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

LIST OF FIGURES

LIST OF TABLES

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 final 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:

- Unification of component-specific 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 offline 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;

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

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

✓ 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;

✓ 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;

✓ 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.

✓ 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 workflow descriptions in D1.6 and D1.7;
- D5.1 - Maritime use case detailed definition [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 specification [4]: D5.6 realises the specifications 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 (final): 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. offline components), which are not explicitly related to the fulfilment 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 refined 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. Unified framework for capturing experiments results, linking the big data challenges, the prototype components and the performance criteria;
3. Intertwined types of assessments: The different components are assessed both through computational and expert-based methods;
4. Scenario-based data preparation: The data encode specific 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 different workpackages and to deliver a unified 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 first 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 scientific teams and the future users of the datAcron prototype (identifying "what" the datAcron components should answer). Figure 1 illustrates the idea.

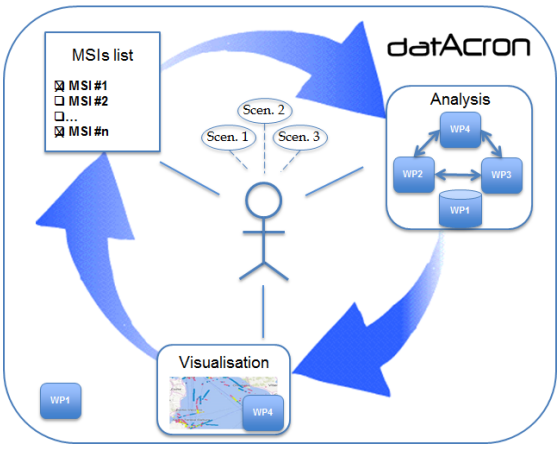


Figure 1: Overview of the user-centred datAcron evaluation design [15]

The MSIs once detected by the different datAcron components designed within WPs 1, 2, 3, and 4 will be rendered to the operator through the visualisation component. Put within a specific operational context defined by maritime use cases and scenarios, the operator would accomplish a specific 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 different datAcron components to fit 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 offline. In comparison, SC12, SC21, SC31 and SC32 do not involve *IVA* offline and SC22 does not involve *IVA* online. Although the datAcron components are developed to fit 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 final evaluation activities reflects 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 offline", 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 offline	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 defined value	MSI 11	-	-	-	-	-	-	-
Course based MSIs	Change of course	MSI 12	-	-	1	-	-	-	-
	Mismatch course vessel destination	MSI 13	-	-	-	-	-	-	-
	Mismatch course user defined 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 affected	MSI 21	-	-	-	-	-	1	-
	Aground	MSI 22	-	-	-	-	-	1	-
	Engaged in fishing	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.

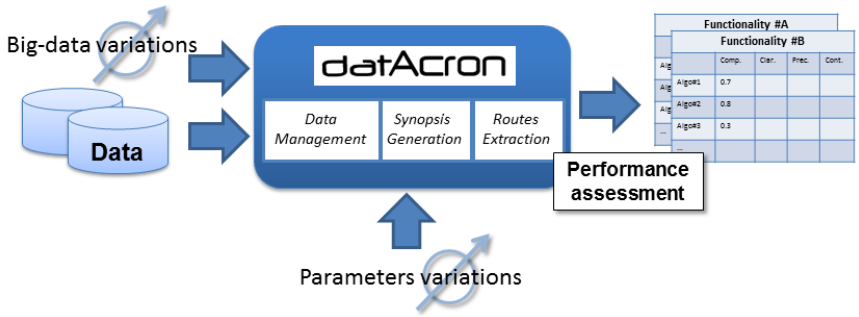
2.1.2 Semantic levels of assessment

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).

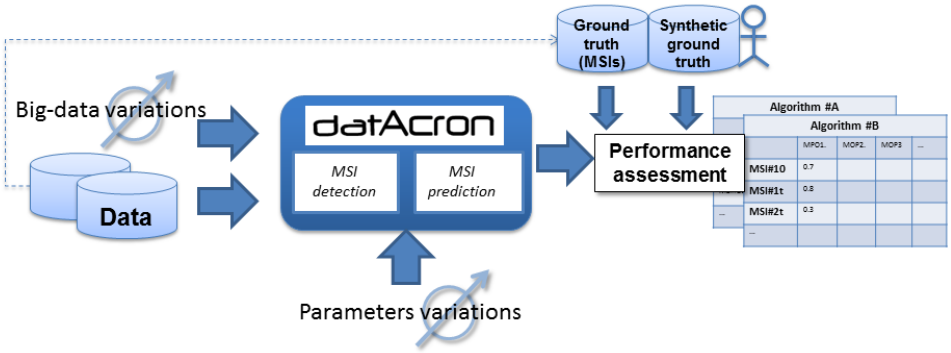
Such a decomposition allows to specify the role of the human in the evaluation: Not involved at under-MSI functionalities assessments, involved as an **expert** at MSI-level assessment and involved as a **user** (or **operator**) at the scenario-level.

2.2 Capturing assessment results in a unified framework

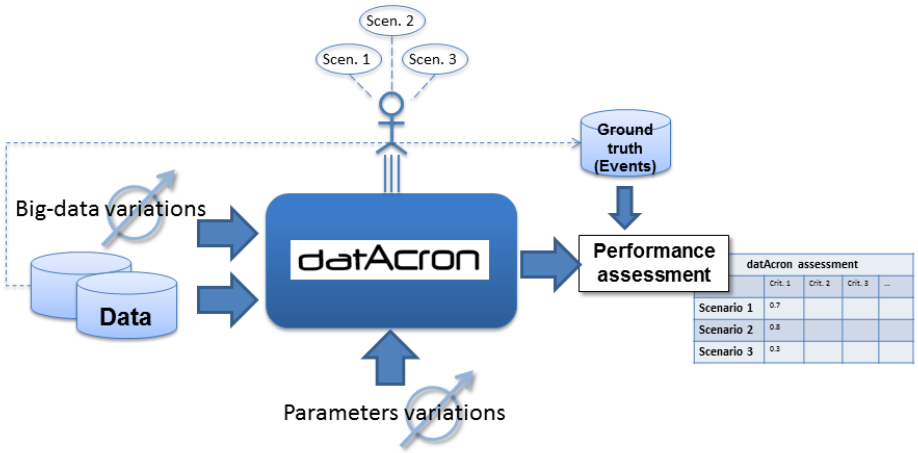
For providing a comprehensive overview of the assessment results from all workpackages, for easing their comparability as well as for allowing the estimate of the response of datAcron prototype to different big data variations, a unified framework is proposed according to which the delivered evaluation results are sorted. The development of the unified 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 different kinds of data variation, different evaluation criteria and measures used by the different component designers. This evaluation space provides a unified framework which allows to make the results of assessment of the different 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 different evaluation measures. Secondly, the criteria are clustered into the different 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 fulfilment process nor part of the evaluation process. This group of evaluation criteria includes to the evaluation criteria *timeliness*, *scalability*, *compression rate* and *classification 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 classification quality which was quantified by true and false positive and negative detections of Synopses Generator and Complex Event Recognition.
  - An operator is part of the task fulfilment. In this case, the performance of datAcron prototype is evaluated via performance criteria quantifying the clarity and the effectiveness.

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 effects 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. Briefly by focusing on the data that are taken into account by the different components;
2. Components are evaluated separately, before interaction effects between the components are evaluated;
3. Different 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 fulfilment 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 different 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 offline components.

In the following, datAcron prototype components are called "online" if they process streamed data, whilst "offline" 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 fixing 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 different 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 fulfilling 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 specific 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 different components. The sequential development and evaluation of components of the datAcron prototype have been conducted in parallel for efficiency 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 quantification of performance improvements.

2.3.2 Expert-based assessment workflow

Figure 3 summarises the workflow of the assessment involving experts that will be detailed in Sections 4 and 5. The reference dataset provided by NARI [20, 22] and briefly described in Section 2.4), either as whole or parts of it, has been processed by the different datAcron components (*LED*, *SG*, *SI*, and *CER* mainly), and the resulting output data have been evaluated at two levels, called respectively *MSI 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.

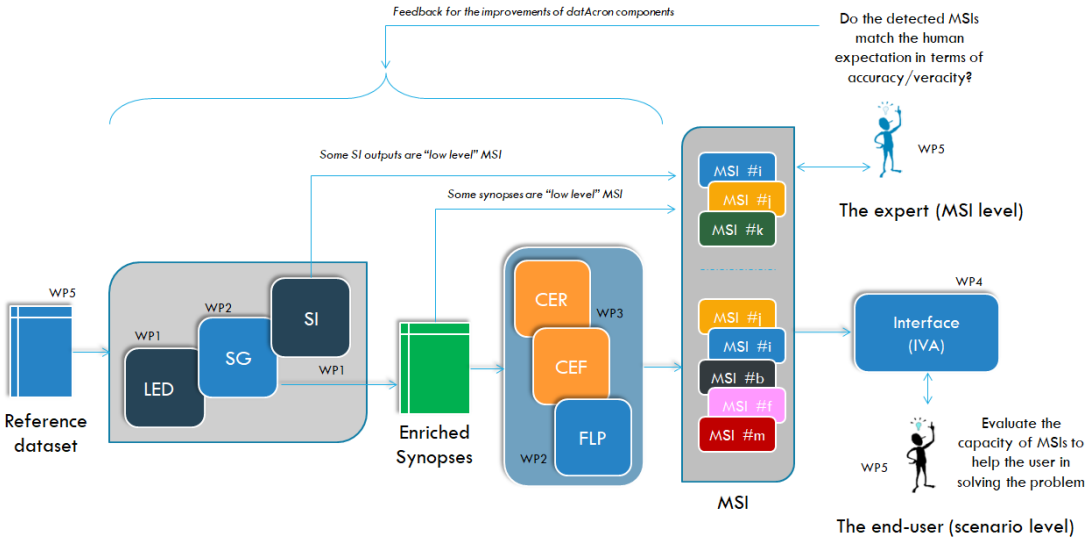


Figure 3: Maritime data workflow and expert-driven assessment principle

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 of *accuracy/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 modification methods (degradation, enrichment), targeted for specific 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 different 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 effectively 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 specific 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 the *CER* 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 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 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 different 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 unified 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)
Classification 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 (Confidence) Effectiveness	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 predefined 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:

- $NARI_{SGv0.7}$: 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.
- $NARI_{SGv0.8}$: 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.
- $IMISG_{SGv0.7}$: 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.
- $IMISG_{SGv0.8}$: 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.
- $\overline{IMISG}_{SGv0.7}$: 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 $IMISG_{SGv0.7}$ extended by the addition of synthetic messages, calculated by interpolating between existing messages.

Algorithms' parameters variations

$\triangle\Theta$ *:* (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.

$\triangle T$*:* (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 classified 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 different 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, different conclusions appear to be pertinent. Firstly, the accuracy of v0.7 is Pareto efficient 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 differences need to be considered as possible factor which limits the comparability of the results. Secondly, the compression ratio is not correlated with this effect. 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 differences are small, given possible differences 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 significant 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 significantly**. Comparing the reduction of latency from single thread to eight thread configuration, 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 unaffected. Both results are intuitive, as [MSI#12] - change in heading is functionally dependent on delta theta.

Contrary, the impact of delta T is different from NARI to IMISG. While the compression ratio remains virtually unaffected in NARI, both in average and MSI wise, the compression ratio in IMISG increases significantly with larger delta T. This finding 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 (final) 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 of *CER* differing both in the input and in the pattern used for the recognition. Each version is assumed to be implemented both in an offline 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 definition of the event recognition predicate is different 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 affected

[MSI#22] Aground

[MSI#23] Engaged in fishing

[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 predefined 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 + fishing 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 $NARI_1$+ fishing 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 + fishing 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 $IMISG_1$) + fishing 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)
- Classification 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 different 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*, different 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 identified 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 fishing 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% affecting 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 affected 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 (final), 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 offers in D3.2 [3] and D3.5 [2] the development and evaluation of forecasting indicators additional to the predefined MSIs. These include exemplary implementations allowing for answering the following two questions:

- Which vessel completes a specific 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 fishing? [2]

Data variations

Volume variations:

- $NARI_1$: Critical point stream of $NARI_0$: "The first 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]

confidence 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].
- Classification quality: Precision, spread (min), distance (min) [3], [2].

Description of experiments

Table 52 summarizes the performed experiments characterized by the application of different 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 fishing in the near future. For both events to be predicted, a strong positive correlation between the algorithm parameter confidence 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 confidence 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 difference 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 benefit from a larger distance. In IMISG dataset the correlation between the confidence 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 confidence 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 different 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 off 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 fishing of more than 378 minutes. This makes the use of *CEF* supposedly also beneficial in online scenarios like collision avoidance. For IMISG dataset this effect 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 of *CEF* 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 offers additional functionalities with respect to the predefined 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 differ 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 benefit 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 effect 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 benefiting 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 first 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 final 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 different 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 fifteen

[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 differs from usability assessment, the later evaluating the ability of human operators to interact correctly and profitably 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 fishing 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 different 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 filtering (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, fishing 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 (filtering, 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 exemplification 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 first MSI, *stop*, he evaluated two different 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 affected 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 affected the quality of the detection.

4.2 MSI assessment, first period

The first 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 (specifically, points 1, 3-5 in the methodology), he visualised the results in the GIS software and identified 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 Traffic 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 confirmed 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.

