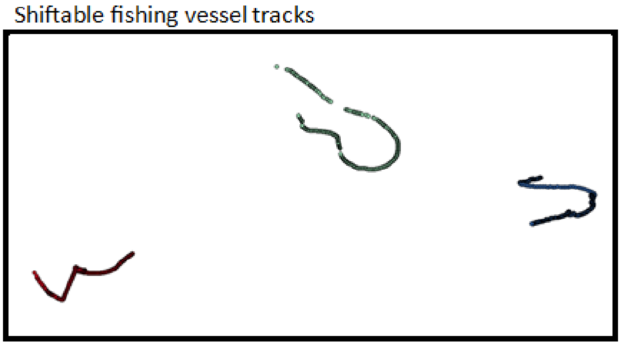

Figure 14: Data preparation of fishing vessel trajectories to be shifted in time and space

4.3.3 Tugging case

One case of tugging has been spotted in the available data, at the entrance of the Brest port. The two trajectories have been isolated and stored separately in a dedicated table. It is now possible to take them and by applying the correct time shifting, display a tugging case whenever needed. The purpose is to provide another situation to the expert user that might look like a collision but that is perfectly normal in the day-to-day business of a port. The case in question is presented in Figure 13.

4.3.4 Trajectory shifting

A trajectory may be shifted in time and space whenever its behaviour, type or activity suites a scenario. Since we focus on fishing vessel trajectories, a total of 15 fishing vessel tracks displaying fishing patters have been selected. These trajectories may be dropped whenever and wherever needed, in order to simulate dangerous or illegal behaviours as described in the use case: e.g., fishing in the middle of the Brest bay, fishing in a restricted area or in the middle of a Traffic Separation Scheme. The purpose is to draw the attention of the expert and make him or her notice the unusual behaviour. One of those cases is presented in Figure 14.

4.4 Variety variations in datasets

Additionally to the variety variations described above, we also defined some variety variations in the data. Variety variations rely on the use of a set of heterogeneous data sources used to complement the AIS data in the understanding task of the maritime situation. Indeed, an abnormal-looking situation could be resolved by means of another source of information, and a normal-looking situation from the sole AIS point-of-view could result in an anomaly when confronted to other data sources.

Those complementary data sources must be aligned in time and space with the AIS data and can cover a large spectrum of data types including for example weather conditions, fleet registers or maritime areas boundaries (those data are a subset of those present in the reference dataset [13]). Introducing variations in variety consists, starting from the use of AIS data alone (minimal variety), to progressively include additional datasets and thus increase the variety.

5 Processed data

From a maritime use case point of view, the main outputs of datAcron algorithms to be assessed with maritime experts are events detected by synopses and complex events detection and prediction as they correspond to the different MSIs previously defined.

Besides the quality of input data (discussed in the previous section), the outputs of these algorithms are mainly affected by internal parameters (such as possible thresholds) or by statistical or logical models choices. Additionally, the interpretation of the detection and prediction of MSIs by the expert depends on contextual information such as the fishing areas, the maritime routes or stationary areas, together with additional semantics as for instance labelling the maritime routes with ports names.

In the following, we provide some details about how this contextual information about patterns-of-life of vessels has been extracted.

5.1 Extraction of maritime routes

5.1.1 The TREAD software

The maritime routes have been extracted following the methodology presented in [11], called Traffic Route Extraction and Anomaly Detection (TREAD), which learns a statistical model for maritime traffic from AIS data in an unsupervised way. Following the work of [20, 19, 10], the knowledge of the traffic is based on the notion of vessel objects, that are created and updated from AIS messages. A bounding box of latitude and longitude values defines the area of study.

Within the selected bounding box, various geometrical objects are extracted from the data such as the stationary areas (called POs), the entry areas (called ENs) and the exit areas (called EXs), referring to the entry or exit of the vessel relatively to the bounding box. The geometrical objects called waypoints (WPs) are linked by another category of objects, namely the route (called Rs). A route represents an oriented link between an inbound waypoint and an outbound waypoint and can therefore link two POs, one EN and one PO, one PO and one EX, or one EN and one EX. On the basis of the knowledge extracted from those routes, anomalies can be spotted, provided the assumption that the statistical model corresponds to a stationary distribution of normal traffic. The feature data are considered as single trajectory points following the point-based approach as defined in literature (such as in [7]) rather than the trajectory-based approach (such as in [8]).

