Grant Agreement No: 687591

28/12/2018

Big Data Analytics for Time Critical Mobility Forecasting

datAcron

Maritime final validation

Deliverable Form	
Project Reference No.	H2020-ICT-2015 687591
Deliverable No.	5.6
Relevant Work Package:	WP 5
Nature:	Report
Dissemination Level:	Public
Document version:	1.3
Due Date:	31/12/2018
Date of latest revision:	28/12/2018
Completion Date:	28/12/2018
Lead partner:	NARI
Authors:	Maximilian Zocholl, Anne-Laure Jousselme, Richard Dréo, Cyril Ray, Clément Iphar, Francesca de Rosa, Elena Camossi, Guillaume Keraudren, Florian Rozé
Reviewers:	George Vouros
Document description:	This deliverable provides a description of evaluation results obtained for the maritime domain.
Document location:	Documents/datAcron/WP5/Deliverables/Final

HISTORY OF CHANGES

Version	Date	Changes	Author	Remarks
0.1	05/12/2017	Mapping components to data variations	CMRE	Internal notes
0.2	08/12/2017	Summary of D2.1	CMRE	Internal notes
0.3	13/12/2017	Summary of D3.2	CMRE	Internal notes
0.4	18/04/2018	Summary experimental rehearsal	CMRE	Internal notes
0.5	19/09/2018	Summary of D2.3	CMRE	Internal notes
0.6	26/09/2018	Summary of D3.4	CMRE	Internal notes
0.7	01/10/2018	Summary of D3.5	CMRE	Internal notes
1.0	25/10/2018		NARI	D5.6 Initial version
1.1	27/10/2017	Structure proposal	CMRE	Working version
1.2	10/12/2018		CMRE, NARI	Version for review
1.3	28/12/2018		CMRE, NARI	Final version

EXECUTIVE SUMMARY

This deliverable labelled D5.6 summarises the evaluation activities conducted in datAcron for the maritime domain as defined by Task 5.5 "System Evaluation and Impact Measurement". The objective of this task was twofold: firstly, the execution of the experimental plan previously sketched in deliverable D5.3 and refined prior to the experiments. The plan relies on the proper preparation and curation of maritime data as described in deliverable D5.4 as well as the prototype set-up as reported in deliverable D5.5. Secondly, the qualitative and quantitative analysis of the measured outcomes of the experiments from a maritime perspective. The evaluation activities were performed in tight collaboration with the workpackages 1, 2, 3 and 4.

A particular mentioning deserve the two collaborations on the one hand between, CMRE and Ecole Navale (NARI) intensified during the 6 month visit of Cyril Ray, to additional weeks together with Richard Dréo and the 3 months visit of cadets Guillaume Keraudren and Florian Rozé, and on the other hand between CMRE and Demokritos during the 2 months visit of Manolis Pitsikalis at CMRE.

An important aspect of the work reported in this deliverable is the development of the methodology for the evaluation of the datAcron prototype. The original approach proposed allowed a comprehensive and structured coverage of the four big data challenges, the evaluation criteria (both computational and involving humans), as well as the traceability and reproducibility of the results. The methodology developed during the datAcron project can be further reused and improved to support evaluation activities of future collaborative research projects.

The present document is structured as follows. The introduction summarises the purpose and objectives of the document and its relation with datAcron objectives and other deliverables. Section 2 provides an overview of the evaluation methodology: We followed a human-centric approach with the Maritime Situational Indicators (MSIs) playing a pivotal role; the evaluation space is defined to capture the evaluation results in a unified framework and partition the space for reducing the number of necessary experiments; two type of assessments have been performed simultaneously during the project, either purely computational without involvement of maritime surveillance experts and independent assessments of the components involving experts; the reference datasets are described which allowed to challenge the datAcron components along the four big data dimensions and to support the final maritime use case experiments with a collision avoidance scenario. In Section 3 the results from work packages driven evaluation activities have been captured in a unified framework. Outcomes of the Synopses Generator (SG), Complex Event Recognition (CER), Complex Event Forecasting (CEF) and Future Location Predictor (FLP) are summarised. In Section 4, we describe the assessment of the accuracy of MSIs by expert. A first period of assessment allowed for improvement of both the SG and CER, while the second period focused on the CER component. Section 5 the scenario-level evaluation is described. Section 6 highlights the main outcomes of the assessment of the datAcron prototype conducted for the maritime domain. Finally, we conclude the report summarising the main findings and lessons learned.

TABLE OF CONTENTS

Contents

1 Introduction 1
- 1.1 Purpose and Scope . . . 1
- 1.2 Relating to the objectives of datAcron . . . 1
- 1.3 Relation to other deliverables . . . 2

2 Overview of the evaluation methodology 4
- 2.1 Human-centric approach . . . 4
  - 2.1.1 Aligning datAcron components to maritime surveillance needs . . . 5
  - 2.1.2 Semantic levels of assessment . . . 7
- 2.2 Capturing assessment results in a uni ed framework . . . 7
  - 2.2.1 The evaluation space . . . 9
  - 2.2.2 System decomposition and dependency analysis . . . 9
- 2.3 Simultaneous computation-based and expert-based assessment . . . 11
  - 2.3.1 Connection between the two types of assessment . . . 11
  - 2.3.2 Expert-based assessment work ow . . . 12
- 2.4 Reference datasets for evaluation . . . 12
  - 2.4.1 Two AIS data sources for big data challenges . . . 13
  - 2.4.2 Data preparation for expert-based experiments . . . 13

3 Capturing component assessment results 15
- 3.1 Synopses Generator (SG) . . . 15
- 3.2 Complex Event Recognition (CER) . . . 17
- 3.3 Complex Event Forecasting (CEF) . . . 20
- 3.4 Future Location Predictor (FLP) . . . 22

4 Expert-based assessment at the MSI-level 24
- 4.1 Methodology for the assessment of MSIs . . . 24
  - 4.1.1 Expert-based assessment . . . 24
  - 4.1.2 Method followed by the expert . . . 25
- 4.2 MSI assessment, rst period . . . 26
  - 4.2.1 Assessment oStop Event Detections . . . 26
  - 4.2.2 Assessment oUnderway Event Detections . . . 27
  - 4.2.3 Assessment oHigh SpeedEvent Detections . . . 28
- 4.3 MSIs assessment, second period . . . 31
  - 4.3.1 MSI considered . . . 32
  - 4.3.2 Results . . . 33
  - 4.3.3 Assessment of MSI #3: On a maritime route . . . 35

5 Assessment at the Scenario Level 38
- 5.1 Scenario level experiments . . . 38
  - 5.1.1 Types of experiments . . . 38
  - 5.1.2 Scenario selection . . . 39
- 5.2 Experiment 0: Rehearsal . . . 39
- 5.3 Experiment 1: The Variety Game . . . 42
  - 5.3.1 Icons Minigame . . . 42
  - 5.3.2 Maritime Situational Indicators Minigame . . . 43
  - 5.3.3 Information Variety Game . . . 44
- 5.4 Experiment 2: MSIs for MSA . . . 45

5.4.1 Relation between experiment 2 and other experiments 46
5.4.2 Assumptions 46
5.4.3 Criteria and Measures 47
5.4.4 Experimental Design 47
5.4.5 Data labelling 48
5.4.6 Data collected 48
5.4.7 Results interpretation 49
5.4.8 Analysis 51
5.4.9 Conclusions on Experiment 2 57
5.5 Experiment 3: Prototype assessment 58
5.5.1 Analysis 59
5.5.2 Conclusions on Experiment 3 62
5.6 Experiment 4: Expert accuracy MSI assessment 63
5.7 MSI robustness to reduced veracity 64
5.7.1 Experiment description 65
5.7.2 Results 65
5.7.3 Remarks 65

6 Main outcomes 66
6.1 Performance of datAcron individual components 66
6.2 Expert-based accuracy assessment of MSI detections 67
6.3 Robustness of CER to veracity degradation 68
6.4 Situation awareness with datAcron prototype 69
6.5 Conclusions about the methodology 70
6.6 Evaluation perspectives 71

7 Conclusions 72

8 Annex 73
8.1 Robustness to missing data of the CER 73
8.2 Capture of results in the evaluation framework 74
8.3 Data recorded for situation description and con dence in Experiment 2 109
8.4 Data recorded for maritime situational awareness 113
8.5 Agenda of the nal experiments week 117