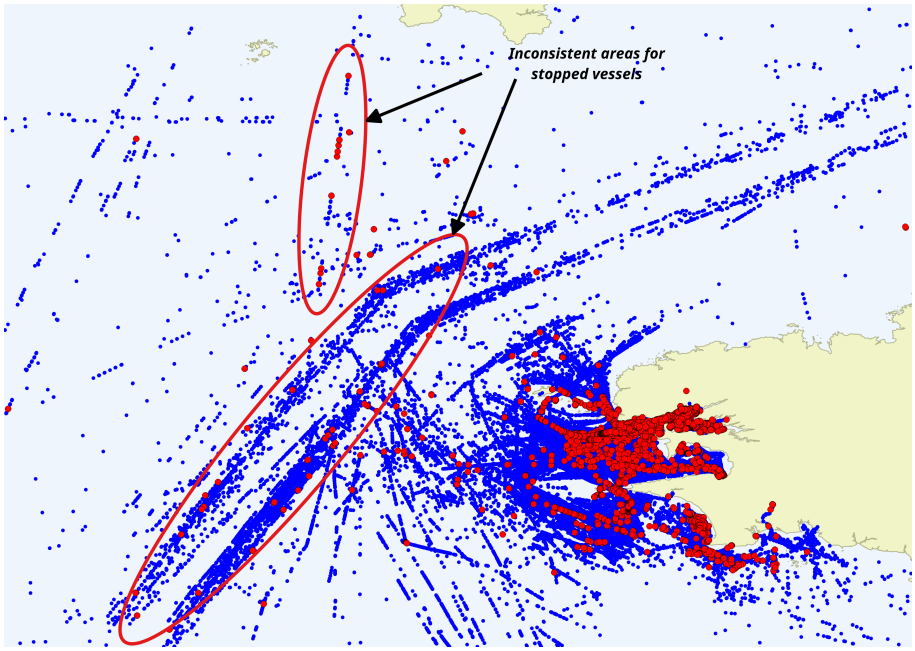


Figure 4: Expert accuracy assessment of *stop* events detected by *SG* in the Channel between Brest (FR) and UK. Blue points represent AIS positions of all other vessels. Red dots represent *stop* events detected by the SG. Inconsistencies are highlighted, including *stop* events detected in straight trajectory and in the Ushant TSS.

When evaluating the accuracy of *stop* events detected by *CER* (which uses *SG* output), the expert considered the same area in the channel between between Brest and UK. In this case, he compared the results with a baseline dataset he computed, according to the procedure described above (points 1-5). To prepare the ground-truth dataset to be used for comparison, he used the analysis capabilities of the GIS (filtering). Specifically, he followed the approach below (where the software is mentioned to exemplify the steps):

1. Import the *CER* results of *stop* events in the (relational) database (PostgreSQL);
2. Query the raw AIS data to select the vessels positions matching the vessel positions included in *CER* results (i.e., matching: vessel identifier (MMSI), starting time, ending time) and save them in a new table for convenience. He updated this table with the spatial capabilities necessary to enable the visualisation in the GIS (add a spatial column in the corresponding table using PostGIS geometry types);
3. Visualise the two baseline datasets in the GIS QGIS, and filter them to show only points with speed over 2.7 knots;
4. In the GIS, overlay the baseline dataset and the result dataset comparing dynamic raw data corresponding to the time range of the detected event with the processed data, to highlight inconsistencies.

In Figure 5, *stop* events reporting a speed over 2.7 knots are highlighted in green, demonstrating that errors identified in the previous example have been propagated. The corresponding qualitative confusion matrix is given in Table 9.

4.2.2 Assessment of *Underway* Event Detections

In this case, the expert was asked to evaluate the accuracy of *underway* events detected by *CER*. Using a similar approach to the one described in the previous section for the same component,

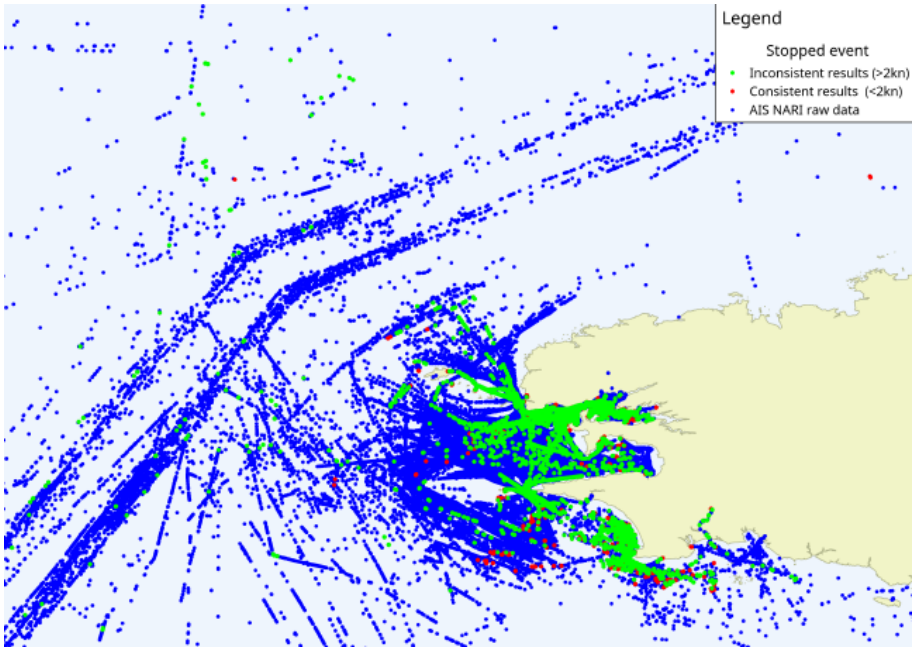


Figure 5: Expert accuracy assessment of *stop* events detected by *CER*. In green, *stop* false positive detections, corresponding to AIS contacts reporting speed over 2.7 knots. Red points are consistent *stop* events.

		CER	
		Stop	*non-Stop*
Expert judgement	*Speed* $<$ *2.7 knots*	TP: Red dots in Fig. 5	FN: —
	Speed $\leq$ *2.7 knots*	FP: Green dots in Fig. 5	TN: —

Table 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).

the expert prepared a baseline dataset of vessels moving with a speed over 2.7 knots (cf. points 1-5 in the methodology). AIS raw data have been visualised and filtered in the GIS with respect to the reported speed, and overlaid with results, highlighting inconsistencies. The accuracy of the detection is shown in Figure 6, where red dots represent consistent *underway* detections (speed is above 2.7 knots), and green dots are inconsistent *underway* events with a speed below 2.7 knots. Moreover, several ships sailing on the maritime route are not detected as underway, while they are clearly matching with the event's criteria.

The expert applied the same approach also in a smaller area, the Douarnenez harbour, detecting underway events within the harbour. The results are reported in Figure 7. The expert's approach is summarised in Table 10.

4.2.3 Assessment of *High Speed* Event Detections

The expert evaluated the accuracy of *high-speed* detections output by *CER*. The *high speed* events correspond to ships that exceed a speed threshold in a given area. *High speed* detections are visualised in Figure 8.

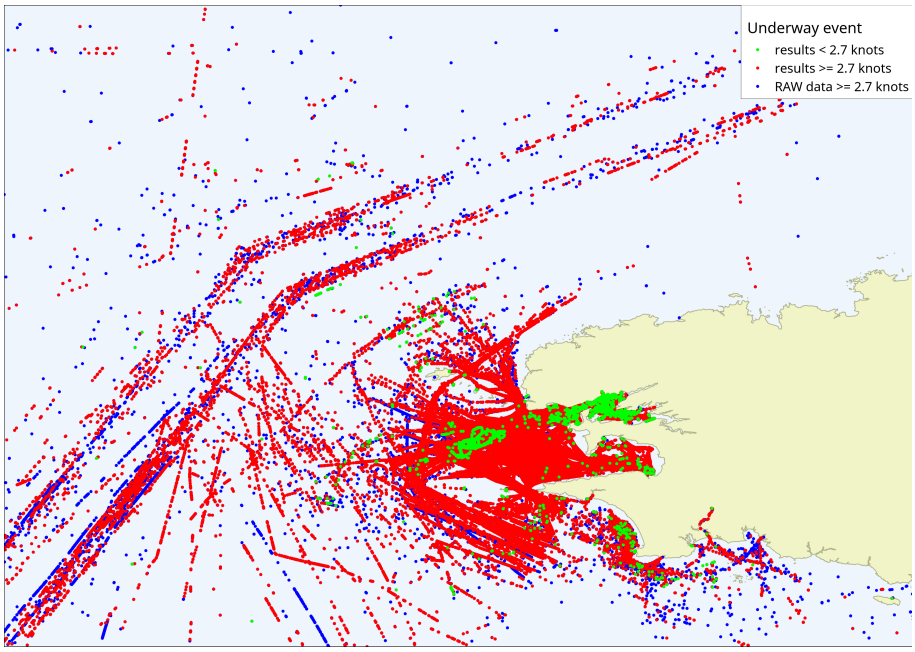


Figure 6: Expert accuracy assessment of *underway* events in the Ushant TSS. Red points are reported positions with speed above 2.7 knots, thus consistent with underway events, while green points are detected underway events with inconsistent speed (< 2.7 knots). Blue points are raw data with speed consistent with underway event, but not detected.

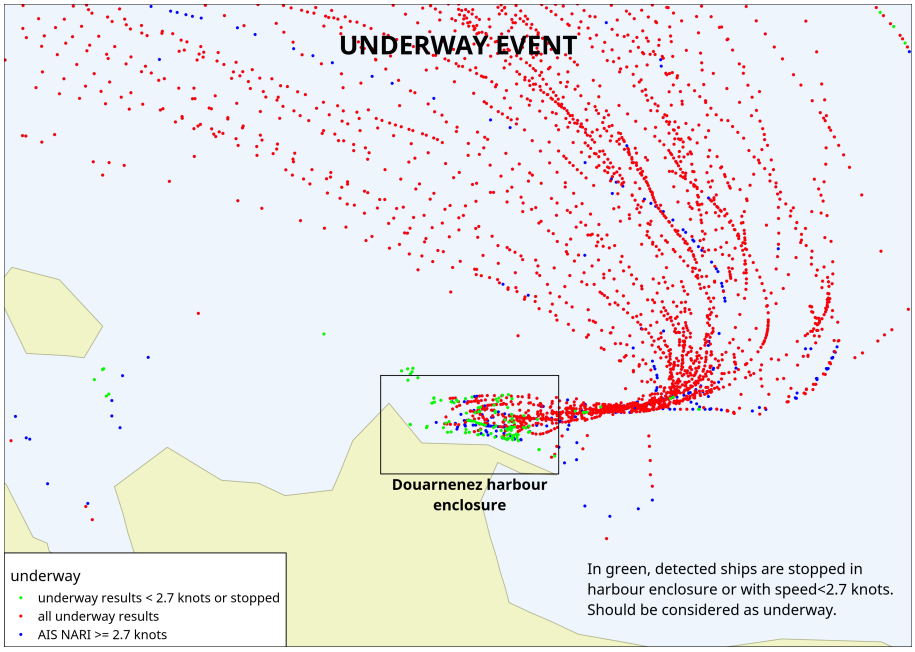


Figure 7: Expert assessment of accuracy of *underway* events in the Douarnenez harbour. Inconsistent events are depicted in green.

		CER	
		Underway	*non-Underway*
Expert judgment	*Speed* $\leq$ *2.7 knots*	TP: Red dots in Fig. 6 and Fig. 7	FN: —
	Speed $<$ *2.7 knots*	FP: Green dots in Fig. 6 and Fig. 7	TN: Blue dots in Fig. 6 and Fig. 7

Table 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).

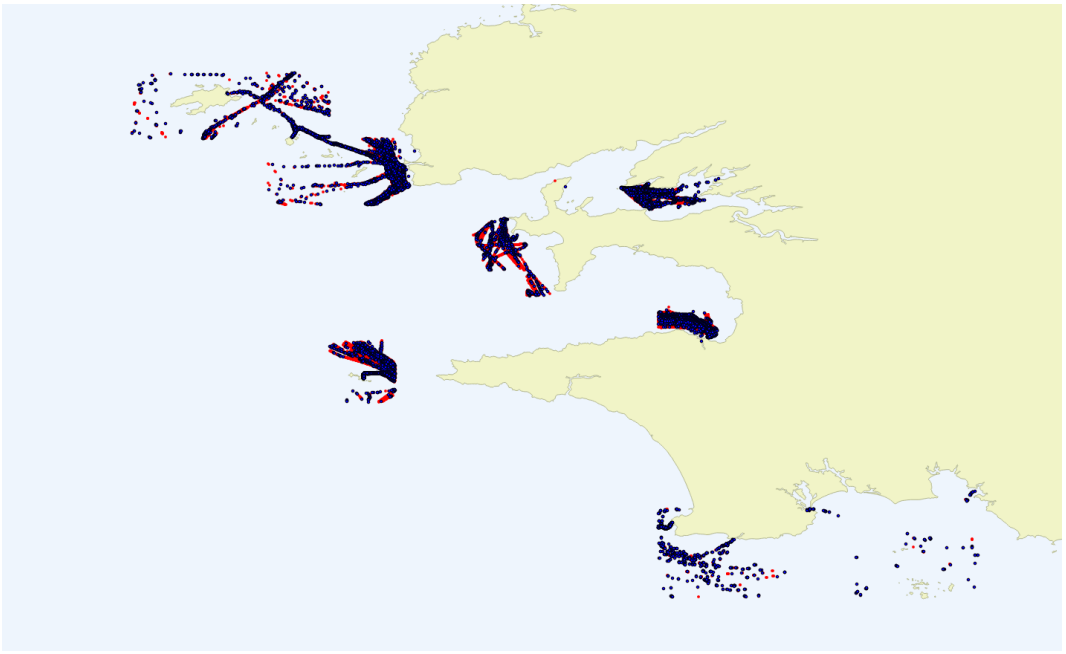


Figure 8: Visualisation of *high speed* detected events. Red dots represent detected *high speed* events, while blue dots are normal traffic.

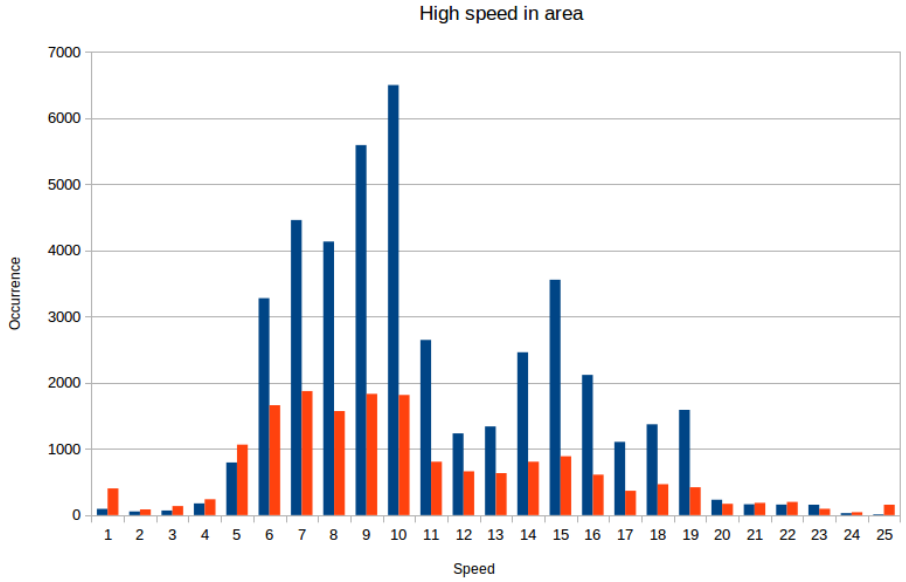


Figure 9: Speed occurrence (in knots) in the original AIS dataset (in blue) and in the compressed data (in red). Extreme speed values are over-represented in the compressed dataset.