In order to manage the computation and as presented in section 5.1.1, four kind of objects are created and manipulated: the *vessels* (a contact is associated to an MMSI number, which is a unique identifier for a vessel), the *stationary areas* (clusters of points with low speed), the *entry or exit areas* (clusters of points with significant speed, located at the limits of the bounding box) and the *routes* (sequences of positional contacts connecting two stationary areas, a stationary area and an entry or exit area, or and entry and an exit area). The characterisation of the stationary areas, which include the port areas, are presented in the next two subsections.

5.1.2 Stationary areas

Stationary areas (POs) include ports, offshore platforms or anchorage areas. Objects in this class are clustered contacts having a speed lower than a given threshold. The clustering is computed using the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm [3]. The activities in those areas (ports, offshore platforms, fishing grounds, anchorage or waiting areas) can be further deduced using the type of vessel and the duration of presence in the area.

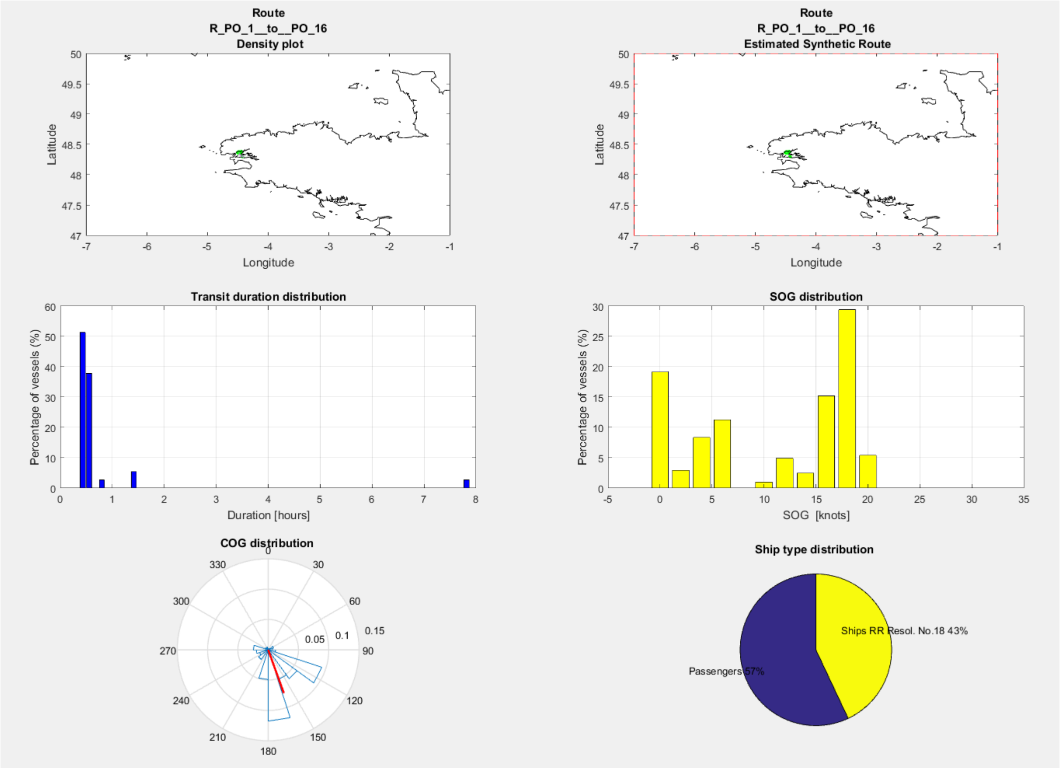

Figure 15: Statistical values of a TREAD route

5.1.3 Vessel routes

After the construction of the areas vessel routes are computed by clustering the vessel data flows and allowing to distinguish between local routes (connecting two ports) and transit routes (connecting an entry area and an exit area, a port and an exit area, or an entry area and an exit area). Not only do those route objects register the transiting vessels, but they also register statistically relevant information about both static and kinematic features of the vessels such as the speed, COG, type of vessels which transited along that route.