LIST OF FIGURES

1 Overview of the user-centred datAcron evaluation design [15] 4
2 Three semantic levels of functionalities and corresponding assessment [4] 8
3 Maritime data work ow and expert-driven assessment principle 12
4 Expert accuracy assessment of stop events detected by SG 27
5 Expert accuracy assessment of stop events detected by CER. In green, stop false positive detections, corresponding to AIS contacts reporting speed over 2.7 knots. Red points are consistent stop events. 28
6 Expert accuracy assessment of underway events in the Ushant TSS. Red points are reported positions with speed above 2.7 knots, thus consistent with underway events, while green points are detected underway events with inconsistent speed (< 2.7 knots). Blue points are raw data with speed consistent with underway event, but not detected. 29
7 Expert assessment of accuracy of underway events in the Douarnenez harbour. Inconsistent events are depicted in green. 29
8 Visualisation of high speed detected events. Red dots represent detected high speed events, while blue dots are normal tra c. 30
9 Speed occurrence (in knots) in the original AIS dataset (in blue) and in the compressed data (in red). Extreme speed values are over-represented in the compressed dataset. 31
10 Areas Studied 32
11 CER adrift event detections (red dots) and shing activities identi ed by the expert (coloured circles). Inconsistencies arise when both detection overlap. . . . 33
12 Tugging events 35
13 Within Area detections combining JRC shing areas (in green) and Natura 2000 (in violet). Red dots correspond to detected events by the CER and blue dots are raw data 36
14 Maritime surveillance experts during the experiments of November 2018 held at CMRE. 38
15 Experimental Rehearsal Design 41
16 Near-distance situation taxonomy 46
17 Experiment 2 pairwise scenario design 47
18 Assessments of MSI detection in Experiment 4 63
19 Expert-based Comparison of datAcron Results vs. Enriched and Annotated Data 64
20 Impact of data degradation on the True Positive, False Positive and False Negative detections of "changingSpeed" 73
21 Impact of data degradation on the detection of "ChangingSpeed" 73
22 Impact of data degradation on the True Positive, False Positive and False Negative detections of "gap" 74
23 Impact of data degradation on the detection of "gap" 74
24 Impact of data degradation on the True Positive, False Positive and False Negative detections of "highSpeedNearCoast" 75
25 Impact of data degradation on the detection of "highSpeedNearCoast" 75
26 Impact of data degradation on the True Positive, False Positive and False Negative detections of "lowSpeed" 76
27 Impact of data degradation on the detection of "lowSpeed" 76
28 Impact of data degradation on the True Positive, False Positive and False Negative detections of "movementAbilityA ected" 77
29 Impact of data degradation on the detection of "movementAbilityA ected" . . . 77

30 Impact of data degradation on the True Positive, False Positive and False Negative detections of "movingSpeed" 78
31 Impact of data degradation on the detection of "movingSpeed" 78
32 Impact of data degradation on the True Positive, False Positive and False Negative detections of "sarCourse" 79
33 Impact of data degradation on the detection of "sarCourse" 79
34 Impact of data degradation on the True Positive, False Positive and False Negative detections of "stopped" 80
35 Impact of data degradation on the detection of "stopped" 80
36 Impact of data degradation on the True Positive, False Positive and False Negative detections of "tuggingSpeed" 81
37 Impact of data degradation on the detection of "tuggingSpeed" 81
38 Impact of data degradation on the True Positive, False Positive and False Negative detections of "underWay" 82
39 Impact of data degradation on the detection of "underWay" 82
40 Impact of data degradation on the True Positive, False Positive and False Negative detections of "unusualSpeed" 83
41 Impact of data degradation on the detection of "unusualSpeed" 83
42 Impact of data degradation on the True Positive, False Positive and False Negative detections of "withinArea" 84
43 Impact of data degradation on the detection of "withinArea" 84

LIST OF TABLES

1 Mapping from components to the maritime scenarios [27]. 6
2 Mapping MSIs to datAcron components. 7
3 Summary of WP self-assessment of components: Synopses Generator (SG), Future Location Predictor (FLP), Complex Event Recognition (CER), Complex Event Forecasting (CEF), Interactive Visual Analytics (IVA), Real-time Visualisation (VIZ). 15
8 Qualitative confusion matrix of the expert driven evaluation of SG results for Stop event annotations, referring to Figure 4. True Positive (TP), False Positives (FP), True Negative (TN), False Negative (FN). 26
9 Qualitative confusion matrix of the expert driven evaluation of CER results for Stop event annotations, referring to Figure 5. True Positive (TP), False Positives (FP), True Negative (TN), False Negative (FN). 28
10 Qualitative confusion matrix of underway event annotations, referring to Figure 6 and Figure 7. True Positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN). 30
11 Confusion matrix of route association. Rows = expert labelling. Columns = computed labelling. Routes are classes from A to V. Class Z is none of the routes. Class Y stands for routes not manually labelled. 36
12 Experiments setup summary 40
13 Thresholds for MSI manual labelling 49
14 Summary of results of experiment 2 51
15 Summary - True Positive, False Positive and False Negative Rates 51
16 H0a,b - True Positive, False Positive and False Negative Rates 52
17 H0c,d - True Positive and False Negative Rates 52
18 H1 - Average time between TP prediction and event occurrence. 53
19 H2 - Average time between TP prediction and event occurrence. 53
20 H1 - Average con dence in prediction and detection. 54
21 H2 - Average con dence in prediction and detection. 54
22 Average Situational awareness self assessment. Situational awareness rating cp. [25]: 1-Low, 7-High. 56
23 Control situations - True Positive, False Positive and False Negative Rates 56
24 Control situations - Average time between TP prediction and event occurrence. . 57
25 Control situations - Average con dence in prediction and detection. 57
26 Summary of results of experiment 3 59
27 $\overline{H\,0c;d}$ - True Positive, False Positive and False Negative Rates 59
28 $\overline{H\,0f}$ - Average time between TP prediction and event occurrence. 60
29 $\overline{H\,0i;j}$ - Average con dence in prediction and detection. 60
30 $\overline{H\,0k}$ - Average Situational awareness self assessment. Situational awareness rating cp. [25]: 1-Low, 7-High. 62
31 SG mapping of deliverables to evaluation framework. 85
32 SG v0.7 - Average over [MSI#6,7,12,16] and slow motion for $NARI_{SGv\,0:7}$, $IMISG_{SGv\,0:7}$ and $\overline{IMISG}_{SGv\,0:7}$ (D2.1). 86
33 SG v0.8 - Average over [MSI#6,7,12,16] and slow motion for $NARI_{SGv\,0:8}$ and $IMISG_{SGv\,0:8}$ (D2.3). 87
34 CER mapping of deliverables to evaluation framework. 88
35 CER - Average over [MSI#2,6,8,19,26,28] and [MSI#2,5,6,8,9,19-28]. 89
36 CER - [MSI#2]: Within a given area. 90
37 CER - [MSI#4]: Proximity to other vessels. 91

38 CER - [MSI#5]: In stationary area. 92
39 CER - [MSI#6]: Null speed (stopped). 93
40 CER - [MSI#8]: Mismatch speed area, here high speed near coast (highSpeedNC). 94
41 CER - [MSI#9]: Mismatch speed vessel type (vesselIST). 95
42 CER - [MSI#19]: Under way. 96
43 CER - [MSI#20]: At anchor or moored. Avg. recognition time as minimum of anchored and moored. 97
44 CER - [MSI#21]: Movement ability a ected (maa). 98
45 CER - [MSI#22]: Aground. 99
46 CER - [MSI#23]: Engaged in shing, here trawling. 100
47 CER - [MSI#24]: Tugging. 101
48 CER - [MSI#25]: In SAR operation (inSAR). 102
49 CER - [MSI#26]: Loitering. 103
50 CER - [MSI#27]: Dead in water, drifting (adrift). 104
51 CER - [MSI#28]: Rendez-vous. 105
52 CEF mapping of deliverables to evaluation framework. 106
53 CEF mapping of deliverables to evaluation framework (continued). 107
54 Future Location Prediction (FLP) for short- and long-term prediction. 108
55 Recorded Data Expert 1 110
56 Recorded Data Expert 2 111
57 Recorded Data Expert 3 112
58 Situational awareness data recorded on experiment 2 and 3 for expert 1. Situational awareness rating cp. [25]: 1-Low, 7-High. 114
59 Situational awareness data recorded on experiment 2 and 3 for expert 2. Situational awareness rating cp. [25]: 1-Low, 7-High. 115
60 Situational awareness data recorded on experiment 2 and 3 for expert 3. Situational awareness rating cp. [25]: 1-Low, 7-High. 116

1 Introduction

The introduction delineates purpose and scope of deliverable D5.6, constitutes the relation between the performed activities and the objectives of the datAcron project and lists related deliverables.

1.1 Purpose and Scope

The deliverable D5.6 reports the results of work package 5, as described in Grant Agreement-687591, p.34-35:

This task culminates the WP overall objective performing the experiments and driving conclusions on the measured outcome. Experiments execution, and evaluation of the results will be documented in a nal deliverable. The results of this task will be reported in deliverable D5.6.

This report culminates the overall objective performing the experiments and driving conclusions on the measured outcome, as far as the maritime domain is concerned. Experiment execution, and evaluation of the results will be documented.

In scope are:

- Uni cation of component-speci c big data variations, measurement criteria and measure to a common evaluation framework;
- Capture of work package assessment activities and results;
- Design and execution of experiments on datAcron components with maritime domain experts;
- Measurement and analysis of the maritime experiments results;
- Interpretation and documentation of experimental results.