In this case, the expert could not rely on an agreed speed threshold to create a baseline dataset as he did in the previous examples. Therefore, he first decided to qualitatively analyse the occurrence of speed in the original AIS raw dataset, and to compared it with the dataset input to *CER* (cf. point 6 in the evaluation methodology). This dataset was a compressed version of the original AIS dataset, processed using *SG* to preserve the most representative base events. Later, statistical analysis of speed in the area and computation of theoretical speed of fishing vessel has been done using the European fleet register to correct threshold retained by *SG* and *CER* and improve assessment with reference thresholds.

The result of this comparison is shown in Figure 9. This simple visual comparison of the two occurrence distributions enabled the expert to highlight an important issue in the compressed dataset, because extreme speed values (< 6 knots, ≤ 21 knots) are over-represented with respect to the original dataset. According to this analysis, the synopses generator is re-labelling the reported speed, instead of filtering out irrelevant AIS contacts as it was assumed by *CER*, therefore it cannot be safely applied as a pre-processing step for event detection.

4.3 MSIs assessment, second period

The objective of this second step was to evaluate the final computations of events (only WP3) prior to the final experimentation. The assessment has been realised along month M34 (October 2018) on the MSIs detected by CER component. The objective of this second period was to complete and validate the definition of MISs and to validate associated thresholds. The assessment is based on accuracy and follows the expert-based approach experimented during the first round. We focus on events related to collision scenario in order to validate MSI before the experiment.

Areas of interest: In order to facilitate the assessment, the choice was made to focus on a 5 days-dataset (February 2016, From 20th to 25th). Two representative areas of the original

dataset were selected. These include several events that are supposed to be detected by CER (tugging, fishing, stopped, etc...). The areas (illustrated in Figure 10) are:

- Area 1: Brest port surroundings
- Area 2: Open seas area

Two sets of complex events were provided by CER to assess (one per area).

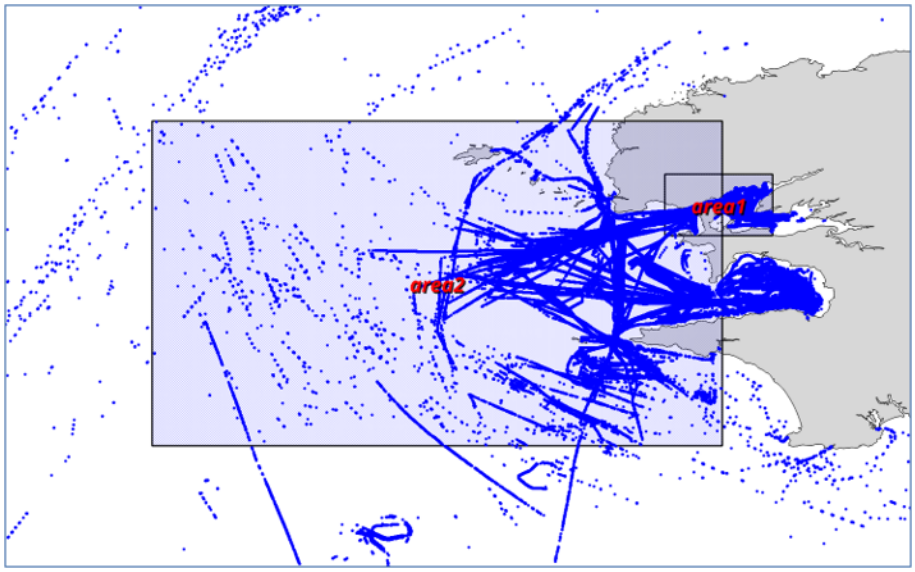


Figure 10: Areas Studied

4.3.1 MSI considered

The assessment report was done on the stand-alone version of CER at the end of October. This CER version provides the most complete results, but was not the one used for the final experiment. CER detected the following events (in bold the events assessed during this second evaluation):

- MSI #2 **Vessel within area** (within JRC fishing area and Natura 2000 combined)
- MSI #6 **Stopped vessel**
- MSI #7 Changing speed vessel
- MSI #8 **High speed near coast**
- MSI #9 Speed not matching vessel type
- MSI #11(a) Speed incompatible with user defined threshold (**speed greater than a max**, **Speed lower than a min**, travel speed)
- MSI #11(b) Low speed vessel
- MSI #16 Communication gap
- MSI #19 **Vessel under way**
- MSI #20(a) **Vessel at anchor**
- MSI #20(b) **Vessel moored**

- MSI #22 Vessel aground
- MSI #23 Trawling (trawling, trawling course, trawling in Natura 2000)
- MSI #24 **Tugging**
- MSI #25 In SAR operation
- MSI #26 Vessel loitering
- MSI #27 **Vessel adrift**

4.3.2 Results

The assessment has been done by the expert in the context of the collision scenario, so the goal was to evaluate the detection of MSIs by the *CER* component versus raw data[5]. Due to the dataset volume, an exhaustive evaluation of each AIS position was not possible. Then, the choice was made to focus on veracity/accuracy of detections using the qualitative expert-based assessment (cf. Section 4.1). The role of the expert in this assessment was not specifically to highlight the good results of the *CER* component but rather to look for confusing maritime situations. The assessment of results is classified into three categories linked to collision scenarios: (1) Distinguish events during fishing activity; (2) Evaluate speed-based detection and; (3) Evaluate mobilities within areas.

Distinguish events during fishing activity

The objective here is to evaluate the *CER* event detection in the situation of fishing vessels in activity. The expert focused on this evaluation because distinguish between fishing activity and events like adrift is quite tricky.

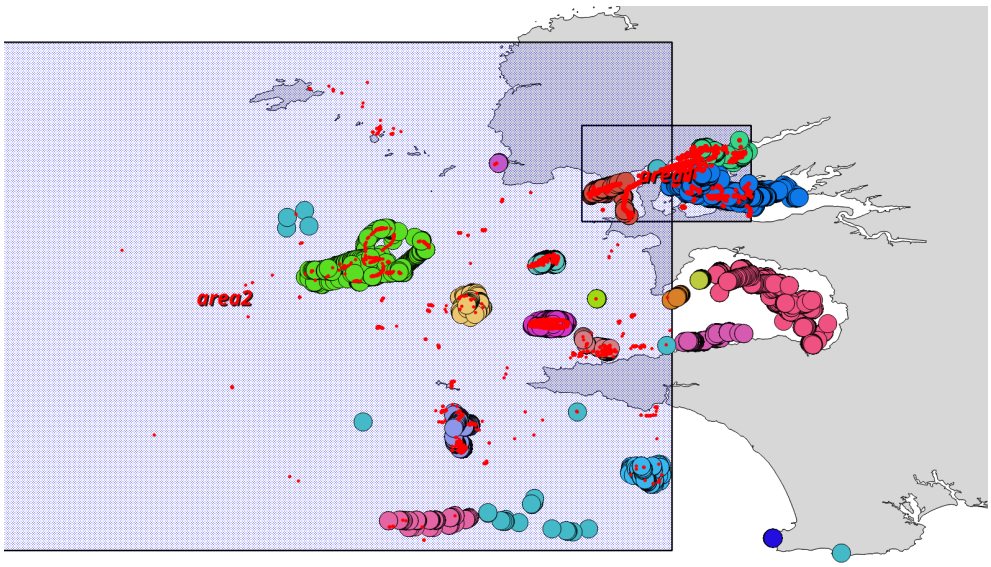


Figure 11: *CER* adrift event detections (red dots) and fishing activities identified by the expert (coloured circles). Inconsistencies arise when both detection overlap.

- Adrift (MSI #27): The expert focused on all fishing vessels having speed consistent with fishing activity and he assessed the event adrift (cf. Figure 11). Many detections correspond

[5] *CER* detection has been applied on top of *SG* 0.8 output

to ships transmitting wrong heading (=511). It seems that another type of wrong detection occurs when ships change course, in particular in the port vicinity. The expert suggested that the necessary time for a big ship to reach the desired course can generate a difference between COG and heading.

- Vessel at anchor or moored (MSI #20): The expert did a comparison between *CER* results and ships declared with the status *at anchor* (in raw data) or having a low speed (using 3 Knots as described in AIS technical specification). The evaluation showed that all the detected situations have speed consistent with the event but not all situations are detected. The expert identified some detected events that actually correspond to almost stopped fishing vessels working with fish trap or net.
- Stopped (MSI #6): While few *CER* detections are slightly over the speed threshold in open seas, detections around port areas are consistent with the speed parameter threshold (0.5 knots).

Evaluate speed-based detection for security

Speed is a highly important parameter for the security of vessels including prevention of collisions (which is the scenario retained for the final experiment, cf. section 5). An analysis of few *CER* events directly based on the speed has been done by the expert.

- Speed greater than max and speed lower than a min (MSI #11): The expert compared *CER* results of ships having a speed below or above 4 knots (speed threshold used). He identified that detections are good (true positive). He emphasised a detection ratio of about 10% versus reference dataset. Which is consistent with compression ratios of the *SG* component. He remarked that some positions over the threshold are not detected.
- High speed near coast (MSI #8): The expert focused on ships having a speed above 5 knots and within a range of 300 meters (agreed threshold) around the coast. He reported a globally good detection status. He however identified 2 wrongly detected trajectories (having high speed but not close enough to the coast) in Area 1 (cf. Figure 10).

Evaluate mobilities within areas for collision scenario

Collision scenario has been initially designed to distinguish collisions from similar situations (where the expert can be mistaken) considering ships underway in specific areas (e.g. around TSS). Similar situations (to collision) are near collision, tugging and rendez-vous. One of the objective was therefore to focus on the relevant MSIs, i.e. within area (MSI #2), underway (MSI #19), tugging (MSI #24) and rendez-vous (MSI #28). Let us note that rendez-vous has not been provided, assessment has been done on the 3 others.

- *Within Area* (MSI #2): The expert evaluated *CER* detections of all vessels (not only fishing vessels) within JRC fishing area and Natura 2000 areas (combined). The events detected are mainly correctly detected (true positive), however not all of them have been detected as illustrated in Figure 13. Without *SG* outputs in that evaluation, it is not possible to confirm a detection ratio. Moreover Figure 13 also clearly shows false positives (red points outside green or violet areas).
- Underway (MSI #19): The expert assessed together with *CER* designer the underway event. Both identified bad detection results mainly due the speed field. Two problems have affected the detection of the underway event: There is a lot of messages with zero speed in the original dataset (and also messages with default speed value). These messages have a different value in the output of the synopses. Indeed, the *SG* component states that "*For each critical point, extra attributes that refer to its instantaneous speed over ground and heading over ground (actually, on the sea surface) are included as computed during the*

summarization process". Secondly the *SG* component performs noise reduction based on these computations. This has badly affected the *CER* detections which sometimes achieve to detect 'underway" event when using the original dataset while it fails in the case of *SG* outputs (in particular, detection may start but never ends). As a consequence the event provides complete trajectories (fishing vessels in activity) sailing with very low speed (typically below 3 knots) and exhibits missed detection.

- *Tugging* (MSI #24): The expert focused on Area 1 looking for tugging events, specifically between Brest strait and the port where he identified three events. The *CER* component correctly detected two events corresponding to (1) a vessel in distress and (2) a regular tugging from an anchor area to the port. The third tugging event (cf. Figure 12) where the vessel in distress Besiktas Orient was tugged by Abeille Bourbon[6] has not been detected.

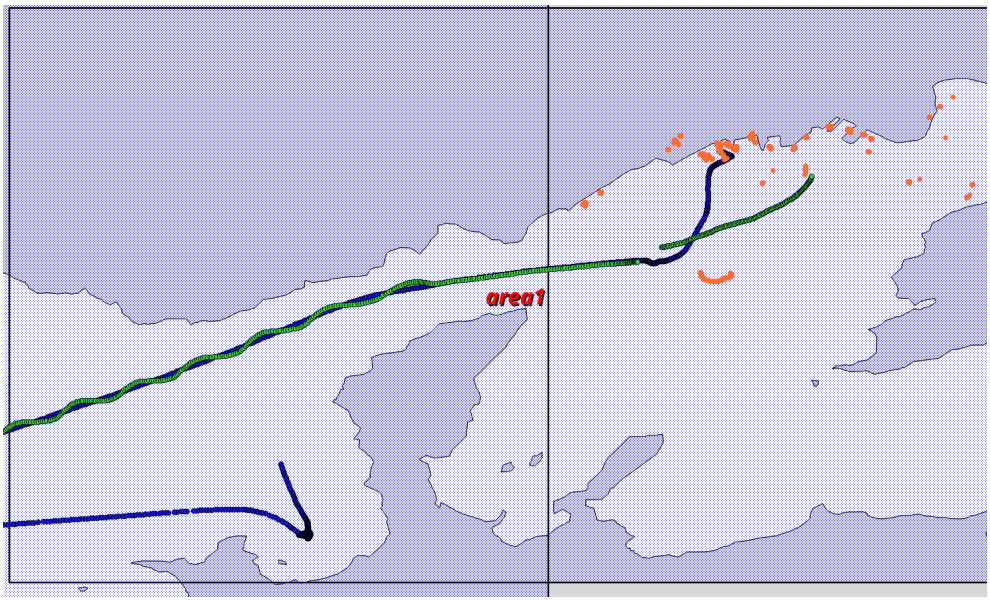


Figure 12: *Tugging* events: trajectories correspond to a missed detection of tugging event (by the CER component), green dots represents vessel in distress (Besiktas Orient) being tugged by Abeille Bourbon.