When a new contact is detected within the bounding box, its characteristics are compared to the already existing set of routes and if the features are considered compatible with one of them, the vessel is added to the list of vessels following the route and the corresponding contacts are treated consequently. If no match is determined, the contact is used for initialising a new route, which is activated once the minimal number of contacts is reached (the minimal number of contacts is a parameter of the algorithm, which may be customised as needed).

In order to have a graphical representation of the route features, the TREAD tool allows to plot some of the kinematic and geographical values, as shown in Figure 15 for the Brest area. The density plot and the synthetic route are presented, as well as the distribution of the transit duration, the distribution of the speed over ground (allowing to perform statistical analyses on the speed value such as extracting the mean, the median, or the mode in the frame of a distribution characterisation), the circular distribution of the course over ground values (allowing the performance of statistical computation such as the mean direction of the trajectory, and to have a visually comprehensive view of the state of the trajectory) and the distribution of the ship type (enabling the characterisation of the route outbound waypoint nature).

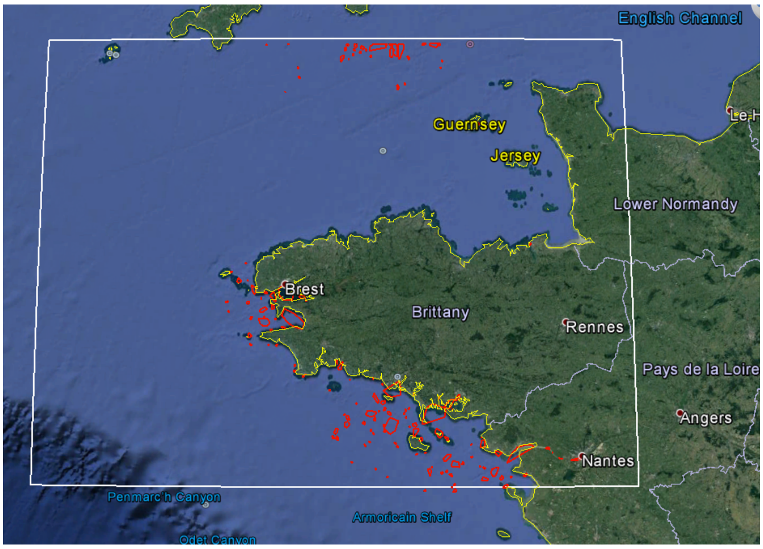

Figure 16: Output of the TREAD software, showing all the stationary areas in the bounding box. Map data: Landsat/Copernicus, Google Data SIO, NOAA, U.S. Navy, NGA, GEBCO

5.2 Port association

In order to perform a comprehensive study of the maritime traffic and a proper exploitation of the maritime routes extracted, additional semantics is useful to ease the interpretation of the results by the user. In particular, TREAD routes only connect different types of areas but does not provide information about the nature of this route such as "Brest-Southampton" or "BrestPort-FishingAreas". The type of a stationary area can be a port, a waiting area, an anchorage area, a fishing ground or a offshore platform which can be further estimated through some classification method which will not be addressed in this work.

The names of ports are extracted from the World Port Index. Given the latitude and longitude value, the nearest port in the World Port Index is allocated to the stationary area that is close enough to the shore (or that overlaps the landmass).

Once the name of the port has been associated to a stationary area, it is possible to estimate the destination of vessels following a route ending in this port. In addition, it is also possible on the one hand to infer the activities of the port through the type of the incoming and outgoing vessels, and on the other hand, if the nature of the activities in the port are known, to detect anomalies in case an unexpected type of vessel is approaching the port infrastructures or following a route ending or beginning in this port.

A map of all stationary areas as computed by TREAD is presented in Figure 16, with a zoom of the Brest area in Figure 17. As one can see, the size of those areas can vary considerably.