Out of scope are:

- Variations of big data types which are not processed by at least one datAcron component;
- Assessment of o ine components, as they are not relevant for any maritime scenario;
- Assessment of functionalities that are not implemented;
- Set-up of a lab to mix the developed algorithms and visualisation analytics with real-life systems data. This topic is addressed by D5.5.

1.2 Relating to the objectives of datAcron

Referring to the objectives of the datAcron project, in what follows we report the achievement of Task 5.5:

O.1 Spatio-temporal data integration and management solutions;

3 A spatio-temporal integrated dataset has been published [19].

O.2 Real-time detection and forecasting accuracy of moving entities' trajectories;

3 Results on the detection capabilities were captured, extended by domain-driven evaluations, continuously shared for the improvement of capabilities and documented in this deliverable.

O.3 Real-time recognition and prediction of important events concerning these entities;

3 Results on the recognition and prediction capabilities are captured and documented in this deliverable.

O.4 General visual analytics infrastructure supporting all steps of the analysis through appropriate interactive visualisations;

3 The visual analytics infrastructure with interactive visualisations was part of the experimental setup which is documented in this deliverable.

O.5 Producing streaming data synopses at a high rate of compression.

3 The accuracy of synopses was assessed by experts both for Synopses Generator version 0.7 and 0.8. Results were communicated to the project partners.

1.3 Relation to other deliverables

This deliverable D5.6 relates to the following deliverables in the described way:

- D1.6, D1.7 - Structuring of evaluation is based on the datAcron system and work ow descriptions in D1.6 and D1.7;
- D5.1 - Maritime use case detailed de nition [15]: D5.6 relates to D5.1, which reports the results of Task 5.1, by referring to the initially proposed methodology, the described big data variations and evaluation criteria as well as the scenario descriptions, especially SC11.
- D5.2 - Maritime data preparation and curation (interim) [18]: D5.6 uses the data prepared and curated during the execution of Task 5.2, described in D5.2 and delivered via D5.6;
- D5.3 - Maritime experiments speci cation [4]: D5.6 realises the speci cations developed in Task 5.3 and described in D5.3 with concrete experiments. The experiments are executed and the results are reported in D5.6;
- D5.4 - Maritime data preparation and curation (nal): D5.6 and the experiments executed in the scope of D5.6 build up on the data prepared in Task 5.4 and reported in D5.4;
- D5.5 - Maritime datAcron prototype set-up: D5.6 makes use of the prototype set-up developed in Task 5.4 and reported in D5.5.

Further, this deliverable D5.6 summarises all self-assessment results addressing online components in the maritime domain as reported by other work packages in the following deliverables:

- WP2: D2.1 [8], D2.3 [5],
- WP3: D3.2 [3], D3.4 [26], D3.5 [2].

Contrary, the following deliverables are not taken into account in D5.6. Those are deliverables which either do not address online components (e.g. o ine components), which are not explicitly related to the ful lment of tasks in the maritime domain (e.g. aviation domain), which are not implemented in the datAcron prototype (e.g. A subset of Complex Events in the scenario assessment), or which are none of the three. This concerns explicitly:

- WP1: D1.11.
- WP2: D2.2.
- WP3: D3.1 [16], D3.3 [17].

2 Overview of the evaluation methodology

The evaluation results provided in this document are outcomes of an original methodology sketched at the beginning of the project (deliverable D5.1 [15]), mainly developed in the experimental plan documented in deliverable D5.3 [4], and re ned upon execution with details provided in this current document. We remind the outline while more details can be found in these respective documents.

The methodology is characterised by the four main features:

1. Human-centred design: The expert is used in scenario-level (for awareness) and MSI-level assessments (for accuracy);
2. Uni ed framework for capturing experiments results, linking the big data challenges, the prototype components and the performance criteria;
3. Intertwined types of assessments: The di erent components are assessed both through computational and expert-based methods;
4. Scenario-based data preparation: The data encode speci c operational challenges in maritime surveillance tasks (here collision avoidance) as well as big data challenges.

The goal of the evaluation methodology is to articulate the evaluation activities of the di erent workpackages and to deliver a uni ed vision of the results.

2.1 Human-centric approach

One of the key aspects of the methodology has been to emphasise the role of experts and users in the evaluation of the datAcron prototype. To this end, we rst proposed a list of Maritime Situation Indicators (MSIs) which aimed at representing generic user information needs in a maritime surveillance task, and which acted has an interface between the datAcron scienti c teams and the future users of the datAcron prototype (identifying what the datAcron components should answer). Figure 1 illustrates the idea.

Figure 1: Overview of the user-centred datAcron evaluation design [15]

The MSIs once detected by the di erent datAcron components designed within WPs 1, 2, 3, and 4 will be rendered to the operator through the visualisation component. Put within a speci c operational context de ned by maritime use cases and scenarios, the operator would accomplish a speci c task supported by some situational indicators as detected or predicted by datAcron components. The experimental plan further designed is aligned with this general idea and details the practical implementation.

2.1.1 Aligning datAcron components to maritime surveillance needs

In order to coordinate the development of the di erent datAcron components to t the maritime surveillance application, some aligning was required. This alignment mainly aimed at identifying which scenario (among the one originally proposed in D5.1) would best help demonstrating the datAcron prototype functionalities as well as at mapping user needs (represented by MSIs) to the development of components.

1. Scenario-components mapping: The mapping of components to scenarios with the current status of datAcron implementation, is given in Table 1. The structure of this table was proposed by the project coordinator in order to help the coordination between WPs during the development of the datAcron prototype. The mapping shows that the work packages leaders evaluated scenario 11 (SC11) to involve the largest range of components. Explicitly, SC11 involves SI, CER and IVA online and o ine. In comparison, SC12, SC21, SC31 and SC32 do not involve IVA o ine and SC22 does not involve IVA online. Although the datAcron components are developed to t the purpose of several maritime surveillance uses cases and scenarios, in light of the mapping above as well as the expert based assessment of the scenario relevance (Section 5.1.2), it has been decided to focus on scenario SC11 of collision avoidance, and the dataset for the nal evaluation activities re ects has been prepared accordingly.

2. MSI-components mapping: The mapping between the MSIs and the datAcron components is displayed in Table 2. The mapping was regularly updated in milestone meetings for highlighting the current coverage of the components relatively to the originally proposed list of MSIs. We remind that this list was proposed to drive the development of datAcron components for maritime operational use, but the complete coverage has never been a goal. Still, 6 out of the 28 MSIs have been implemented, among which 2 are not part of the components and 2 are implemented in more than a single components.

WP	Component	SC11	SC12	SC21	SC22	SC31	SC32
WP1							
WP1	SI	Necessity: Y; Next Release:31.01.2018; Version:V0.1; Comment: (a) Results in M25 presentations as well as documented in paper under submission	Necessity: N	See SC11	See SC11	See SC11	See SC11
WP2	SG						
WP2	FLP						
WP3	CER	Necessity: Y; Next Release: 31.01.2018; Version: v0.1; Comment: (a) Description in D3.2,Sect. 3.2.2 (b) Results in D3.2	See SC11	See SC11	See SC11	See SC11	See SC11
WP4	Viz	Necessity: N	Necessity: N	Necessity: N	Integral to "IVA o ine", see below	Necessity: N	Necessity: N
WP4	IVA online	Necessity: Y; Next Release: 02.02.2018; Version: v0.9; Comment: (a) Description in D4.4.1,Sect. 3 (b) Results in D4.4.1	See SC11	See SC11	Necessity: N	See SC11	See SC11
WP1	DM	Necessity: N	Necessity: N	Necessity: N	Necessity: N	Necessity: N	Necessity: N
WP2	TP						
WP2	TDA						
WP3	CER/CEF						
WP4	IVA o ine	Necessity: Y; Next Release: 31.01.2018; Version: V2.0.0-SNAPSHOT; Comment: (a)Description in D1.2,sect.3.6 (b) Results in D4.2.1	Necessity: N	Necessity: N	Necessity: Y; Next Release: 31.01.2018; Version: V2.0.0-SNAPSHOT; Comment: (a)Description in D1.2,Sect.3.6	Necessity: N	Necessity: N

Table 1: Mapping from components to the maritime scenarios [27].