4.3.3 Assessment of MSI #3: *On a maritime route*

We present here the results of the assessment of the MSI #3, standing for the presence of a vessel on a maritime route. Table 11 presents the confusion matrix of the route association. For this MSI a slightly different methodology has been followed: Each of the 190 AIS contact previously labelled by maritime experts has been assigned to a route (A to V) or to none of the routes (Z). The results provided by the algorithm presented in [6], which assigns a contact to a route using the geometrical shapes of the routes, the position and the course over ground of the vessel, were then compared to the experts annotations. The output of the MSI #3 detector is one of the routes labelled by the experts (A to V), another route (Y) or none of the routes (Z). In Table 11, the rows represent the labelling by the experts, the columns represent the result of the algorithm and each value in the matrix the number of matching. Out of the 190 contacts, 117 (62%) have been assigned to the correct route.

This result depends on (1) the distance for associating vessels to routes, (2) the quality of the routes pre-extracted (itself depending on the raw AIS historical data) in their fidelity to

[6]https://www.youtube.com/watch?v=QZxSnvkp-i4

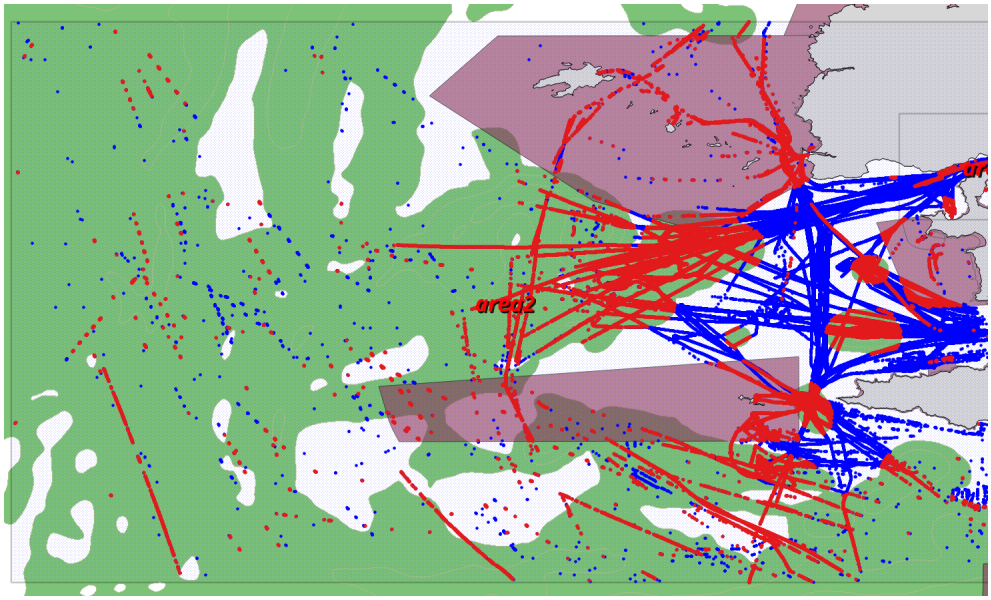


Figure 13: *Within Area* detections combining JRC fishing areas (in green) and Natura 2000 (in violet). Red dots correspond to detected events by the *CER* and blue dots are raw data

Table 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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	Z	Y	Σ
A	**1**		1																				1		3
B																							2		2
C			**6**				1																6		13
D				**2**									1												3
E					**3**	3				1															7
F					1	**1**																	2		4
G			1				**6**										5								12
H							1	**6**										2					1		10
I									**1**										1						2
J										**4**													1		5
K											**1**								1						2
L																							1		1
M																									0
N				1						1				**5**									1		8
O						1	1																		2
P																**1**									1
Q											1			1			**9**								11
R																		**1**					1		2
S											1														1
T																				**1**					1
U						1																			1
V																						**2**			2
Z			3		1	3	4	5			5		2	1		1	1		1		1		**67**	2	97
Σ	1	0	11	3	5	9	13	11	1	6	8	0	3	7	0	2	15	3	3	1	1	2	83	2	190

the current traffic and (3) the synthetic representation of the route. The 62% of matching rate should be read as the ability of the algorithm to associate a vessel to route **as good as would have done an expert analysing the same contacts**. Also, the annotation captures only the appreciation of an expert regarding the association of a vessel to a route. Besides the quality of

the route, the display plays also a role in the annotation, and another expert could have provided another labelling.

Concluding remarks

This section presents the methodology developed for datAcron in order to provide an expert-based assessment at the MSI-level. It shows how necessary is such an approach for the design, improvement and evaluation of event modelling and algorithms assessment. The expert-based assessment has produced many meaningful feedback helping the definition or improvements of synopses generator and complex event detection. The methodology originally combines qualitative and quantitative evaluation realised by a maritime expert having technical skills to process data himself.

This expert-based preparation and analysis of maritime data is also a means to organise annotation of data to further improve the quality of assessment. This work stresses first how challenging the annotation of dataset is and second the prior annotation realised for the final experiment (cf. Section 5.6). Next section comments the influence and quality of such annotation in the assessment process.

The weak point of the current methodology relies on the subjectivity of the expert(s), but also on their difficulty to process very large volume of data. It that sense, this work would require additional investigation to further automate and guide the analysis of data.

5 Assessment at the Scenario Level

The objective of the assessment at the scenario level is to evaluate the capability of operators to achieve their mission with a visualisation enriched by detected and predicted or forecast events (MSIs). It relies on a scenario-driven design of experiments (user in context of use).

5.1 Scenario level experiments

The assessment at the scenario level has been organised in two periods. The first experiments were organised at CMRE (with NARI and FRHF) from the 22^{nd} to the 29^{th} of March 2018 and involved three maritime experts. The second experiments involved three maritime experts, two cadets from the French navy and the maritime expert which assessed the accuracy of MSIs along the project. The experiment took place at CMRE, with visiting personnel from NARI and FRHF, from the 5^{th} to the 9^{th} of November 2018. The agenda of this second week of experiments is given in Annex 8.5 and some photos of experts in action is provided in Figure 14.

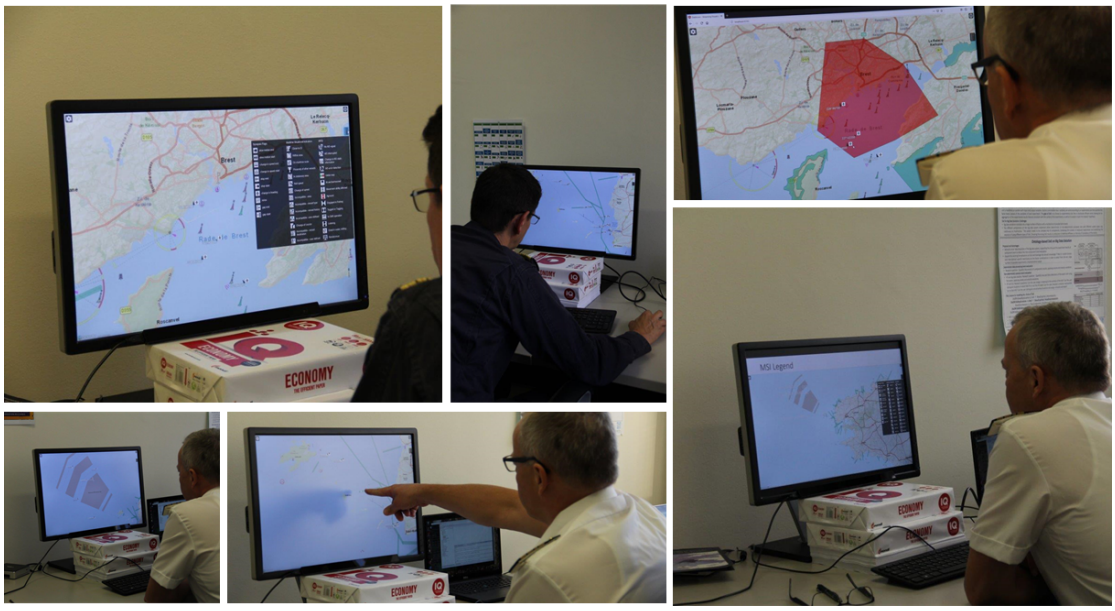


Figure 14: Maritime surveillance experts during the experiments of November 2018 held at CMRE.

5.1.1 Types of experiments

The assessment at the scenario level has been divided into 5 experiments which are summarised in Table 12. as follows:

Experiment 0 is a rehearsal for experiments 2 and 3.

Experiment 1 assesses the perceived relevance and utility of AIS and MSI information in the framework of a serious game.

Experiment 2 assesses how the MSIs impact the operators detection of near distance events with respect to only AIS messages.

Experiment 3 assesses the overall implemented functionalities of datAcron prototype, both on the interactive visual analytics component and on the detection capabilities.

Experiment 4 assesses the accuracy of the MSIs detection through an in-depth expert-driven analysis of all detections of datAcron prototype on the dataset used for experiments 2 and 3.

5.1.2 Scenario selection

The focus of the scenario level assessment on one scenario guarantees a yield of experimental results with the largest level of significance, given a limited number of operators. The selection of one scenario amongst the ones defined in Deliverable D5.1 [15] has been done with the help of experts feedback. A survey was conducted with ten (10) military and civilian experts in maritime surveillance, from six (6) different countries (Denmark, Finland, France, Germany, Norway, Romania) who were asked to evaluate the three maritime use cases proposed in D5.1 that include six scenarios requiring operational monitoring of fishing activities. Each scenario addresses a specific maritime security or safety mission, covering different areas:

- *Use Case #1: Secure fishing*, addressing either *prevention of collisions* involving fishing vessels or support to search and rescue operations of *vessels in distress*.
- *Use Case #2: Maritime sustainable development*, tackling the impact of fishing activities on maritime ecosystems towards a responsible exploitation of maritime resources. These two scenarios address the *monitoring of marine protected areas*, in particular to protect them from Illegal Unreported Unregulated (IUU) fishing, and the estimation of *fishing pressure* on areas.
- *Use Case #3: Maritime security*, including two scenarios addressing detection of trafficking, smuggling, *migrants and human trafficking* and other *illicit activities* conducted at sea.

In a first set of questions, the experts evaluated the relevance to the respective national concerns, specifying whether they either personally experienced the problem described in the scenario or are aware of recent events. From these answers, it appears that all use cases are generally relevant to their national concerns (above 70%) and quite recurrent (50% have heard about recent events).

In a second group of questions the experts evaluated whether the scenarios are realistic and challenging. The experts generally judged the scenarios realistic and challenging. The most challenging ones appear to be the scenarios related to the maritime security and the secure fishing use cases, far above maritime sustainable development.

Based on this feedback, it had been decided to focus the experiments on the collision avoidance scenario (SC1.1). The scenario is challenging, of interest and it happens quite often. The experiments run explicitly refer to the experimental descriptions proposed in [4], especially to the experiments for collision avoidance. The description includes one offline setting and one online setting of the collision avoidance evaluation and one performance evaluation. These "Steps to perform" are detailed and implemented in the same order by experiments 1, 2 and 3.

5.2 Experiment 0: Rehearsal

The rehearsal conducted at CMRE facilities in March 2018 involved CMRE, NARI, FRHF and maritime surveillance experts, and focused on two main objectives.

Table 12: Experiments setup summary

				Variations			Data captured							
	Scenario	Expert's task	Data	Variety	Veracity	VI	Visual	Dynamic	Bel.	H_i	Time	ET	SA	Output type	Purpose
EXP #1 - The Variety Game															
EXP #1.1 - Icons	-	Match icons and MSIs names	Set of icons & set of MSI names	-	NO	NO	Board	-	NO	YES	NO	NO	NO	Picture of the board	Capture understandability of icons
EXP #1.2 - MSIs	-	Subjective relevance of MSIs	Set of MSIs cards	-	NO	NO	Board	-	YES	NO	NO	NO	NO	Picture of the board	Capture MSIs a priori relevance
EXP #1.3 - Info impact	SC0	Discriminate between H_1, H_2, H_3	Set of cards for info items	YES	NO	NO	Board	Artificial	YES	YES	NO	NO	NO	Picture of the board	Capture information items relevance
EXP #2 - MSIs for MSA															
EXP #2.1 - MSIs for MSA (1)	SC1-SC2-SC3	Discriminate between H_1, H_2, H_3	X_0	NO	NO	NO	IVA	Real time × 3	YES	YES	YES	YES	NO	Confidence graded, H_i, Time	Assessment of contextual MSIs dynamic display vs AIS only
EXP #2.2 - MSIs for MSA (2)	SC1-SC2-SC3	Discriminate between H_1, H_2, H_3	X_0^*	NO	NO	NO	IVA	Real time × 3	YES	YES	YES	YES	NO	Confidence graded, H_i, Time	Assessment of contextual MSIs dynamic display vs AIS only
EXP #3 - IVA for MSA															
EXP #3.1 - IVA zoom-pan	SC1	Discriminate between H_1, H_2, H_3		NO		YES	IVA	Real time × 3	NO	NO	??	NO	YES	SA and utility questionnaires	Assessment of IVA functionality: Zoom & Pan
EXP #3.2 - IVA filtering	SC2	Discriminate between H_1, H_2, H_3		YES		YES	IVA	Real time × 3	NO	NO	NO	YES	YES	SA and utility questionnaires	Assessment of IVA functionality: Info filtering
EXP #3.3 - IVA full	SC3	Discriminate between H_1, H_2, H_3	$\overline{X}_0$	YES	YES	YES	IVA	Real time × 3	NO	NO	NO	??	YES	SA and utility questionnaires	Assessment of integrated prototype
EXP #4 - Expert MSI assessment															
EXP #4.1 - Expert MSI assessment (1)	SC1-SC2-SC3	Validate or invalidate MSI detections	$\overline{X}_i$	NO	YES	NO	GIS	Static	?	YES	NO	NO	NO	Partial confusion matrix	Assessment of MSIs accuracy & detection veracity robustness
EXP #4.2 - Expert MSI assessment (2)	SC1-SC2-SC3	Validate or invalidate MSI detections	$\overline{X}_i$	NO	YES	NO	GIS	Static	?	YES	NO	NO	NO	Partial confusion matrix	Assessment of MSIs accuracy & detection veracity robustness
EXP #4.3 - Expert MSI assessment (3)	SC1-SC2-SC3	Validate or invalidate MSI detections	$\overline{X}_i$	NO	YES	NO	GIS	Static	?	YES	NO	NO	NO	Partial confusion matrix	Assessment of MSIs accuracy & detection veracity robustness

Firstly, the setup of a maritime prototype. This included especially the testing of data processing workflow (as depicted in Figure 3) and the development of methods for data enrichment and variations, which results are reported in Deliverable D5.5 [12].