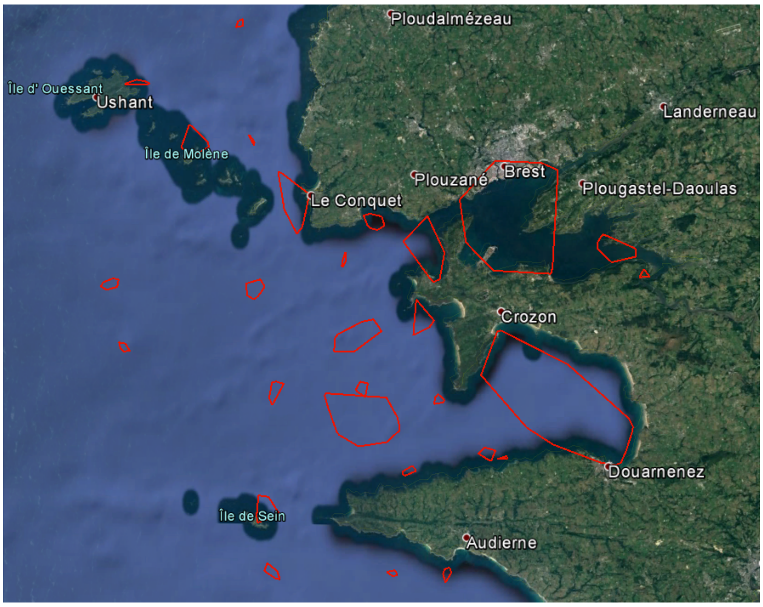

Figure 17: Zoom of the stationary areas of the Brest Roadstead area. Map data: Landsat/Copernicus, Google Data SIO, NOAA, U.S. Navy, NGA, GEBCO

6 Conclusions

The report concludes the task 5.2 of the maritime work package. Prepared historical data and the AIS stream have been presented in Sections 2 and 3. The methodology and data structures designed for the evaluation are detailed in Sections 4 and 5.

The contribution of this work package is twofold. The first result presented in Section 2 and Section 3 concerns the preparation of maritime data. While efforts, especially in Europe, have been initiated to centralise maritime data information, most of the data are of of heterogeneous type and format and still independently sourced and maintained. In datAcron, we provided an heterogeneous dataset that combines in space and time a large variety of maritime information. The dataset contains four categories of data: navigation data, vessel-oriented data, geographic data, and environmental data. It covers a time span of six months, from October 1st, 2015 to March 31st, 2016 and provides ships position within Celtic sea, the Channel and Bay of Biscay (France). Beyond datAcron, the dataset has been designed to support researches focusing on methods for the detection and prediction of trajectories, mobility patterns and complex events related to moving entities at sea. A public (sub)dataset (depending on licences) has been published on a European repository under Licence CC-BY-NC-SA-4.0 with predefined integration and querying principles [13, 14]. In order to stress datAcron algorithms and architecture under high volume and velocity, a data stream has been also set up. The stream covers all European coasts and seas.

Secondly, a methodology and related algorithms and data structure have been proposed to challenge the algorithms with the veracity issues of large moving objects databases (Section 4). This has been done by defining qualitative metrics and by computing reference information describing intrinsic quality of navigation data. Such statistical analyses have been further used for the generation of scenarios datasets (raw or synopses/MSI level) designed to exhibit typical maritime situations. To do so, high quality data degradation functions based on navigation data assessment have been designed and implemented. This methodology also comprises the computation of complementary patterns such as maritime routes and stationary areas. This is a key factor for the assessment of results by experts under the conditions of use of the datAcron prototype.

References

[1] Edwige Boudiere, Christophe Maisondieu, Fabrice Ardhuin, Mickael Accensi, Lucia Pineau-Guillou, and Jeremy Lepesqueur. A suitable metocean hindcast database for the design of marine energy converters. *International Journal of Marine Energy*, 3-4:e40–e52, 2013.

[2] Elena Camossi, Anne-Laure Jousselme, Cyril Ray, Melita Hadzagic, Richard Dreo, and Christophe Claramunt. Maritime experiments specification, H2020 datacron d5.3. 2017.