	Maritime Situational Indicator		IMISG	LED	SG	SI	FLP	CER	CEF
Position based MSIs	Close to critical infrastructure	MSI 1	-	-	-	-	-	-	-
	Within a given area	MSI 2	-	1	-	1	-	1	-
	On a maritime route	MSI 3	-	-	-	-	-	-	-
	Proximity to other vessels	MSI 4	-	-	-	1	-	-	-
	In stationary area	MSI 5	-	-	-	-	-	1	-
Speed based MSI	Null speed	MSI 6	-	-	1	-	-	1	-
	Change of speed	MSI 7	-	-	1	-	-	-	-
	Mismatch speed area	MSI 8	-	-	-	-	-	1	-
	Mismatch speed vessel type	MSI 9	-	-	-	-	-	1	-
	Mismatch speed vessel history	MSI 10	-	-	-	-	-	-	-
	Mismatch speed user de ned value	MSI 11	-	-	-	-	-	-	-
Course based MSIs	Change of course	MSI 12	-	-	1	-	-	-	-
	Mismatch course vessel destination	MSI 13	-	-	-	-	-	-	-
	Mismatch course user de ned value	MSI 14	-	-	-	-	-	-	-
Message based MSIs	No AIS reception	MSI 15	-	-	1	-	-	-	-
	AIS reception interrupted	MSI 16	-	-	1	-	-	-	-
	Change in AIS static information	MSI 17	-	-	-	-	-	-	-
	AIS error detection	MSI 18	1	-	-	-	-	-	-
Complex MSIs	Under way	MSI 19	-	-	-	-	-	1	-
	At anchor or moored	MSI 20	-	-	-	-	-	1	-
	Movement ability a ected	MSI 21	-	-	-	-	-	1	-
	Aground	MSI 22	-	-	-	-	-	1	-
	Engaged in shing	MSI 23	-	-	-	-	-	1	-
	Tugging	MSI 24	-	-	-	-	-	1	-
	In SAR operation	MSI 25	-	-	-	-	-	1	-
	Loitering	MSI 26	-	-	-	-	-	1	-
	Dead in water, drifting	MSI 27	-	-	-	-	-	1	-
	Rendez-vous	MSI 28	-	-	-	-	-	1	-

Table 2: Mapping MSIs to datAcron components.

2.1.2 Semantic levels of assessment

The evaluation methodology builds up on the decomposition of the datAcron prototype into three semantic levels of functionalities centred around the MSI concept, proposed in D5.3 and recalled in Figures 2(a), 2(b) and 2(c).

Such a decomposition allows to specify the role of the human in the evaluation: Not involved at under-MSI functionalities assessments, involved as an *expert* at MSI-level assessment and involved as a *user* (or operator) at the scenario-level.

2.2 Capturing assessment results in a uni ed framework

For providing a comprehensive overview of the assessment results from all workpackages, for easing their comparability as well as for allowing the estimate of the response of datAcron prototype to di erent big data variations, a uni ed framework is proposed according to which the delivered evaluation results are sorted. The development of the uni ed framework is based on the prior work reported in D5.1 and D5.3, as well as on the reported results of WP1, 2, 3 and 4.

(a) Assessment of under-MSI functionalities [4]

(b) MSI-level assessment [4]

(c) Scenario-level assessment [4]

Figure 2: Three semantic levels of functionalities and corresponding assessment [4]

2.2.1 The evaluation space

With the dimensions of (1) big data challenges (volume, velocity, variety and veracity), (2) evaluation criteria, and (3) datAcron components, an evaluation space is open in which the results of the experiments have been captured: On the one hand, from the self-assessment of the components performed at the workpackage level (see Section 3) and on the other hand, from the assessments involving experts as reported in Sections 4 and 5. The evaluation space is built upon the di erent kinds of data variation, di erent evaluation criteria and measures used by the di erent component designers. This evaluation space provides a uni ed framework which allows to make the results of assessment of the di erent components (assessed both at the workpackage level and expert-experiments) comparable, from a maritime domain perspective.

As for the input dimension of big data challenges, all evaluation results are distinguished according to the reported data challenges or variations in input to the component:

- Volume: variation of data volume for a given time period
- Velocity: variation of data frequency
- Variety: variation of data source type
- Veracity: variation of data quality

The output dimensions make the evaluation results distinguishable according to the evaluation criteria and measures applied to the output of the datAcron component, which address either a single component of datAcron, either a subset of components or the entire datAcron prototype. Firstly, the evaluation criteria are operationalised by di erent evaluation measures. Secondly, the criteria are clustered into the di erent dimensions of evaluation criteria.

The evaluation criteria and measures can further characterised according to the type of assessment performed (see Section 2.3):

- Computed evaluation criteria and measures without contribution of domain experts or operators, neither in the part of the task ful lment process nor part of the evaluation process. This group of evaluation criteria includes to the evaluation criteria timeliness, scalability, compression rate and classi cation quality in Table 3.
- Domain expert or operator related criteria and measures refer to those criteria and measures that were applied to situations in which one of the two cases is given:

  A domain expert is performing evaluation activities. This refers to the evaluation criterion of classi cation quality which was quanti ed by true and false positive and negative detections of Synopses Generator and Complex Event Recognition.

  An operator is part of the task ful lment. In this case, the performance of datAcron prototype is evaluated via performance criteria quantifying the clarity and the e ectiveness.

2.2.2 System decomposition and dependency analysis

In order to reduce the complexity of the evaluation process and to ensure that the experiments cover a relevant subset of the evaluation space, some system decomposition and dependency analysis has been performed, which are reported in this section.

Additionally, the positive side e ects of this approach are a reduction of evaluation risks imminent to prototype developments, as well as a deeper analysis of single components performance which allows again an understanding of inter-component performance. This is achieved in three steps, illustrated in the rest of the section:

1. Brie y by focusing on the data that are taken into account by the di erent components;
2. Components are evaluated separately, before interaction e ects between the components are evaluated;
3. Di erent levels of data aggregation, i.e. results of inter-components processing, are evaluated separately and successively.

Data complexity reduction datAcron components ingest and process a subset of the broad range of data described in D5.1, D5.2, D5.3 and D5.4, which have a large importance for the successful ful lment of typical tasks of maritime surveillance. Out of this set, the datAcron prototype processes especially:

- Dynamic AIS data (P1).
- Port information (C1).
- Nautical charts (C2).
- Fishing areas (C4).
- Maritime routes (F1).

For the development of experiments, focusing on the relevant data types of each component allows to restrict the number of experiments necessary for a full characterisation of the di erent components.

Component-wise assessment The component-wise decomposition of the datAcron prototype relies on the datAcron architecture as depicted in Figure 4 in Deliverable D1.6. Components are distinguished in online and o ine components.

In the following, datAcron prototype components are called online if they process streamed data, whilst o ine components cannot process streamed data. Online components that were addressed by the evaluation activities are:

- Low-level Event Detection (LED)
- Synopses Generator (SG)
- Complex Event Recognition (CER)
- Future Location Prediction (FLP)
- Interactive Visual Analytics (IVA)

Inter-component assessment Combinations of components are then assessed, until the full datAcron prototype is assessed. In order to assess individual criteria and to avoid double-counting of possible errors, we designed the experiments so that the evaluation space is partitioned properly. In particular, the scenario-level assessment is challenging at it involves all the components: The failure or lack of quality of one of them will impact the result of the following ones, without being able to identify which one fails. For instance, if wrongly detected MSIs are displayed to the operator, it will be impossible to know if the situation has been wrongly assessed by the operator due to the wrong MSI detection or due to the irrelevance of the MSI. Expert-based experiments reported in Section 5 overcome this issue by xing the veracity dimension of the MSIs in experiment 2.

This component decomposition is combined with the preceding decomposition (semantic levels). For instance, for the scenario level assessment, the mapping of components to scenarios is used prior to focusing on online components and the respective data ingested by those.

2.3 Simultaneous computation-based and expert-based assessment

Two main types of assessment were conducted simultaneously:

- Computation-based assessments which do not require expert knowledge, were performed by at the component level by the components designers themselves and were reported in di erent deliverables under WP1, 2, 3 and 4 (referred below as self-assessment), and
- Expert-based assessments which are independent assessments of the components performance involving maritime surveillance experts. Expert-based assessment was performed into two phases or periods:

  Period 1, during the execution of the project: components' results on the reference dataset were received and analysed, while the feedback was provided to the WP leaders for further improving their solutions;

  Period 2, during the weeks of experiments in March 2018, and mostly in October and November 2018.

2.3.1 Connection between the two types of assessment

Both the under-MSI and the MSI-level assessment results are reported from the self-assessment of the components designers and from expert-based assessments through maritime domain experiments conducted by WP5. The scenario-level assessment extends the scope of evaluation from a single output assessment, e.g. the assessment of one type of critical point or one MSI, to an evaluation of multiple MSIs and under-MSI products simultaneously used by an expert user ful lling a typical maritime domain task, i.e. collision avoidance. The scenario-level assessment is a pivot for the maritime domain evaluation as it creates an evaluation context which makes use of all under-MSI and MSI functionalities and enables a qualitative (and sometimes quantitative) assessment of the impact of the outputs from the under-MSI and MSI quality in an operational context (See especially Section 5.5).

The self-assessment by components designers of workpackages 1, 2, 3, 4 and the expert-based evaluation activities of workpackage 5 were interconnected by the following mechanism:

Firstly, the expert-based evaluation activities of WP5 were focusing on processes which are speci c for the maritime domain, thus which are only executable with domain knowledge. All assessments not requiring domain knowledge were primarily treated by the respective workpackage.