Secondly, the testing of the experimental setup to be further applied and possibly adapted to experiments (exp. 2 and 3) of November 2018. The specification of the experimental setup included the design and the selection of suitable icons for displaying MSIs to the expert user ([12], section 4.4). It also included the testing of the three questions listed below (which lead to the selection of the third question for the evaluation of the datAcron prototype in experiments 2 and 3). Further, datasets with different numbers of vessels, zooming-levels and replay speeds were displayed to the operators in order to evaluate whether they were capable of distinguishing the MSI icons at the respective zooming-level and whether an accelerated replay affects the verdict of the operator. The traffic density created by the simultaneous visualisation of 2 near-distance situations with 2 vessels each plus 4 moving vessels which were not involved in the near-distance situation was described by the operators to be challenging but not impossible to occur in reality. As a result, these guidelines were used for the creation of the datasets for experiments 2 and 3.

The setup of experiment 0 is depicted in Figure 15. It is composed of three sub-experiments, each addressing a different research question:

1. AIS-MSI-compliance: The user is asked whether the MSIs shown are corresponding to the AIS messages shown. This sub-experiment is limited to MSIs involving single vessels.
2. MSI-Situation-compliance: The user is asked whether the MSIs shown are corresponding to the AIS messages shown. Contrary to the first sub-experiment, the user is shown situations with two vessels each.
3. Collision prediction: A dataset was prepared which includes eight situations involving two vessels each and shown almost simultaneously to the operator. Four situations end in a collision, four situations do not end in a collision. The operator is asked to classify the situations in collisions and non-collisions.

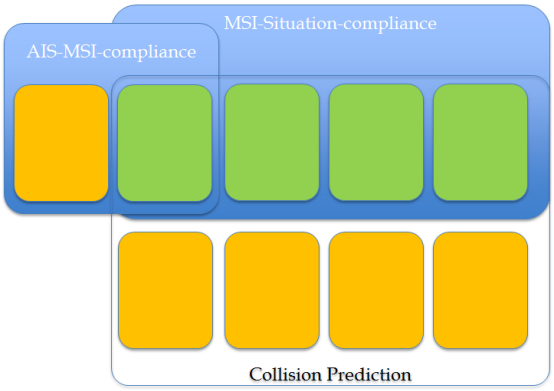


Figure 15: Experimental Rehearsal Design

The results gathered during experiment 0 are both of theoretical and of practical importance for the development and execution of experiment 2 and 3 as well as for the interpretability of the obtained results. As the experiments include an operator, the results include insights on the nature of datasets, on the technical details of the experimental environment and the type of response variables to collect. This includes especially:

- The acceleration of data shown to the operator by factor 3. This allows to present 3

separate datasets instead of 1 dataset to the operator in 30 min without exhaustion of spare mental capacity.

- The size of the display is restricted by the technical specifications of the eye-tracking equipment. The extract of the dataset is chosen respecting both the display and the collision avoidance task.
- The MSI icons are developed in order to communicate the detections to the operator, partly base on COLREGs day shapes and night lights [1]. A familiarisation phase of the operators to the MSI icons is performed at the beginning of experiment 1, thus before experiment 2 and 3.
- The research questions for experiment 2 and 3 build up on research question of experiment 0, investigating the benefit of datAcron MSIs on the task fulfilment of collision avoidance and detection.
- The responses of the operators are collected in a speak-aloud session, allowing for the expression of all associations which are in the following categorised the described methodology which makes the analysis of the experimental results repeatable.

5.3 Experiment 1: The Variety Game

The general purpose of the *Variety Game* is to capture the impact of information variety on human belief assessment. More specifically, the Variety Game is completely disconnected from the numerical implementation of the datAcron prototype while reproducing in a simple way the basic functionalities. It typically answers the question of whether information conveyed by MSIs is useful, compared to contextual information and raw AIS data.

The information variety is understood along the three groups of information items:

1. Contextual information
2. Raw AIS data
3. Maritime Situational Indicators

This specific experiment is divided in three sub-experiments, that look at the following aspects:

1. Maritime Situational Indicators Icons
2. Maritime Situational Indicators perceived relevance
3. Information variety

The game is composed by two initial minigames, hereafter referred to as the *Icons Minigame* (Exp1.1) and the *MSI Game* (Exp1.2), followed by the main game (*Information Variety Game* - Exp1.3).

5.3.1 Icons Minigame

This first experiment consists of a *gamification* of the one exercised during the March rehearsal. As reported in Deliverable 5.5, the March experimentations included some assessment of newly defined icons through a paper questionnaire asking experts to match the MSI icons with their respective verbal descriptions. This assessment is necessary to understand and capture a possible

impact of the icons interpretation in the chain of evaluation.

The main purpose of the Icons Minigame is to familiarise the player with the icons and their corresponding Maritime Situational Indicators in order to support the subsequent experiments. However, the design of the game could allow to capture and quantify the suitability of the selected icons to represent the MSIs.

The elements provided to the players are:

- Board with a matrix of slots, each displaying the verbal expressions of MSIs presented to the player;
- A set of tokens displaying the MSI icons.

The players were requested to associate each token to an MSI slot. No time constrains were imposed to perform the task and shuffling of the tokens were allowed.

The data captured consist of one picture of the board at the end of the Icon Minigame recording the tokens associations and an optional intermediate picture. The latter was taken when the players affirmed that they did not know how to associate the remaining icons. In this case the game facilitator asked the player to pursue the association exercise, but took a picture of the board, to be able to record the drop in the player's self-confidence (meaning that after that step, the match between icons and MSIs could be considered differently in subsequent analysis).

A preliminary analysis allowed to highlight the following results:

- The exercise lasted more than expected (thirteen minutes vs fifteen minutes);
- The icons that were associated more easily and with higher confidence, were the ones showing standard navigational status of a ship [1];
- Some family of icons (e.g. the MSIs that relate to AIS issues) were easily mapped to the corresponding family of MSIs, however the precise assignment was no easy, as the minimal differences were not easily noticed;
- Most of the players expressed the need for stylised and clear icons, which should be preferred to detailed ones, that could be confusing and misleading in operational environments.

Further analysis and an increased number of samples (*i.e.* players) will allow to quantify the degree of matching between the icons and their corresponding verbal expression. This is an interesting side product of the Variety game.

5.3.2 Maritime Situational Indicators Minigame

The purpose of this minigame is to further familiarise the players with the icons and corresponding MSIs and to capture their prior subjective relevance and utility assessment of the MSIs to perform the task at hand.

As the perceived relevance of information is task dependent, it has been important to start introducing the player to the role he had to perform in the game. More specifically, it was explained that the task is to monitor the area of responsibility with respect to potential safety (e.g. collision avoidance) and security threats (e.g. IUU). The elements provided to the player are:

- A board similar to the previous one, with information items being the contextual information layers, AIS fields and all MSIs.
- Three sets of coloured tokens;

- Three sets of coloured numerical tokens.

The player is requested to rank each information item using the coloured tokens (*i.e.* green if highly relevant, yellow if moderately relevant, red if not relevant). Then, within each colour category the information is ranked from the most relevant to the least relevant, using the coloured numerical tokens. During the minigame shuffling and equal rankings were allowed. No time constrains were imposed to the player to perform the ranking exercise. The data captured consist of one picture of the board at the end of the minigame. The minigame roughly lasted 20 minutes.

A preliminary analysis allowed to highlight the following results:

- Most players ranked the AIS information of high relevance;
- As expected, from the results it appears that the most relevant MSIs are the ones connected to safety.

Although it was not the main goal of this minigame, further quantitative analysis will allow to define the *perceived relevance* of the different information item, which could be used as baseline for a comparison with the information relevance as per the results of the Exp1.3.

5.3.3 Information Variety Game

The purpose of the game is to capture the impact of information variety on players' belief assessment, possibly understanding which are the information items with a higher impact.

One basic assumption of the game is that the player is familiar with maritime surveillance (they have been invited based on their experience) and the MSIs (thanks to the previous minigames). In this experiment only the impact of variety is assessed, therefore the veracity dimension is fixed, *i.e.*, the information items do not exhibit any uncertainty.

During the game the player is presented with a scenario and a role. More specifically, the player has to perform a monitoring task in a given area to prevent collisions and detect potential transshipment of goods.

The experiment methodology is a variation of previous games, namely the Risk Game [14] and Reliability Game [7]. As in the previous games the player is presented with incoming information (*i.e.* contextual information, AIS, MSIs) regarding an event which is unfolding between two ships. The information is presented through cards on which a stylised monitor screen is reproduced. The belief state of the player is recorded through the position of the belief tokens on the game board. The game is divided into two rounds, one in which no MSIs are provided and one in which MSIs are provided to the player. The full game length is about forty minutes and at the end of each round a picture of the game board is taken in order to collect the belief data. Moreover, the confidence in the assessment is recorded.

One important result of such an experiment is the observation of how the MSIs appear to be more useful in an initial phase of the assessment (*i.e.* when the first pieces of information are received), while their relevance appears to decrease once the player attention has already been focused on the situation and that he/she is looking mainly to detailed vessel kinematic information (*i.e.* speed, course). More specifically, it appears that MSIs are more useful to drive the attention towards a situation of interest, but their relevance with respect to the understanding of the specific situation decreases the deeper the assessment process goes. In general, the MSIs support operators task of monitoring broad areas.

The new Variety Game designed has shown to be efficient in capturing the impact of information variety on belief assessment. Unfortunately, the small number of players of the Variety

Game (6) does not allow to provide quantitative results on the relevance and utility of the MSIs, but the experiment still highlights that they might ease the operator's task compared to raw AIS data. Further analysis will evaluate the evolution of the degree of belief toward the three hypotheses as the information items are provided to the player. This should help identifying the most relevant items.

5.4 Experiment 2: MSIs for MSA

Experiment 2 has the objective to measure the effect of MSIs on the operators' awareness of collision situations. Therefore, the experiment is constructed in the way that it can help to answer the research question whether or not the MSIs improve the maritime situational awareness compared to the absence of MSIs.

Therefore, the following hypothesis H0 is tested:

> MSIs are not changing the prediction and detection of collisions by the maritime surveillance operator.

The aim of the investigation is the founded rejection of this hypothesis, in order to replace it by the alternative hypothesis, stating that MSIs are changing the prediction and detection of collisions. The hypothesis is split up, specified and tested into the following sub-hypothesis:

- H0a: MSIs are not changing the predictability of collisions.
- H0b: MSIs are not changing the detectability of collisions.
- H0c: MSIs are not changing the predictability of near-distance situations.
- H0d: MSIs are not changing the detectability of near-distance situations.
- H0e: MSIs are not changing the average time between prediction and collision.
- H0f: MSIs are not changing the average time between prediction and near-distance situation.
- H0g: MSIs are not changing the confidence in the prediction of collisions.
- H0h: MSIs are not changing the confidence in the detection of collisions.
- H0i: MSIs are not changing the confidence in the prediction of near-distance situations.
- H0j: MSIs are not changing the confidence in the detection of near-distance situations.
- H0k: MSIs are not changing the situational awareness of near-distance situations.

For testing the different hypotheses, 3 scenarios are displayed to different operators. Each corresponding dataset includes two near-distance situations, *i.e.* collision, near-collision or rendezvous between two vessels plus additional vessels. The task of the operator is to avoid collisions if possible or to detect collisions after they took place. For this, the operator is asked to think aloud and to describe the visualised situations. All statements are recorded on paper by the facilitator with the point in time and a confidence value. After an introduction where the task is explained following a protocol, the interaction with the operator is reduced to a minimum with the exception of asking confidence values for the stated situational assessments.

Figure 16 proposes a taxonomy for near-distance situations. While rendezvous is an intrinsically intentional manoeuvre of two or more vessels, close quarter situations are unintended and dangerous situations which can result both in a near-collision/near-miss or an actual collision. For a functional representation of close-quarter situations see [11].

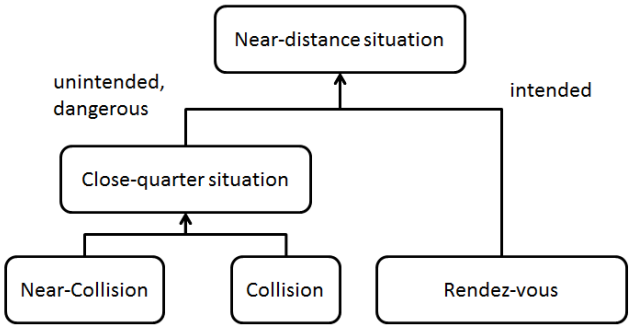


Figure 16: Near-distance situation taxonomy

5.4.1 Relation between experiment 2 and other experiments

All experimental designs have commonalities which allow an interpretation of the collected data beyond the scope of one single experiment, as well as differences which allow to answer the different research questions.

The relations between experiment 1 and 2 are depicted in the following, firstly commonalities and secondly differences. Both experiments are aligned on the fact that the operator is asked to fulfil the task of collision avoidance and both experiments offer the same three hypotheses of near-distance situations, namely collision, near-collision and rendezvous. Further, the information available to the operator includes in both experiments dynamic AIS data and MSIs.

The experiments differ firstly and most importantly in the fact that experiment 2 imposes a time constraint on the operator. The visualisations cannot be interrupted by the operator and all situations are constantly evolving. This obliges the operator to react instantaneously. Secondly, the three expected hypotheses in experiment 2 are not named explicitly beforehand, thus the operator does not know if there are different or additional hypotheses than in experiment 1. Nonetheless, it can be assumed that the 3 hypotheses of experiment 2 are known to the operator, as experiment 1 proceeded experiment 2. Thirdly, in experiment 2 MSI data is not added successively to the AIS data, but there are scenarios purely with MSI and purely without MSI. This allows for the measurement of the impact of the availability of MSIs. Finally, the visualisation component is required in experiment 2 while it is not in experiment 1 which uses a board game and no numerical display.

5.4.2 Assumptions

1. The three types of situations displayed to the operator are supposed to be known, given that Experiment 1 proceeded Experiment 2.

2. Every two situations, e.g. Collision1 and Collision2 are both representative for their type of situation, e.g. Collision.

3. Every two situations of the same type are similar, compared to situations of other type. Thus, the selection which situation is shown with and without MSIs can be done arbitrarily. Supposed that a situation shown with MSIs attracts the attention of the operator stronger than a situation shown simultaneously, but without MSIs. In order to preclude this bias, the assignment of situations shown with and without MSIs is not done randomly but dataset-wise, *i.e.* scenario dataset 1 without MSIs, scenario dataset 2 with MSIs, etc.