[3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[4] Anne-Laure Jousselme and Patrick Maupin. *Uncertainty Representations for Information Retrieval with Missing Data*, pages 87–104. Springer International Publishing, Cham, 2016.

[5] Anne-Laure Jousselme, Cyril Ray, Elena Camossi, Melita Hadzagic, Christophe Claramunt, Karna Bryan, Eric Reardon, and Michael Ilteris. Maritime use case description, H2020 datacron d5.1. 2016.

[6] Christos Kalyvas, Athanasios Kokkos, and Theodoros Tzouramanis. A survey of official online sources of high-quality free-of-charge geospatial data for maritime geographic information systems applications. *Information Systems*, 65:36 – 51, 2017.

[7] Rikard Laxhammar. Anomaly detection in trajectory data for surveillance applications, 2011. Licenciate Thesis, Örebro University, Örebro, Sweden.

[8] Brendan Morris and Mohan Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1114–1127, 2008. doi: 10.1109/TCSVT.2008.927109.

[9] Fabrizio Natale, Maurizio Gibin, Alfredo Alessandrini, Michele Vespe, and Anton Paulrud. Mapping fishing effort through ais data. *PloS one*, 10(6):e0130746, 2015.

[10] Giuliana Pallotta, Michele Vespe, and Karna Bryan. Traffic route extraction and anomaly detection (tread): Vessel pattern knowledge discovery and exploitation for maritime situational awareness, 2013. NATO Formal Report CMRE-FR-2013-001, NATO Unclassified; NATO: Brussels, Belgium.

[11] Giuliana Pallotta, Michele Vespe, and Karna Bryan. Vessel pattern knowledge discovery from AIS data - A framework for anomaly detection and route prediction. *Entropy*, 5(6):2218–2245, 2013.

[12] Cyril Ray, Elena Camossi, Anne-Laure Jousselme, Melita Hadzagic, Christophe Claramunt, and Ernie Batty. Maritime data preparation and curation, H2020 datacron d5.2. 2016.

[13] Cyril Ray, Richard Dreo, Elena Camossi, and Anne-Laure Jousselme. Heterogeneous integrated dataset for maritime intelligence, surveillance, and reconnaissance (version 0.1), February 2018. Data set. Licence CC-BY-NC-SA-4.0. Zenodo. doi: 10.5281/zenodo.1167595.

[14] Cyril Ray, Richard Dreo, Elena Camossi, Anne-Laure Jousselme, and Clement Iphar. Heterogeneous integrated dataset for maritime intelligence, surveillance, and reconnaissance. *Data in Brief*, 2018. Submitted.

[15] Cyril Ray, Clement Iphar, Aldo Napoli, Romain Gallen, and Alain Bouju. Deais project: Detection of ais spoofing and resulting risks. In *OCEANS'15 MTS/IEEE, Genoa, Italy*. IEEE, 2015.

[16] Loic Salmon, Cyril Ray, and Christophe Claramunt. Continuous detection of black holes for moving objects at sea. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on GeoStreaming*, IWGS '16, pages 2:1–2:10, New York, NY, USA, 2016. ACM.

[17] Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002. doi: 10.1037//1082-989X.7.2.147.

[18] Michele Vespe, Maurizio Gibin, Alfredo Alessandrini, Fabrizio Natale, Fabio Mazzarella, and Giacomo C. Osio. Mapping eu fishing activities using ship tracking data. *Journal of Maps*, 12(sup1):520–525, 2016.

[19] Michele Vespe, Giuliana Pallotta, Ingrid Visentini, Karna Bryan, and Paolo Braca. Maritime anomaly detection based on historical trajectory mining. In *Proceedings of the NATO Port and Regional Maritime Security Symposium*, 2012.

[20] Michele Vespe, Ingrid Visentini, Karna Bryan, and Paolo Braca. Unsupervised learning of maritime traffic patterns for anomaly detection. In *Proceedings of the 9th IET Data Fusion and Target Tracking Conference*, 2012.