Secondly, the evaluation activities followed the order of completion (and availability) of the di erent components. The sequential development and evaluation of components of the datAcron prototype have been conducted in parallel for e ciency reasons. This means that the evaluation on upstream components started before the development of the downstream components was necessarily concluded. As depicted in Figure 5 of D1.11, the Low-level Event Detection (LED), for instance, is upstream from the Synopses Generator (SG) and the Semantic Integrator (SI) is downstream from both of them.

Thirdly, the evaluation activities followed the path with the largest potential for improvement. By these intermediate evaluations, the development of components was supported: For instance, after analysing CER results, it was decided that an evaluation of the SG was required, as the substandard results of CER were bred by SG. For developments resumed after the completion of the evaluation, the results of these evaluations constitute a contribution to the development of the respective component. If the results are outdated, all intermediate evaluation results are quoted for documenting the contribution to the development process and for the quanti cation of performance improvements.

2.3.2 Expert-based assessment work ow

Figure 3 summarises the work ow of the assessment involving experts that will be detailed in Sections 4 and 5. The reference dataset provided by NARI [20, 22] and brie y described in Section 2.4), either as whole or parts of it, has been processed by the di erent datAcron components (LED , SG, SI, and CER mainly), and the resulting output data have been evaluated at two levels, called respectivelyMSI level (Figure 3, top-right) and Scenario level (Figure 3, bottom-right). The results of these evaluations are reported in Section 4 and Section 5, respectively.

Figure 3: Maritime data work ow and expert-driven assessment principle

As illustrated in Figure 3, in both cases the assessment involved maritime experts, respectively considering the following research questions:

- Do the detected MSIs match the human expectation in terms ofaccuracy/veracity? (MSI level)
- Can the MSIs help the expert solve the problem targeted by the scenario (e.g., identify, understand, prevent a given maritime situation in a live experiment?) (Scenario level)

Some low-level MSIs concern spatio-temporal queries results, eventually integrated with semantics (as for SI output) or trajectory synopses annotations (e.g., MSIs #2, #6, #8, #9, #10, #11, #12, #15, #16, #17), while high-level MSIs correspond to more complex events of interest (e.g., MSIs #23, #25, #26, #28). Low-level MSIs expert's evaluation has been mainly done on the reference dataset (sampling the data and focusing either on one day, one week or one month, because these simple events are numerous). For high-level MSIs and for scenario level assessment, the reference dataset has been enriched with reference information obtained via data modi cation methods (degradation, enrichment), targeted for speci c experimental purposes. The details of the dataset preparation are reported in Deliverable D5.5 [12] and summarised in the next section.

2.4 Reference datasets for evaluation

Working with real data is essential for the credibility of processing and results. However, such data often comes with lots of intrinsic veracity issues. An accurate assessment of the datAcron

prototype and results can only be experimented on controlled datasets. In the context of task 5.2, we designed and prepared an adapted maritime dataset for datAcron partners [23]. This maritime dataset is composed of two parts:

1. A reference (batch) dataset, limited in volume but composed of a large variety of maritime data;
2. A real-time stream of AIS messages with volume and high velocity.

2.4.1 Two AIS data sources for big data challenges

The maritime dataset and the stream have been prepared using two dynamic AIS sources, both were considered for exercising the algorithms and validating the prototype:

1. IMISG AIS data, which covers all European coasts, is a basis for the stream and the reference dataset;
2. NARI AIS data, which covers western coasts of France, including parts of Celtic sea, North Atlantic ocean, English Channel and Bay of Biscay, is a basis for the reference dataset.

The IMISG data has high volume, high velocity and unknown veracity, while the NARI, which focus on a small area, has low volume and velocity, high variety and masterable veracity. Accordingly, the two sets of data were used for di erent purposes as far as the big data challenges are concerned:

- Volume and velocity challenges were addressed mainly using the IMISG data (results are reported mainly in Section 3),
- Variety and veracity challenges were addressed mainly using the NARI data (results are reported in Sections 4 and 5).

2.4.2 Data preparation for expert-based experiments

In order to setup the datAcron prototype for supporting a maritime surveillance task, and to e ectively assess the quality of the results provided by the datAcron components (e.g. synopses generation, event detection and forecasting) at the MSI and scenario levels [4], we prepared a series of datasets including the aforementioned reference dataset.

1. Reference dataset (6 months of AIS data complemented by contextual data covering western coasts of France, including parts of Celtic sea, North Atlantic ocean, English Channel and Bay of Biscay) Provided to the partners to exercise datAcron components; Details can be found in the deliverable 5.5 [12]. A shareable version of this dataset has been published [19, 21].
2. Datasets for scenario-level evaluation (Subsets of the reference dataset, enriched with speci c events and annotated by experts) Designed for the prototype setup, to exemplify situations of interest for a collision prediction task;
3. Datasets for veracity variations (a series of degraded versions of a subset of the reference dataset) Used to test the robustness of some components, in particular theCER as reported in Section 5.7.

The preparation of the datasets consist into three steps: (1) excerpt of the full reference dataset of a spatio-temporal area of interest, (2) enrichment with synthetic or shifted events, (3) expert-annotation of events (real and injected). Reference or ground truth information is generally di cult to obtain in large scale moving object data streams and databases (e.g., European scale). Indeed, the annotation of real moving object datasets with MSI is extremely challenging as it is time consuming, the proper annotation tools do not exist yet, and the ground truth

highly depends on the operator task (contrary to more objective labels such as vessel type).

Therefore, the reference dataset has been enriched with synthetic or real trajectories to reflect specific situations. For synthetic trajectories, additional information describing intrinsic quality of the reference dataset has been first extracted and further used to generate small datasets (raw or synopses/MSI level) exhibiting typical vessel behaviours, aligned with the scenario definitions [15] and the experimental plan [4] proposed. Datasets created for experimental validation are thus either purely synthetic and automatically generated based on some motion models, or pseudo-synthetic by modifying existing real data with a controlled process. In the first case they may be biased by the model applied, while in the second case they preserve some characteristics of the original data. In both cases, the dataset was quite conform to real vessel motions and maritime events. This approach aimed thus at providing simulated behaviours credible to operators.

This methodology designed for the evaluation of maritime situations with experts has been first rehearsed in March 2018 with WP4 partners. The resulting prototype setup and the outcomes of this rehearsal session based on collision scenarios are reported in the deliverable D5.5 [12].

Several situations have thus been created and integrated as part of the different scenarios. The situation presented to the operator during the experiment considered a set of real AIS data enriched with specific events either simulated or shifted in time and/or space from real data. These events 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 (e.g. the individuation of fishing patterns);
- The simulation of a rendez-vous behaviour.

In order to modify the original dataset, several techniques have been developed to reflect in particular the well documented lack of veracity in AIS data. Indeed, AIS data are not perfect as errors, falsifications and spoofing cases [24], inaccuracy or incompleteness with missing data fields [13] 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 along to the corresponding dimensions of veracity.

The proposed methods can be classified into four families: (1) noise addition, (2) data modification, (3) data removal and (4) data addition, and were presented in [23]. These data degradation functions have been designed and implemented to degrade a given dataset automatically.

The remaining of the document reports the evaluation results of the datAcron prototype: In Section 3, results from computational-based self-assessment of the different component designer are captured and summarised. In Section 4, the accuracy of the MSIs is assessed by expert and the methodology followed is briefly sketched. Section 5 reports the results of experiments involving maritime surveillance experts put in operational context of use on the collision avoidance scenario. Finally, the robustness of the CER component to successive degradation of data veracity (missing data) is studied in Section 5.7. This last experiment does not involve experts.

3 Capturing component assessment results

This section summarises results stemming from evaluations conducted by all WP in charge of the development of the respective component, captured in the evaluation space introduced in Section 2.2. In Table 3, the columns list the four big data challenges of variety, veracity, volume and velocity. The rows show the di erent groups of evaluation criteria with evaluation measure in brackets. Each cell of the table contains the components and the deliverables in which the self-assessments of these components address the respective big data variation and evaluation measure. For instance, the measure compression rate is used for evaluating the performance of the Synopses Generator under volume and velocity variations whose results are reported in D2.1 and D2.3. In the following as well as in Section 8, the self-reported results are captured in accordance with the uni ed evaluation framework and concluded from a maritime domain perspective.

Criterion (Measure)	Variety	Veracity	Volume	Velocity
Timeliness (Latency), Scalability	CEF(D3.5)		SG(D2.1,D2.3) CER(D3.2,D3.4) CEF(D3.5)	SG(D2.1,D2.3) CEF(D3.5)
Compression rate			SG(D2.1,D2.3)	SG(D2.1,D2.3)
Classi cation quality (Accuracy, F1)	CER(D3.4) CEF(D3.5)	CER(D3.4) FLP(D2.4)	SG(D2.1,D2.3,D3.4) OTC(D2.5) CEF(D3.5)	SG(D2.1,D2.3,D3.4) FLP(D2.3,D2.4) CEF(D3.5)
Clarity (Con dence) E ectiveness	IVA, VIZ (D5.5, D5.6)			

Table 3: Summary of WP self-assessment of components: Synopses Generator (SG), Future Location Predictor (FLP), Complex Event Recognition (CER), Complex Event Forecasting (CEF), Interactive Visual Analytics (IVA), Real-time Visualisation (VIZ).