4. Every two situations of the same type are different, so that the familiarity with the firstly shown situation does not allow insights into the development or result of the situation

shown in the second place. Thus, a learning effect between the two situations is supposed to be negligible.

5.4.3 Criteria and Measures

Criteria (from D5.3): Timeliness, Accuracy, Clarity Measures:

- Timeliness: Time between cognition of a possible collision situation and hailing the vessel and the actual time of collision.
- Accuracy: number of True(T)/False(F) Positive(P)/Negative(N), *i.e.* TP, TN, FP, FN verdicts of the operator with respect to the situation.
- Clarity: Confidence.

Reminder of the prototype setup: *VIZ* (with zooming and panning disabled), eye tracking.

5.4.4 Experimental Design

Three scenario datasets are prepared each including two near-distance situations between two vessels plus additional vessels. Scenario 1 includes one collision and one rendezvous. Scenario 2 includes one collision and one near-collision, and scenario 3 includes one near-collision and one rendezvous.

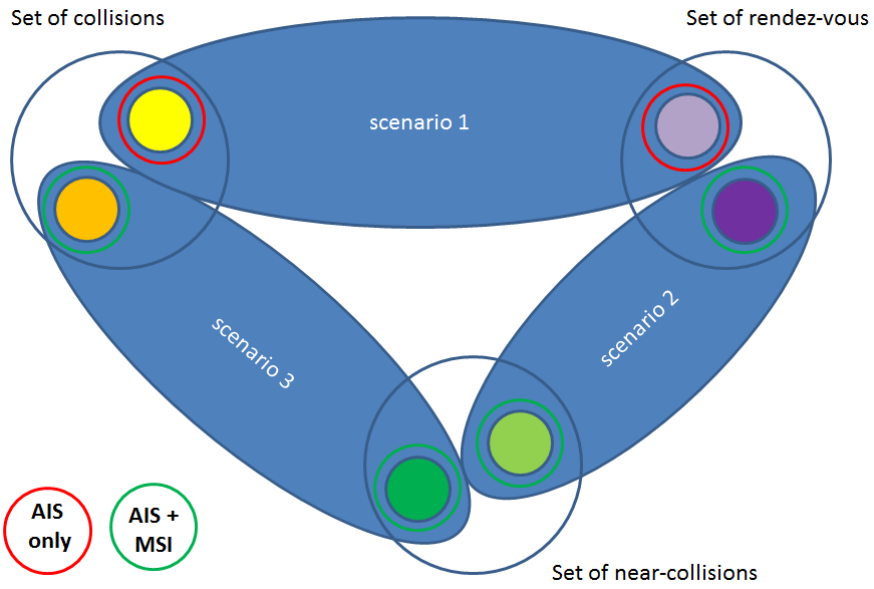


Figure 17: Experiment 2 pairwise scenario design

This experimental design is chosen to be the most appropriate design for answering the research questions out of the three designs tested during the experimental rehearsal on 22.-30. March 2018 and is described in section 5.2.

The research question is addressed by comparing two situations of the same type, one with and the other one without MSIs. This setup is chosen for the collisions, as this situation corresponds to the scenario #11, and the rendezvous. The two near-collisions are both shown with MSIs in order to assert the stated assumption, that two situations of the same type are similar, compared to situations of other types. The near-collision is chosen as it is supposed to lie between Collision and rendezvous.

The situations are described by AIS data and MSIs specified in the following:

- Scenario 1 (CoRe) without MSI: Collision1 vs. rendezvous1
- Scenario 2 (NeRe) with MSI: Near collision1 vs. rendezvous2
- Scenario 3 (CoNe) with MSI: Collision2 vs. Near collision2

The datasets are shown in time laps speeded up by factor 3. experiment 2:

- Scenario 1: collision in minute 20 (6.7), rendezvous in minute 24 (8).
- Scenario 2: rendezvous in minute 15 (5), near-collision in minute 25 (8.3).
- Scenario 3: near collision in minute 23 (7.7), collision in minute 26 (8.7).

Each type of event occurs twice, so that conclusions are drawn both on the level of a specific situation and on the level of the type of situation. Each situation, e.g. Collision1 occurs once, so that the operator is shown different events in order to avoid learning effects.

5.4.5 Data labelling

In order to assess only the impact of the MSIs information on the operator situation awareness, **regardless their correct detection by the datAcron components**, the dataset has been labelled by experts, providing some "perfect" MSI detections to the operator. The dataset was thus then enriched by "true" MSI detections. For the second prototype setup of November 2018, the annotation of data was done by two maritime domain experts, to whom a set of thresholds was given for the triggering of the indicators. The annotation is performed for every single AIS position and for all the MSIs of interest. Each raw of the AIS dataset corresponding to the scenario is enriched with additional column of binary values, where 0 means that the MSI is not activated and 1 means that a MSI is activated. Details can be found in Deliverable D5.5.

A subset of MSIs is chosen for the labelling process, based on two criteria out of which at least one needs to be matched. Firstly, each datAcron component has to be represented by a subset of MSIs. For this a plausible set of MSIs that matches the scenario is chosen including MSIs #s2, #3, #4, #12, #16, #19, #23 and #28. Secondly, all MSIs are taken into account, that are detected by the datAcron components for at least one of the three scenarios. This set includes MSIs #2, #4, #6, #7, #12, #15, #16, #19, #28. The union of the two sets results in a catalogue against which the AIS data of the scenarios is checked manually with respect to the thresholds given in Table 13.

5.4.6 Data collected

The data collected through experiment 2 is of different types:

1. Written records of the verbal comments of the operators with the corresponding timestamp;
2. The confidence level of the operator in his/her situation assessment;
3. The situation assessment as one the possible type of events;
4. Situation awareness through a survey paper form;
5. The focus of attention through the eye-tracking software.

All printable results are included in the Appendices of the present document and in Deliverable D4.9.

Table 13: Thresholds for MSI manual labelling

#	MSI	threshold
2	Within a given area (TSS)	-
3	On a maritime route	-
4	Proximity to other vessels	100m
6	Null speed	$<$ 0.5 knts
7	Change of speed	25% between change of speed start and end
12	Change of course	4 degrees
15	No AIS reception	-
16	AIS reception interrupted	$>$1800s
19	Under way	-
23	Engaged in fishing	-
28	rendezvous	-

5.4.7 Results interpretation

The qualitative interpretation of the verbal comments made by the operators is performed into two steps. Firstly, in the classification step, the situational descriptions are assigned to one of the more general situational types of near-distance situations depicted in 16 for guaranteeing the comparability of the situational descriptions. Secondly, in the comparison step, the generalised situational types are compared to the situational type of the reference dataset. This step is performed both for the prediction phase and the detection phase. The prediction phase starts at the beginning of each scenario dataset and ends before the perceived moment of event occurrence marked in **bold** letters in the tables 55, 56, etc. The detection phase includes the perceived moment of event occurrence as well as all following assessment of the situation marked in *italic* letters in the tables 55, 56, etc. and ends with the end of the scenario dataset.

The classification step: As expected by not communicating the set of possible situations to the operator, additional hypotheses are mentioned by the operator, which are described in the following. In order to allow a comparison of the mentioned situation types, the situation types are classified according to the taxonomy depicted in Table 16. This generalisation includes the following situational descriptions:

- Near-distance situation:
  - Near situation or near distance situation. Mentioned in experiment 2 in scenario 2 in minute 3:30 for describing the rendezvous situation and scenario 3 in minute 1:30 for describing the situation before the collision.
  - Proximity situation. Mentioned in experiment 2, scenario 2 minute 5:50 for describing the situation before the near collision.
  - Loitering in close position. Mentioned in experiment 2 in minute 7:30 for describing the rendezvous situation in scenario 1.
- Close-quarter situation:
  - Dangerous situation. Mentioned in experiment 2 in scenario 3, minute 6:30 for describing the situation before the collision.
  - Not a normal situation. Mentioned in experiment 2, scenario 1 in minute 4 for describing the situation before the collision.

  - Crossing situation. Mentioned in experiment 3, scenario 2 in minute 2 for describing the situation before the rendezvous.
  - Close contact situation. Mentioned in experiment 2 in scenario 1, minute 2:30 for describing the situation before the rendezvous.
  - Near-collision situation. Mentioned in experiment2, scenario 3, minute 7 for describing the situation before the collision.

- Near-collision situation:
  - Overtaking manoeuvre. Mentioned in experiment 2 in minute 7:30 for describing the near-collision situation in scenario 2 and in the post situation description of the near-collision in scenario 3.
  - Near-miss situation. Mentioned in experiment 2, scenario 2 in minute 7 describing the near-collision. situation.
  - Close point of contact. Mentioned in experiment 3, scenario 2 in two post-situational assessments for describing a rendezvous and a near collision situation.

- Collision situation:
  - Imminent collision. Mentioned in experiment 2 in scenario 2 in minute 4:00 for describing the situation before the rendezvous.

- Rendezvous situation:
  - Handover. Mentioned in experiment 2 in post situation description of the rendezvous situation in scenario 1.
  - Hand shaking. Mentioned in experiment 2 in post situation description of the near-collision situation in scenario 2.
  - Loitering in close position. Mentioned in experiment 2 in minute 2 for describing the situation before the rendezvous.

The comparison step: In the comparison step the generalised situational assessments of the operator from the prior classification step, are compared to the reference dataset and rated as true positive, true negative, false positive or false negative. This rating step is performed both for the prediction and the detection phase:

- True positive (TP): The description of the situation by the operator coincides with or includes the situation type of the reference data with respect to the near-distance taxonomy in Table 16. For the inclusion, a distinction is made between prediction and detection. E.g. if a close-quarter situation (cqs) is predicted and the reference data includes a collision, the prediction is rated as a true positive:
  - Prediction: cqs $\sqcap$ c->TP, cqs $\sqcap$ nc->TP, nds $\sqcap$ rdv->TP.
  - Detection: noc $\sqcap$ nc->TP, noc $\sqcap$ rdv->TP.

- True negative (TN): No negative statements are asked from the operator, thus no true negatives are recorded. Assuming that the absence of positive statement is interpreted as a negative statement, a list of true negative detections can be computed for all pairs of vessels included in the respective dataset.

- False positive (FP): The description of the situation made by the operator is very specific, *i.e.* collision, near-collision, rendezvous or tugging and does not coincide with the situation type of the reference data with respect to the near-distance taxonomy in table 16, e.g. a tugging is described where no tugging takes place.

- False negative (FN): A situation in the reference data is not described as such, e.g. a collision is not described at all.

5.4.8 Analysis

Tables 14 and 15 summarise the absolute and the relative results of experiment 2. From Table 15 it becomes clear that for scenarios with MSIs the true positive rates are equal or larger to the scenarios without MSIs. This effect is further discussed in the next section.

Table 14: Summary of results of experiment 2

	prediction				detection			
Situation	TP	FP	FN	$\sum$	TP	FP	FN	$\sum$
Collision without MSIs	1	0	2	3	1	0	2	3
rendezvous without MSI	1	2	0	3	2	1	0	3
rendezvous with MSI	2	1	0	3	2	1	0	3
Near-collision with MSI	3	0	0	3	2	1	0	3
Collision with MSI	3	0	0	3	2	0	1	3
Near-collision2 with MSI	3	0	0	3	3	0	0	3

Table 15: Summary - True Positive, False Positive and False Negative Rates

	prediction				detection			
Situation	TPR	FPR	FNR	$\sum$	TPR	FPR	FNR	$\sum$
Collision without MSIs	0.33	0	0.67	1	0.33	0	0.67	1
Collision with MSI	1	0	0	1	0.67	0	0.33	1
rendezvous without MSI	0.33	0.67	0	1	0.67	0.33	0	1
rendezvous with MSI	0.67	0.33	0	1	0.67	0.33	0	1
Near-collision with MSI	1	0	0	1	0.83	0.17	0	1

1. Effect of MSIs on prediction and detection

For comparing the effects of scenarios with MSIs on the prediction and the detection of collisions and near-distance situations, four hypotheses are tested:

- H0a: MSIs are not changing the predictability of collisions.
- H0b: MSIs are not changing the detectability of collisions.
- H0c: MSIs are not changing the predictability of near-distance situations.
- H0d: MSIs are not changing the detectability of near-distance situations.

Out of the three scenario datasets, a subset is chosen in order to test the hypothesis. For H0a and H0b only CoNoMSI from scenario 1 and CoMSI from scenario 3 are compared. For H0c and H0d CoNoMSI and RdvNoMSI from scenario 1 are compared to CoMSI from scenario 3 and RdvMSI1 from scenario 2.

Assuming the existence of a specific situation, *i.e.* a collision, the analysis is limited to TP, FP and FN while TN are neglected. For both hypothesis TP, FP and FN are discriminate with respect to prediction and detection. Different Criteria and measures are used for quantification,

as described in 5.4.3, *i.e.* accuracy/true positive and false negative rate, time to detection and confidence.

True positive, false positive and false negative rates with and without MSIs for H1 and H2 are given in tables 16 and 17.

Table 16: H0a,b - True Positive, False Positive and False Negative Rates

	prediction				detection			
Situation	TPR	FPR	FNR	$\sum$	TPR	FPR	FNR	$\sum$
Collision without MSIs	0.33	0	0.67	1	0.33	0	0.67	1
Collision with MSI	1	0	0	1	0.67	0	0.33	1

Table 16 indicates a positive effect of the MSI on the task fulfilment of collision avoidance and detection. Both in the prediction and the detection phase, the presence of MSIs increases the true positive rate and reduces the false negative rate.

Table 17: H0c,d - True Positive and False Negative Rates

	prediction				detection			
Situation	TP	FP	FN	$\sum$	TP	FP	FN	$\sum$
Near-distance situation without MSIs	0.33	0.33	0.33	1	0.5	0.17	0.33	1
Near-distance situation with MSI	0.83	0.17	0	1	0.67	0.17	0.17	1.01

Table 17 indicates a positive effect of the MSI on the classification of near-distance situations, both for the prediction and the detection. Both for the prediction and the detection of near-distance situations the true positive rate is higher while MSIs are present and false positive and negative rates are lower.

Concluding the discussion on the effect of MSIs on the prediction and detection of collisions and near-distance situations, the hypothesis can either be reject or not:

- H0a: MSIs are not changing the predictability of collisions. (rejected)
- H0b: MSIs are not changing the detectability of collisions. (rejected)
- H0c: MSIs are not changing the predictability of near-distance situations. (rejected)
- H0d: MSIs are not changing the detectability of near-distance situations. (rejected)

All observed effects that lead to the proposition of hypothesis rejection indicate an improvement of the prediction and the detection of both collisions and near-distance situations, either measured by the improvement of two measures or more. Nonetheless the results are not statistically significant, given the small number of operators.

2. Effect of MSIs on the average prediction time

For comparing the effect of MSIs on the average prediction time, the following time related hypothesis are tested:

- H0e: MSIs are not changing the average time between prediction and collision.

- H0f: MSIs are not changing the average time between prediction and near-distance situation.

The time spans between event prediction and event occurrence with and without MSIs are listed in tables 18, 19. In both cases, the time is denoted in minutes of the reference dataset. As the reference dataset is accelerated by factor 3, the real time span is three times the stated values.

Table 18: H1 - Average time between TP prediction and event occurrence.

	prediction
Situation	Average Time † to * (min)
Collision without MSIs	3
Collision with MSI	0.5

Table 18 resumes the average time span between the TP prediction of a collision and its occurrence, based on the row "Time †to *" of tables 55, 56, 57 and the rows CoNoMSI and CoMSI. It shows that the time span between the TP prediction of a collision and its occurrence is larger without MSIs.

Table 19: H2 - Average time between TP prediction and event occurrence.

	prediction
Situation	Average Time † to * (min)
Near-distance situations without MSIs	4
Near-distance situations with MSI	2

Table 19 resumes the average time span between the TP prediction of a near-distance situation and its occurrence, based on the row "Time †to *" of tables 55, 56, 57. The time between the TP prediction and the occurrence of the near-distance situation is larger without MSI.

Concluding the comparison of the average time between the prediction of collisions and near-distance situations and the occurrence of the predicted event, both tested hypothesis can be rejected.

- H0e: MSIs are not changing the average time between prediction and collision. (rejected)
- H0f: MSIs are not changing the average time between prediction and near-distance situation. (rejected)

All observed effects that allow for the proposition of the rejections indicate an extension of the prediction time for scenarios without MSIs, both for collisions and near-distance situations. Again, the sample size does not allow for the conclusion of significant correlations.

3. Effect of MSIs on the confidence of predictions and detections

For quantifying the effect of MSIs on the confidence of operators when predicting and detecting collisions and near-distance situations, the confidence related hypotheses are tested as follows:

- H0g: MSIs are not changing the confidence in the prediction of collisions.
- H0h: MSIs are not changing the confidence in the detection of collisions.
- H0i: MSIs are not changing the confidence in the prediction of near-distance situations.

- H0j: MSIs are not changing the confidence in the detection of near-distance situations.

The average confidence of predictions and detections is shown in tables 20, 21. The stated confidence is based on the values in tables 55, 56, 57, if available, dividing prediction and detection values with respect to time point of the actual event occurrence, marked by a *-sign. As prediction confidence value, the last confidence value before the event occurrence is considered that specifies the statement which lead to the rating of TP, FP or NP. For instance, the prediction confidence of Expert 1 in scenario 1, RdvNoMSI is 4. Here, the assessment as a near-distance-situation leads to a TP prediction in minute 3 which is specified by the confidence value 4 in minute 7. Similarly, the detection confidence value is derived from the last confidence value before the end of the reference dataset that specifies the respective TP, FP or NP statement. For Collision without MSIs, TP and FN are aggregated over column CoNoMSI while FP are aggregated over column RdvNoMSI. For Collision with MSI, TP and FN are aggregated over column CoMSI while FP are aggregated over all other columns with MSIs, *i.e.* RdvMSI1, NcMSI1 and NcMSI2.

Table 20: H1 - Average confidence in prediction and detection.

	prediction			detection		
Situation	TP	FP	FN	TP	FP	FN
Collision without MSIs	-	-	-	3	4.5	5
Collision with MSI	4	-	-	4.5	3	5

For the available confidence values both with and without MSIs, the FN detections show high level (5) indicating that the operators are confident that no collision happened, even though this situational assertion is wrong. Hence the MSIs are not reducing the confidence in a wrong situational assessment. Comparing true and false detections between collisions with and without MSIs a slight tendency towards an increase of confidence in TP and a reduction in FP is visible. This trend is supported by a confidence of 4 in the case of TP prediction. Summing up and given the small size of 8 samples, the findings are not representative but indicate a possible increase of confidence for true predictions and detections and reduce the confidence at least for FP detections.

Table 21: H2 - Average confidence in prediction and detection.

	prediction			detection		
Situation	TP	FP	FN	TP	FP	FN
Near-distance situation without MSIs	4	5	-	5	4.3	5
Near-distance situation with MSI	4.5	-	-	4.85	4.3	5

The confidence values shown in table 21 are confirming the tendencies found in table 20. All stated confidence values are in the upper half of possible values, irrespectively of the veracity of the situational description that is specified with the confidence value. Given a slightly larger sample size of 21 observations, it becomes observable that the confidence values range from 3 to 5. The difference in confidence between near-distance situations with and without MSIs for TP predictions and events is not significant. Thus, the hypotheses that the level of confidence is equal with and without MSIs can not be rejected.

Concluding the comparisons of confidence values for collision and near-distance situation prediction and detection, the four hypotheses can not be rejected as the differences between the confidence values are small and become even smaller for larger sample sizes.

- H0g: MSIs are not changing the confidence in the prediction of collisions. (not rejected)

- H0h: MSIs are not changing the confidence in the detection of collisions. (not rejected)
- H0i: MSIs are not changing the confidence in the prediction of near-distance situations. (not rejected)
- H0j: MSIs are not changing the confidence in the detection of near-distance situations. (not rejected)

4. Effect of MSIs on the situational awareness

For the evaluation the effect of MSIs on the maritime situational awareness, the following hypothesis is tested:

- H0k: MSIs are not changing the situational awareness of near-distance situations.

The hypothesis is further decomposed into the different measurement dimensions of situational awareness, as shown in section 8, cp. [25]. Prior to visiting the average scores, listed in Table 22, for each dimension, a qualitative analysis of the operator self assessment is performed:

- Instability of situation: All three operators rate the instability of the situation without MSIs lower or equal than with MSIs. Out of the three operators two rate the instability of the situation without MSIs to be lower than with MSIs.
- Complexity of situation: All three operators rate the complexity of situations without MSIs lower than with MSIs.
- Variability of situation: All three operators rate the variability of situations without MSIs lower than with MSIs.
- Arousal: No correlation is observed. The arousal in situations with MSIs are rated once lower, once equal and once higher than the arousal in situations without MSIs.
- Concentration of attention: All three operators rate the concentration of attention lower for situations without MSIs compared to situations with MSIs.
- Division of attention: Two operators out of three rate the division of attention to be lower for situations without MSIs compared to situations with MSIs.
- Division and concentration of attention: All three experts rate the division and the concentration of their attention in a positively correlated fashion, *i.e.* more the attention is concentrated, more it is divided.
- Spare mental capacity: All three operators state to have at least as much spare mental capacity in situations without MSIs.
- Information quantity: All three operators estimate having gained less information from situations without MSIs than from situations with MSIs.
- Familiarity with situation: All three operators indicate to be very familiar with all situations.

Concluding the comparison of situations with and without MSIs it is possible to reject the investigated hypothesis, as consistent and similar differences in the operator ratings are observable for the majority of measurement dimensions over all operators.

- H0k: MSIs are not changing the situational awareness of near-distance situations. (rejected)

Table 22: Average Situational awareness self assessment. Situational awareness rating cp. [25]: 1-Low, 7-High.

	exp. 2			exp. 3
Dimension of situational awareness cp. [25]	sc.1	sc.2	sc.3	$\overline{sc.2}$
Instability of Situation	3.8	5.2	6.2	4.5
Complexity of Situation	2.5	4.2	5.2	3.2
Variability of Situation	2.5	4.5	5.5	3.8
Arousal	5.5	5.2	5.5	4.8
Concentration of Attention	3.2	4.2	5.5	3.5
Division of Attention	3.2	4.2	5.2	3.8
Spare Mental Capacity	6.2	5	5.5	5.2
Information Quantity	2.5	4.5	5.2	5.2
Familiarity with Situation	6.5	5.8	6.2	6.2

The observations leading to the hypothesis rejection allow for the formulation of different alternative hypotheses that describe a detailed picture of the effect of MSIs on the situational awareness. Specifically, situations with MSIs are perceived as situations with a higher information quantity and operators state to be more concentrated on the situation. The situations with MSIs are perceived to be less stable, more complex and having a higher variability than situations with MSIs. The operators state to have less spare mental capacity and that the division of their attention is divided more importantly in situations with MSIs.

5. Effect of dataset design on results

For estimation the impact of the dataset design on the obtained results, the results on two near-collision situations both enriched with MSIs are compared. In the following, the two near-collision situations are referred to as control situations. The choice of near-collision situations as control situation is funded in the similiarity to collisions in the prediction phase, since both situations are perceived as close-quarter situation and the similarity to rendezvous in the detection phase, since both situations allow the vessels the continuation of their route. The two control situations are included in different scenario datasets, thus happen in different locations with different vessels, different AIS trajectories and at different seconds of the scenario dataset. The effect of the AIS trajectory construction methodology and the subsequent labelling process are assumed to be negligible, if the difference between the two control situations is small compared to the differences between situations with and without MSIs. In order to estimate if the difference between the control situations is small, the same measures are used as for the comparison between scenarios with and without MSIs.

Table 23: Control situations - True Positive, False Positive and False Negative Rates

	prediction				detection			
Situation	TP	FP	FN	$\sum$	TP	FP	FN	$\sum$
Near-collision with MSI	1	0	0	1	0.67	0.33	0	1
Near-collision2 with MSI	1	0	0	1	1	0	0	1

Table 23 describes the similarly rated operator assessment of the two control datasets. The variance of the results are smaller than the variance shown in table 16 and in table 17.

Table 24 shows very similar average time spans between the two different control datasets.

Table 24: Control situations - Average time between TP prediction and event occurrence.

	prediction
Situation	Average Time † to * (min)
Near-collision situations 1 with MSIs	2.5
Near-collision situations 2 with MSI	3

The difference of 0.5 minutes is small compared to the differences of 2.5 minutes in table 18 and 2 minutes in table 19.

Table 25: Control situations - Average confidence in prediction and detection.

	prediction			detection		
Situation	TP	FP	FN	TP	FP	FN
Near-distance situation 1 with MSIs	5	-	-	5	-	5
Near-distance situation 2 with MSI	5	-	-	5	-	-

Table 25 shows again very similar results on the confidence of differently rated predictions and detections of the control situations. Only one false negative detection of a near-distance situation exists, which is due to the classification of the first near-distance situation detection as a rendezvous. Despite this false classification the absolute difference in the average confidence of the rated situational assessments is smaller between the control situations than between situations with and without MSIs.

Concluding, the effect of the dataset design and labelling process, measured by the difference between two near-distance situations as control situations, is small compared to the differences found between situations with and without MSIs. This is valid especially for the prediction and assessment of the situations, for the confidence assigned to those predictions and assessments, as well as the remaining time between situation prediction and situation occurrence. This finding also supports the strong assumption 2 in section 5.4.2. As these findings support the assumption that the effect of the dataset design and labelling processes is small, the applied method is supposed to be suitable also for modelling and labelling the complementary situations and yielding situational datasets whose analysis with respect to an effect of MSIs on the given measures is well-founded.

5.4.9 Conclusions on Experiment 2

Experiment 2 allows for the following conclusions:

As analysed in the preceding paragraph the methodology used is suitable and efficient for capturing the impact of MSIs on the prediction and detection of near-distance situations, including especially collisions, the confidence assigned to those predictions and detections as well as the time between the prediction of a near-distance situation and its occurrence. Thus the findings listed in the following are promoted by the assessment of the used method, but also by the fact that the inclusion of the control situations would even increase these findings, especially the positive impact of MSIs on the prediction and detection of near-distance collisions.

MSIs are useful for the enrichment of AIS information for the prediction and detection of collisions and near-distance situations with the following characteristics:

- MSIs improve the prediction and the detection of both collisions and near-distance situations.

- MSIs are not extending the time between the correct prediction of a near distance situation and its occurrence.
- MSIs are not changing the confidence of experts users in their predictions and detections of near-distance situations.
- MSIs are changing the situational awareness of operators in two ways. Firstly and beneficially, MSIs increase the perceived information quantity of situations and operators state to be more concentrated in situations with MSIs. Secondly and adversely, MSIs increase also the perceived complexity and variability of situations and tendentially reduce the spare mental capacity of operators.

5.5 Experiment 3: Prototype assessment

Building up on the results of experiment 2, especially on the finding that MSIs are improving the prediction and detection of near-distance situations, experiment 3 investigates the more specific hypothesis $\overline{H0}$: "MSIs calculated by datAcron components are not changing the prediction and detection of near-distance situations". Again, the aim is to reject the hypothesis and replace it with the alternative hypothesis, that MSIs calculated by datAcron components are improving the prediction and detection of near-distance situations. As in experiment 2, the hypothesis is split up, specified and tested in the following sub-hypotheses:

- $\overline{H0c}$: datAcron MSIs are not changing the predictability of near-distance situations.
- $\overline{H0d}$: datAcron MSIs are not changing the detectability of near-distance situations.
- $\overline{H0f}$: datAcron MSIs are not changing the average time between prediction and near-distance situation.
- $\overline{H0i}$: datAcron MSIs are not changing the confidence in the prediction of near-distance situations.
- $\overline{H0j}$: datAcron MSIs are not changing the confidence in the detection of near-distance situations.
- $\overline{H0k}$: datAcron MSIs are not changing the situational awareness of near-distance situations.

For testing these hypotheses, the same approach is used as for experiment 2. The difference during the execution of the experiment is related to the change of the experimental unit. In experiment 2 the experimental unit is composed of the MSIs, labelled by operators, the visualisation component of WP4 with disabled interactive functionalities and an operator who is tasked to predict and detect collisions. In experiment 3 the experimental unit consists of MSIs, calculated by the different datAcron components, the visualisation component of WP4 with enabled interactive functionalities and again an operator with the same task as in experiment 2. The interactive functionalities include the free use of layers, filters, zooming and panning. As for the interpretation of the MSI icons a familiarisation phase is given to the operators in order to get used to the interactive functionalities of the *IVA* component. By changing both the interactivity of the visualisation component and the source of MSIs, both the rejection and the failure of the rejection of the hypothesis is either due to one or due to the interaction of both changes of the experimental unit. The experimental design of experiment 3 uses a subset of the experimental design of experiment 2, more specifically, one of two scenario datasets with MSIs, randomly drawn:

- Scenario 2: rendezvous in minute 15 (5), near-collision in minute 25 (8.3).