3.1 Synopses Generator (SG)

Synopses Generator has been reported in Deliverables D2.1 [8] and D2.3 [5]. The experimental setup and the respective results are described in pages D2.1 [8], p.59, D2.3 [5], p.85. Two versions of SG are distinguished, SG v0.7 and v0.8.

MSIs investigated by the component evaluation

[MSI#6] Null speed (Stop)

[MSI#7] Change of speed (change in speed)

[MSI#12] Change of course (change in heading)

[MSI#16] AIS reception interrupted (Gap)

In addition to these prede ned MSI, D2.1 [8] and D2.3 [5] address the development and evaluation of indicators for slow motion events.

Data variations

Volume and velocity variations:

- NARI $_{SGv\,0:7}$: 18.495.677 messages of 5.055 vessels with an average reporting rate of one message per 1061 seconds. The dataset is collected in Brest are during 6 month.
- NARI $_{SGv\,0:8}$: 19.035.630 messages of 5.055 vessels with an average reporting rate of one message per 1061 seconds. The dataset is collected in Brest are during 6 month.
- IMISG $_{SGv\,0:7}$: 61.187.265 messages of 118.003 vessels with an average reporting rate of one message per 1215 seconds. The dataset is collected on European scale during 1 month, especially the Mediterranean sea.
- IMISG $_{SGv\,0:8}$: 60.645.849 messages of 118.003 vessels with an average reporting rate of one message per 1215 seconds. The dataset is collected on European scale during 1 month, especially the Mediterranean sea.
- $\overline{\text{IMISG}}_{SGv\,0:7}$: 41.466.539 messages of 18.034 vessels with an average reporting rate of one message per 5 seconds. The dataset is a 4 hours subset of IMISG $_{SGv\,0:7}$ extended by the addition of synthetic messages, calculated by interpolating between existing messages.

Algorithms' parameters variations

4 : (2.5, 5, 7.7, 10 degrees). The value of the parameter impacts on the detection of change in heading events. It is the threshold that needs to be exceeded between two raw AIS messages.

4 T: (10,15,30,60 minutes). Value of the threshold for detecting gap events#threads: (1,2,4,8 threads). Number of parallel processing units, threads or nodes.

Performance Criteria - Measures - Results

- Compression ratio: Relation between removed AIS messages, i.e. messages that are not classi ed as critical points, and all AIS messages. Both quantities are measured by the amount of messages, yielding the compression ration in percentage, [5], p.69.
- Timeliness, measured in latency. Latency corresponds to the time in SG pipeline, measured in milliseconds [5], p.86.
- Timeliness, measured in throughput, which corresponds to the number of messages processed per second [5], p.86.
- Scalability: Measured in change of latency with respect to the change of the number of parallel processing units.
- Accuracy: Measured in Root mean squared error (RMSE), based on Haversian distance, measured in meter, including critical points.

Description of experiments

Table 31 summarises the performed experiments characterised by the application of di erent data variations, referring to the deliverables in which the results are reported with respect to the respective measurement criteria and measures.

Conclusions on SG The evaluation of SG is performed on two versions, v0.7 and v0.8. Comparing the performance of these two versions in Table 32 and Table 33, di erent conclusions appear to be pertinent. Firstly, the accuracy of v0.7 is Pareto e cient compared to the accuracy of v0.8, measured in root mean square error (RMSE). Since the datasets used for the evaluations of v0.7 and v0.8 are very similar but not identical, the di erences need to be considered as possible factor which limits the comparability of the results. Secondly, the compression ratio is not correlated with this e ect. Both versions are capable of compressing the volume of the initial AIS data with a similar rate. SG v0.7 reaches higher compression rates than v0.8 in 6 out of 14 experiments using NARI and IMISG datasets. Vice versa,SG v0.8 reaches higher compression rates than v0.7 in 5 cases. The di erences are small, given possible di erences in the datasets which were used for the experiments. Nonetheless, the results imply that the compression capability of SG is similar for SG v0.7 and v0.8. Latency and throughput are comparable between v0.7 and v0.8 only for the single thread case, as the collected data is complementary for other cases. None the less, a signi cant reduction of latency is observable from v0.7 to v0.8 for the NARI dataset from ca. 238 to ca. 116 ms and for the IMISG dataset from ca. 2785 to ca. 923 ms. The throughput is reduced from v0.7 to v0.8; For the NARI dataset from ca. 16900 to ca. 8545 messages/second and for the IMISG dataset from ca. 19485 to ca. 11455 messages/second. Given that latency is of greater importance for scalable systems than throughput, the implemented changes from v0.7 to v0.8 are improving the performance of SG signi cantly . Comparing the reduction of latency from single thread to eight thread con guration, the improvement of performance is the lowest for v0.8 on NARI with ca. 86%, followed by v0.7 on IMISG interpolated with ca. 89% and outperformed by v0.8 on IMISG with ca. 96%. No conclusion can be drawn on the improvement of v0.8 with respect to v0.7. Instead,this indicates that the scalability depends on the dataset .

The impact of delta Theta is both in NARI and in IMISG the most substantial on the number of detected [MSI#12] - change in heading. The larger delta Theta, the smaller the number of retained critical points. All other analysed MSIs remain largely una ected. Both results are intuitive, as [MSI#12] - change in heading is functionally dependent on delta theta.

Contrary, the impact of delta T is di erent from NARI to IMISG. While the compression ratio remains virtually una ected in NARI, both in average and MSI wise, the compression ratio in IMISG increases signi cantly with larger delta T. This nding is not surprising for [MSI#16] - AIS reception interrupted, given that the average AIS transmission frequency in IMISG is lower than in NARI. Independently and more interesting, a decrease of the number of detections for [MSI#12] - change in heading becomes obvious for larger delta T, as shown in D2.1, Figure 22, bottom-right, p.68 and D2.3, p.90. This implies, that some [MSI#12] - changes in heading are not detected, if the threshold of another MSI, here delta T for [MSI#16] - AIS reception interrupted is changed. This behaviour is both not intuitive and supposedly undesirable .

3.2 Complex Event Recognition (CER)

CER has been reported in Deliverables D3.4 (nal) and D3.2 (interim); See Table 34. The experimental setup and the respective results are described in D3.4 [26], p.28 and in D3.2 [3], p.10.

Three versions ofCER di ering both in the input and in the pattern used for the recognition. Each version is assumed to be implemented both in an o ine and an online fashion:

- Version 1: Raw AIS data as input. The raw AIS data is not preprocessed in a Synopses Generator -fashion, but the de nition of the event recognition predicate is di erent from Version 2. This version is described in D3.2.

- Version 2: Only trajectory synopses as input. This version is described in D3.2 and D3.4.
- Version 3: Enriched AIS data as input. In this version, the raw AIS data are enriched by trajectory synopses. This version is described in D3.4.

MSIs investigated by the component evaluation

[MSI#2] Within a given area

[MSI#5] In stationary area

[MSI#6] Null speed

[MSI#8] Mismatch speed area

[MSI#9] Mismatch speed vessel type

[MSI#19] Under way

[MSI#20] At anchor or moored

[MSI#21] Movement ability a ected

[MSI#22] Aground

[MSI#23] Engaged in shing

[MSI#24] Tugging

[MSI#25] In Search And Rescue (SAR) operation

[MSI#26] Loitering

[MSI#27] Dead in water, drifting

[MSI#28] Rendez-vous

In addition to these prede ned MSI, D3.2 [3] addresses the development and evaluation of indicators for low speed events. D3.4 [26] addresses additionally indicators for communication gap, low speed, changing speed, in area of interest and travelling speed events.

Data variations

Variety variations implying Veracity variations:

- Enriched AIS stream (AIS data enriched by critical point stream)
- Critical point stream

Volume variations:

- NARI $_0$: Atlantic Ocean around the port of Brest, France, from 1 October 2015 to 31 March 2016 including 4.142.448 critical Simple Derived Events (SDEs) and 1.851.265 spatial SDEs, and concerns 4.953 vessels and 6.894 areas of interest.
- Greakseas: 5.166 vessels sailing in the Greek seas in January 2016 including 7.834 areas of interest, while the dataset includes 1.181.044 critical SDEs and 176.070 spatial SDEs.
- NARI $_1$: Critical point stream of NARI + shing areas + Natura 2000 areas + anchorage areas + near coast areas + shallow areas (with average number of input entities x 1000):
  - window size 2h (ca. 7 AIS entities)
  - window size 4h (ca. 15 AIS entities)
  - window size 8h (ca. 30 AIS entities)
  - window size 16h (ca. 60 AIS entities)
- NARI $_2$: Enriched stream of NARI (including NARI $_0$ and NARI $_1$+ shing areas + Natura 2000 areas + anchorage areas + near coast areas + shallow areas (with average number of input entities x 1000):
  - window size 2h (ca. 25 AIS entities)
  - window size 4h (ca. 40 AIS entities)
  - window size 8h (ca. 70 AIS entities)
  - window size 16h (ca. 145 AIS entities)
- IMISG $_1$: Critical point stream of IMISG + shing areas + Natura 2000 areas (with average number of input entities x 1000):
  - window size 2h (ca. 200 AIS entities)
  - window size 4h (ca. 400 AIS entities)
  - window size 8h (ca. 800 AIS entities)
  - window size 16h (ca. 1500 AIS entities)
- IMISG $_2$: Enriched stream of IMISG (including IMISG AIS data stream and IMISG $_1$) + shing areas + Natura 2000 areas (with average number of input entities x 1000):
  - window size 2h (ca. 450 AIS entities)
  - window size 4h (ca. 900 AIS entities)
  - window size 8h (ca. 1700 AIS entities)
  - window size 16h (ca. 3100 AIS entities)

Algorithms' parameters variations

Window size: (2, 4, 8, 16, 24 hours). The value of the parameter impacts on the data volume processed at once, thus it's change creates for both NARI and IMISG dataset additional levels of data volume.

Window size: (1,2,4,8 cores). The number of cores utilised for processing.

Performance Criteria - Measures - Results

- Timeliness: Average recognition time (sec)
- Classi cation quality: (assumption: enriched stream corresponds to ground truth): Precision, Recall, F1.

Description of experiments

Table 34 summarises the performed experiments characterised by the application of di erent data variations, referring to the deliverables in which the results are reported with respect to the respective measurement criteria and measures.

Conclusions on Complex Event Recognition (CER) For the CER, di erent conclusions can be drawn with respect to timeliness and accuracy measures. With respect to timeliness, the average recognition time of ca. 1.8 sec for the largest window size lets assume that an application of CER in an operational environment is suitable for time critical tasks for datasets with similar characteristics as NARI $_1$. For datasets with higher volume or velocity like IMISG $_1$, the application to tasks with observational character without need of interaction with the vessels seems to be more appropriate .

With respect to accuracy measures, the CER is stressed by variations of veracity, here performed by neglecting the subset of AIS messages which are not identi ed to be critical points by SG.

- CER yields the same results on raw AIS data plus critical points and critical points only for the given datasets for [MSI#6] - Null speed, [MSI#19] - Under way, [MSI#20] - At anchor or moored, [MSI#22] - Aground, [MSI#26] - Loitering and [MSI#28] - Rendezvous. [MSI#23] - Engaged in shing and [MSI#25] - In SAR operation with over 99% for recall and precision reaches a large overlap of results calculated on raw AIS data and critical points.
- Lower or varying precision scores: The performance of [MSI#27] - Dead in water, drifting varies importantly with respect to the ingested dataset. While recall moves around 90-95%, precision varies between 40-85% a ecting the spread of F1 scores.
- Lower or varying recall scores: [MSI#24] - Tugging yields for both datasets a recall of 88%, for [MSI#21] - Movement ability a ected the recall varies for the datasets between 71-99%. [MSI#8] - Mismatch speed area achieves 91% of recall for NARI $_1$.
- Lower or varying recall and precision scores: For [MSI#9] - Mismatch speed vessel type both recall and precision vary similarly between 94 and 97% with lower values for NARI $_1$ and larger values for IMISG $_1$.

3.3 Complex Event Forecasting (CEF)

Relevant Deliverables for CEF are D3.5 (nal), D3.2 (interim); See Table 52 and 53. The experimental setup and the respective results are described in D3.5 [2], p.18 and in D3.2 [3], p.43.

The CEF forecast events are primarily based on critical points. For the forecasting of the movement pattern, the critical points are enriched with heading information [3], p.43.

MSIs investigated by the component evaluation

[MSI#] - .

The CEF o ers in D3.2 [3] and D3.5 [2] the development and evaluation of forecasting indicators additional to the prede ned MSIs. These include exemplary implementations allowing for answering the following two questions:

- Which vessel completes a speci c movement pattern? Two patterns are investigated [3]:

  Turn Communication Gap Turn.

  TurnNorth TurnWest or TurnEast TurnSouth.

- Which vessel is approaching port? (starting from 10 km distance) [2]
- When is vessel about to start shing? [2]

Data variations

Volume variations:

- NARI $_1$: Critical point stream of NARI $_0$: The rst was provided by NARI and contains AIS kinematic messages from vessels sailing in the Atlantic Ocean around the port of Brest, Brittany, France and span a period from 1 October 2015 to 31 March 2016. [2], p.18.
- IMISG $_1$: Critical point stream of IMISG : The second was provided by IMISG and contains AIS kinematic messages from vessels sailing in the entire Mediterranean Sea, as well as in part of the Atlantic Ocean and span an one month period from 1 January 2016 to 31 January 2016. [2], p.18.
- Greakseas$_2$: ca. 6500 vessels sailing in the Greek seas over 3 month.

Algorithms' parameters variations

Features - additional detection features, here using speed and heading information (yes, no). [2]

con dence threshold: (0.1, 0.3, 0.5, 0.7, 0.9). [2]

m - Markov chain order: (0,1,2,4). [2]

m - Markov chain order: (0,1,2). [3]

prediction threshold: (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) [3].

Performance Criteria - Measures - Results

- Timeliness: Execution time (sec) [2].
- Classi cation quality: Precision, spread (min), distance (min) [3], [2].

Description of experiments

Table 52 summarizes the performed experiments characterized by the application of di erent data variations, referring to the deliverables in which the results are reported with respect to the respective measurement criteria and measures.

Conclusions on Complex Event Forecasting (CEF) The CEF is used to forecast which vessel is approaching a port and which vessel will start shing in the near future. For both events to be predicted, a strong positive correlation between the algorithm parameter con dence threshold on the one hand and precision (good when high), spread (good when low) and distance[1] (good when high) on the other hand become visible for NARI dataset. For both pattern an increase of the con dence value from 0,7 to 0,9 increases the spread and the distance more importantly than the precision. The selection of an optimal value is assumed to be task

[1] measuring the di erence in time between the forecasted start of the event and the current point in time

dependent or depending on the preferences of the operator, e.g. safety related task are more likely to bene t from a larger distance. In IMISG dataset the correlation between the con dence threshold and the performance measures precision, spread and distance cannot be found for the port approach forecast. Instead, it becomes more apparent, that spread and distance are positively correlated. This allows for the selection of a high con dence value which yields a high precision score. Depending on the optimality criteria of the task or the operator a high distance or a low spread can be achieved by di erent values for m, the order of the Pattern Markov Chain.

The execution time exceeds 80 seconds for NARI dataset and 180 seconds for IMISG dataset. Compared to the other datAcron components, these latency values are relatively high, e.g. the latency of Synopses Generator v0.8 for IMISG remains below 1 second. Trading o the execution time of e.g. 80 seconds for datasets similar to NARI against a forecasting distance of more than 380 minutes results in a forecasting horizon for shing of more than 378 minutes. This makes the use ofCEF supposedly also bene cial in online scenarios like collision avoidance. For IMISG dataset this e ect diminishes, given that the execution time rises to ca. 200 seconds, while the time between event forecast and event occurrence diminishes to 5 to 15 minutes, but returns guarantees average forecasting horizons, which are twice as large, as the execution time. A possible hindrance for the application ofCEF in online scenarios is the fact, that the spread of the event forecast is typically twice as large, as the time from the event forecast to the event occurrence.

3.4 Future Location Predictor (FLP)

The development and evaluation of FLP is described in D2.4 [9].

MSIs investigated by the component evaluation

[MSI#] - .

The future location predictor o ers additional functionalities with respect to the prede ned MSIs. These include the following task:

- Future location prediction [9]

Data variations

Veracity variations:

- NARI $_0$: raw data.
- NARI $_1$: critical point stream of NARI $_0$.

Algorithms' parameters variations

short term prediction horizon: 10s, 20s, 40s, 1min20s, 2min40s, 5min [9]long term prediction horizon: 1h30min, 3h, 6h, 12h [9]

Performance Criteria - Measures - Results

- Accuracy (Median-, Average-, Maximum-RMSE) [9].

Description of experiments

Table 54 summarises the data variations and the results of the performed experiments with respect to the listed measurement criteria and measures. Additional experiments on the timeliness of training and testing of FLP with measured by latency and throughput. As the performance for training and testing of FLP is supposed to di er from the performance during the operational use, the results can be looked up in D2.4 [9], p.63.

Conclusions on Future Location Prediction (FLP) The short- and long-term future location prediction is assumed to be computational expensive and only available for a subset of vessels for which accuracy results are given in table 54. Typically, two vessels are involved in collisions or close-quarter situations. Thus, assuming that the future locations of two vessels, which are supposed to collide in the future, are predicted and assuming that these location predictions are later used for triggering an estimated collision time, as proposed in D5.3 [4], p.23, then the bene t of the FLP depends on the characteristics of the two vessels, their actual manoeuvre and the environmental conditions. As the median accuracy of the long-term FLP is too small and the prediction horizon is too large, an application in online scenarios like collision avoidance is not indicated. The short-term FLP reaches lower median RMSE values. Assuming the largest prediction horizon evaluated, i.e. 5min, and the inaccuracy of the FLP having only a positive e ect on the prediction of the time to collision, i.e. estimating the collision always too early and never too late, the following vessels are potentially bene ting from being hailed, assuming an immediate crash-stop: refrigerated cargo vessels (<4min), container vessels with maximum speed larger than 25 knots (1min40sec - 5min), cargo liners with maximum speed larger than 18 knots (5min) [10]. For situations in which one or both vessels have a longer stopping time, the 5min prediction horizon is assumed to be too small, as the predicted time to collision is smaller than the necessary time to reduce the speed to zero. This includes: Tankers (>8min), Bulk carriers (>5min), container vessels with maximum speed lower 25 knots (>5min), cargo liners with maximum speed lower 18 knots (>5min) [10].

4 Expert-based assessment at the MSI-level

This section reports on the expert-based evaluation organised in support to the development and validation of Maritime Situational Indicators. The objective is the evaluation of MSIs detection independently of any use case or scenario. This assessment concerns synopses produced by the component SG[2] and events produced by the component CER. The evaluations have been conducted during two periods. During the rst period (from month 16 to month 25), SG and CER outputs have been assessed by the expert on a regular basis in order to improve results of components by providing regular feedback to the components designers. During the second period (during month 34), the focus has been made on CER outputs for the validation of MSIs to be presented to experts during the nal experiment organised at the beginning of month 35 (cf. Section 5).

Below we sketch the methodology followed by the maritime expert that is aimed to be a generic enough to gather both automatic and human assessment. Then the two periods of expert-based MSIs assessment are explained.

4.1 Methodology for the assessment of MSIs

Assessment of MSIs may be evaluated comparing the results with reference or baseline assessments, with known values. To assess in particular the *accuracy*[3] of event detection and prediction algorithms, this reference is often given by or derived from ground truth events.

4.1.1 Expert-based assessment

As mentioned previously, labelled datasets with ground truth information representing extensive and realistic use cases are challenging to obtain. The evaluation is usually undertaken by subject matter experts, who, either manually or assisted by ad-hoc software, check the results and evaluate them according to their own expertise. They thus identify the cases where the algorithm correctly detects an event of interest, and the cases where it fails, distinguishing *true positive* detections from *false negative* ones. The expert may also identify the cases in which the algorithm does not detect any event of interest while there is none, the so called *true negative* detections, from the cases in which it incorrectly detects events, which are called *false positives*. When the expert assesses all the available results, a measure of accuracy of the method may be calculated, and the results with expert annotation may be used for tuning the software solutions. If the number of events is too large for a fully manual assessment, as it is usually the case, a representative sample of events may be evaluated instead, including multiple settings of software use such as close to the coast , in open-sea , or considering di erent periods of time. The approach followed by the expert may be eventually formalised, driving the development of software solutions to assist, or run automatically, user driven evaluation.

The approach to be described below follows the general idea above. On a couple of examples, we describe and analyse the expert-driven assessment of the accuracy of event detection software (SG and CER) undertaken. The expert who achieved the assessment is a former operational expert from the French Navy (NARI). The expert spent 20 years (1997 to 2016) in submarine forces, as a specialist in acoustic recognition. During this period, the expert navigated for fteen

[2] Synopses Generator v0.5, v0.6, v0.7, v0.8 have been evaluated.

[3] The accuracy is here the measure of quality of the algorithm based on true positive and false positive detections. Note that accuracy assessment di ers from usability assessment, the later evaluating the ability of human operators to interact correctly and pro tably with the software.

years on submarines (SSN, SSBN) and frigates. The expert is currently research engineer in signal processing (acoustic) and data analysis (maritime mobilities) at French Naval Academy. In the last years, he developed the expertise in recognising shing vessel behaviours and mobility patterns, especially in Brest (FR) area where the scenarios analysed below take place.

4.1.2 Method followed by the expert

The expert was given the reference AIS dataset in which events of interest (MSIs) representing vessels status were included, especially:

1) Stop, where a vessel stops (MSI #6);
2) Underway, which describes a vessel which is moving, or sailing (MSI #19);
3) High speed, describing a which speed is above the cruise speed, i.e. above a threshold in a given area (MSI #7-11).

These events were detected by either SG, CER or both. The expert's task was to evaluate the accuracy of these components. To accomplish the task, the expert had at his disposal also the raw AIS data stream used as software input, plus many geographical features (electronic nautical charts), environmental data (e.g., sea state) and contextual information (e.g., vessel register). Furthermore, the expert knows the maritime region and typical ships behaviours in the given area.

The expert used an ad-hoc combination of di erent software to support his analysis[4]:

- A storage system enhanced with spatial capabilities to store the datasets and do some basic analytics and spatial operations (e.g., the object-relational database PostgreSQL, including its spatial extension PostGIS);
- A Geographical Information System (GIS) for data visualisation and spatial analysis and ltering (e.g., the desktop GIS QGIS);
- Scripting languages for data analysis (e.g., Python and Matlab).

In the approach used by the expert, the support for data visualisation, in this case provided by the GIS software, was fundamental. The expert performed the following steps:

1. Import raw surveillance data (AIS), processed data (AIS data annotated with events), and supporting datasets, including AIS status codes for vessel types, vessel list, shing vessel list, protected areas datasets, port database, weather conditions) in the database;
2. Develop scripts to elaborate surveillance raw data and produce a baseline dataset, for comparison, and import it in the database;
3. Convert the data in the database in order to enable the spatial representation of the spatial features (AIS coordinates, vessel trajectories);
4. Use the integrated database spatial capabilities to query relevant subsamples of the baseline and processed datasets;
5. Visualise the baseline and processed datasets in the GIS software. Use GIS analysis capabilities (ltering, spatial overlay) to compare the two datasets and highlight inconsistencies;
6. Eventually, develop ad-hoc scripting to calculate statistics on relevant features of the two datasets (e.g., speed), to be analysed and compared.

[4] We report this software here as an exempli cation of the approach rather than as an expert recommendation

		Synopses Generator	
		Stop	non-Stop
Expert judgement	Stopping area	TP: Red dots in harbour in Fig. 4	FN:
	Non-Stopping area	FP: Red dots in TSS and straight trajectories in Fig. 4	TN: blue dots in TSS and straight trajectories in Fig. 4

Table 8: Qualitative confusion matrix of the expert driven evaluation of SG results for Stop event annotations, referring to Figure 4. True Positive (TP), False Positives (FP), True Negative (TN), False Negative (FN).

In the following, we describe and analyse the evaluation undertaken by the expert for evaluating two pieces of software for detecting the three MSIs described above. For the rst MSI, stop, he evaluated two di erent components: SG and CER. CER was originally used in cascade to SG, i.e., SG output were used as CER input; therefore, beside result accuracy, it was important to evaluate how CER results were a ected by SG ones. This is not unusual for stop indicator, because it is a basic event for the discovery of complex events (e.g. rendez-vous events).

In the case of underway events the expert evaluated the results directly produced by CER processing SG output, which he compared with the baselines datasets he produced.

Finally, he analysed CER high speed detections. Since CER was used in cascade to SG (i.e., using SG results as input data), he computed some statistics of SG output to assess the quality of CER's input dataset, to qualitatively evaluate how it a ected the quality of the detection.

4.2 MSI assessment, rst period

The rst period of SG and CER outputs assessment has been done during the M16-M25 period (prior to the experiments). The objective of this assessment was to assist WP2 and WP3 in the design of synopses and events detectors through iterative evaluations.

4.2.1 Assessment of Stop Event Detections

In order to evaluate the accuracy of stop events detected by SG, the expert focused on areas where the probability of having stopped vessels is usually very low, such as maritime routes, or along a vessel trajectory, providing the basis for false positives assessment. Following the procedure we described in previous section (speci cally, points 1, 3-5 in the methodology), he visualised the results in the GIS software and identi ed visually the false positive detections. In Figure 4, the expert highlighted the false positive annotations in an area in the Channel between Brest (FR) and UK of the Ushant Tra c Separation Scheme (TSS) where, unless a vessel is in distress, it is very unlikely to stop.

In Figure 4, the red dots represent stop events (start of stop events, or end of stop events) detected by SG. The expert highlighted the detected inconsistencies, i.e., AIS contacts annotated as stop in the middle of the TSS and AIS positions annotated as stop while the vessel was going in straight line. All these events were initially considered as false positive events by the expert, which was con rmed by checking the raw data. Indeed, those ships were clearly underway. On the contrary, the stops detected near the Brest harbour were considered as more trustful. The approach used by the expert is summarised by the qualitative confusion matrix in Table 8.