For this, the MSIs displayed to the operator in scenario 2 are once the result of the labelling process of domain experts, in the following referred to as “labelled” situations, and once the detection results of datAcron components, in the following referred to as "detected" situations. All other specifications on experiment 2, especially assumptions, criteria and measures, data labelling, data collection and result interpretation are equally valid for experiment 3 which allows the interested reader to continue directly with the result analysis.

5.5.1 Analysis

Table 26 summarises the TP, FP and FN for predictions and detections of near-distance situations. As in experiment 2, FN refers to a missed event in the situational dataset, FP refers to a wrong prediction or detection of the operator and TP refers to the matching of the event in the situational dataset and the prediction or detection of the operator, as described in Section 5.4.7.

Table 26: Summary of results of experiment 3

	prediction			detection		
Situation	TP	FP	FN	TP	FP	FN
rendezvous labelled	2	-	1	2	-	1
Near-collision labelled	3	-	-	2	1	1
rendezvous detected	1	-	2	-	-	3
Near-collision detected	2	-	1	2	-	1

For both labelled and detected MSIs the prediction and detection of near-collisions is performed successfully in at least two out of three cases. The same observation is valid for the labelled rendezvous situations, which makes the true positive rate generalisable for both labelled situations. Only for the detected rendezvous, the results are worse in the sense that only one rendezvous out of three is correctly predicted and in the timespan after the occurrence of the rendezvous, no operator classified the situation as a rendezvous. With labelled data, the same situation is classified three times less FN. In the following, the collected data is used for testing the different proposed hypotheses including the effect of detected MSIs on the prediction and detection, on the average prediction time, on the confidence of predictions and detections, as well as on the situational awareness.

1. Effect of MSIs on prediction and detection

Table 27: $\overline{H0c, d}$ - True Positive, False Positive and False Negative Rates

	prediction				detection			
Situation	TPR	FPR	FNR	$\sum$	TPR	FPR	FNR	$\sum$
Near-distance situation labelled	0.83	-	0.17	1	0.57	0.14	0.29	1
Near-distance situation detected	0.5	-	0.5	1	0.33	-	0.66	1

Table 27 indicates lower true positive rates and higher false negative rates of operator assessments for situations with MSIs detected by datAcron components compared to situations with MSIs labelled by domain experts. This finding is observable both for the prediction and for the detection of near-distance situations. As shown in table 26 both the false predictions and the false detections of rendezvous in the detected situational dataset are the main driving factor of this conclusion.
Concluding the comparison of TP, FP and FN rates for near-distance situation prediction and

detection, the two hypotheses are rejected:

- $\overline{H0c}$: datAcron MSIs are not changing the predictability of near-distance situations. (rejected)
- $\overline{H0d}$: datAcron MSIs are not changing the detectability of near-distance situations. (rejected)

The observations allowing the proposition of the hypotheses rejection favour an alternative hypothesis to be tested which states the decrease of predictability and detectability of near-distance situations which are enriched by MSIs detected by datAcron components.

2. Effect of MSIs on the average prediction time

Table 28 describes the average time span between the TP prediction of the respective near-distance situation and its actual occurrence. As in experiment 2 the average time spans are based on the row "Time †to *" of tables 55, 56, 57:

Table 28: $\overline{H0f}$ - Average time between TP prediction and event occurrence.

	prediction
Situation	Average Time † to * (min)
Near-distance situations labelled	2.3
Near-collision situations detected	1.3

The time span is larger for near-distance situations enriched with MSIs labelled by domain experts. As the situational datasets are shown to the operators accelerated by factor 3, the 1 minute difference between the average prediction time spans correspond to 3 minutes real time. Concluding the comparison of the average time between the TP prediction of a near-distance situation and its actual occurrence, the hypothesis is rejected. The alternative hypothesis proposes a reducing effect of the detected MSIs on the time span between prediction and occurrence of the respective near-distance event.

- $\overline{H0f}$: datAcron MSIs are not changing the average time between prediction and near-distance situation. (rejected)

3. Effect of MSIs on the confidence of predictions and detections

Table 29: $\overline{H0i,j}$ - Average confidence in prediction and detection.

	prediction			detection		
Situation	TP	FP	FN	TP	FP	FN
Near-distance situations labelled	5	-	-	5	5	-
Near-distance situation detected	-	-	-	4	-	-

Given that only positive assertions are counted, table 29 does not allow a comparison of prediction confidence due to the lack of stated values. For all stated values the findings of experiment 2 are confirmed in the sense that for both positive true and false rated events the operators state a relatively high confidence value. For true positive detections, the labelled situations receive marginally higher confidence levels, which are not significant, given the small number of confidence values.

In conclusion, the proposed hypotheses cannot be rejected, which is again in line with the findings in experiments 2.

- $\overline{H0i}$: datAcron MSIs are not changing the confidence in the prediction of near-distance situations. (not rejected)
- $\overline{H0j}$: datAcron MSIs are not changing the confidence in the detection of near-distance situations. (not rejected)

4. Effect of MSIs on the situational awareness

For evaluating the effect of MSIs calculated by datAcron components on the situational awareness in maritime situations, the assertions of the operators are compared for the two versions of the situational dataset 2. The average values listed in table 30 state difference which are partly confirmed on the level of all operators:

- Instability of situation: The instability of the situation is rated once lower, once equal and once higher than the instability with detected MSIs.
- Complexity of situation: The complexity of situations with detected MSIs is rated to be lower by two operators than with labelled MSIs.
- Variability of situation: The variability of the situation is rated lower with detected MSIs by two out of three operators.
- Arousal: The arousal is rated to be equal by two operators and lower for the situations with detected MSIs by one operator.
- Concentration of Attention: Two operators out of three rate their concentration of attention to be lower for situations with detected MSIs, one operators attention is more concentrated for detected MSIs and the availability of interactive functionalities.
- Spare mental capacity: The spare mental capacity is rated to be equal by two and larger by one operator for detected MSIs.
- Information quantity: The information quantity is rated larger by two operators for the situations with detected MSIs and interactive functionalities, lower by one.
- Familiarity with situation: The familiarity with the situation is rated equal by two and higher by one operator for the situations with detected MSIs.

The availability of interactive functionalities and MSIs detected by datAcron components has a positive effect on the maritime situational awareness of operators. Especially, the information quantity is assessed to be larger, the situations are perceived to be less unstable, less complex and less variable. The operators feel less in the status of arousal, need to concentrate less and their attention is less divided. Further, the experts users estimate in average to have more spare mental capacity, even though experiment 3 is conducted consecutively to experiment 2 which would let assume that the spare mental capacity decreases. Thus, the proposed hypothesis can be rejected and replaced by the alternative hypothesis stating that interactive functionalities and MSIs detected by datAcron components increase the situational awareness. As discussed before, the average ratings are not representative for all operators. In comparison to experiment 2, the answers of the operators diverge more importantly.

- $\overline{H0k}$: datAcron MSIs are not changing the situational awareness of near-distance situations. (rejected)

Table 30: $\overline{H0k}$ - Average Situational awareness self assessment. Situational awareness rating cp. [25]: 1-Low, 7-High.

	exp. 2	exp. 3
Dimension of situational awareness cp. [25]	sc.2	$\overline{sc.2}$
Instability of Situation	5.2	4.5
Complexity of Situation	4.2	3.2
Variability of Situation	4.5	3.8
Arousal	5.2	4.8
Concentration of Attention	4.2	3.5
Division of Attention	4.2	3.8
Spare Mental Capacity	5	5.2
Information Quantity	4.5	5.2
Familiarity with Situation	5.8	6.2

5.5.2 Conclusions on Experiment 3

By comparing the results of experiment 3 with the results of experiment 2, the following conclusions can be drawn:

1. Comparing the results of experiment 2 and experiment 3 in table 26 it becomes apparent that both the predictions and the detections of the operators indicate to be better for situations enriched with MSIs labelled by domain experts than with MSIs detected by datAcron components. While the operators predict only 3 near-distance situations True Positive and 3 False Negative for MSIs based on datAcron component results, the manually labelled data of the corresponding datasets leads to 5 True positives and only 1 False Negative. For near-distance situation detections the operators detect only 2 situations True Positive but 4 False Negative. For the manually labelled data, also for the detection of near-distance situations, the manually labelled data leads to better performing operators reaching 4 True Positives, 1 False Positive and 2 False Negative.

2. For the correct prediction of near-distance situations, the MSIs detected by datAcron components seem to have a reducing effect on the time span between prediction and occurrence of the respective event.

3. No difference in the confidence level of operators situational assessment between labelled and detected MSIs can be stated, both because of the rough granularity of the confidence values scale (three values) and because of the small difference of the recorded values.

4. For the maritime situational awareness, the availability of interactive functionalities and MSIs detected by datAcron components have a positive effect. Especially, the information quantity is assessed to be larger, the situations are perceived to be less unstable, less complex and less variable. The operators feel less in the status of arousal, need to concentrate less and their attention is less divided. Further, the expert users estimate in average to have more spare mental capacity.

5. All experts understand and interpret the tugging icon correctly, presumably justifiable by its reference to the corresponding COLREGs day shapes or night lights [1].

6. Also, all experts judge the detection to be wrong, misleading or not corresponding to the vessel tracks indicated by the dynamic AIS data.

7. Perspective: Improve accuracy, precision and display of information.

5.6 Experiment 4: Expert accuracy MSI assessment

In this experiment, the objective was to reproduce a "live" assessment at the MSI level. To this end, the expert gathered all the results produced the datAcron components during the other experiments[7]. The expert then applied the methodology described in Section 4.1 and illustrated by Figure 18. To organise his evaluation the expert had access to all data layers (sea state, weather condition, geographical layers, ...) provided with the reference dataset. This makes a difference with respect to experts involved in the experiments 2 and 3 to whom only fishing areas, TSS, natura 2000 and recommended tracks were made available.

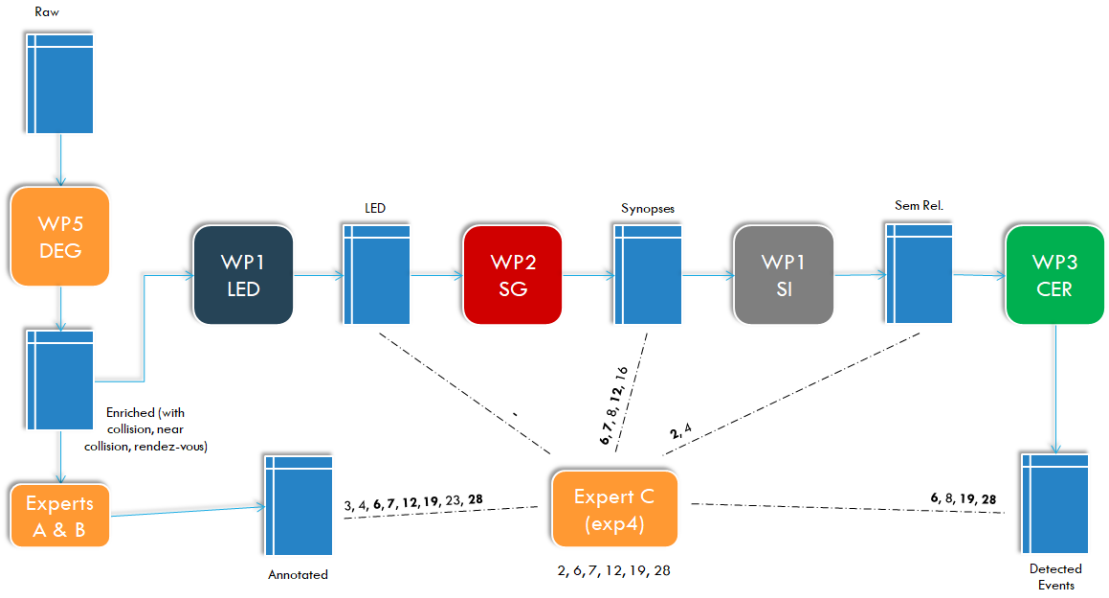


Figure 18: Assessments of MSI detection in Experiment 4

The assessment focused on the MSIs #2, #6, #7, #12, #19, #28 involved in the scenario and computed by *SG, SI, CER*. Detections of datAcron components were compared with raw, enriched and manually annotated data. Results of this comparison are reported in Figure 19. The expert also commented results as follows.

MSI #2 (Within an area) : "The result is good, all events are detected. It appears that SI component limited the computation of *within an area* of the TSS to the polygons corresponding to the separation zones. However from a maritime situation perspective, I was expecting a kind of bounding box around these polygons in order to include the channels of the TSS".

MSI #6 (Null speed) : "While the SG component gives some rather consistent results, we see that the CER does not return the same thing. The big difference I think is the fact that the *CER* has rejected all the detections that are in the port of Brest. As for the rest, none of the *CER* detections correspond to the annotations. But they are not so inconsistent. The thresholds chosen are probably not unrelated to this difference. Note that after the collision, the *CER* does not see the boats stopped, unlike the annotators".

MSI #7 / MSI #12 (Change of speed/course) : "We can notice a big difference between the detected events and the annotations, particularly for the change speed event. Indeed,

[7]Data collected are located in /datAcron/WP5/Data/Scenario Data (Demo 2)

	SI	CER			SG		
	Within an area MSI#02	Null speed MSI#06	Rendez-vous MSI#28	Underway MSI#19	Change of course MSI#12	Change of speed MSI#07	Stopped MSI#06
Message count in annotated dataset	4077	3035	3035	3035	3035	3035	3035
Annotated detections	521	871	34	2224	712	96	871
Datacron detection	523	29	1	431	907	598	204
True positives	521	0	1	422	396	33	171
False positives	2	29	0	9	511	565	33
True negatives	3554	2135	3000	802	1812	2374	2131
False negatives	0	871	33	1802	316	63	700
Detection probability	1	0,00	0,03	0,19	0,56	0,34	0,20
Missed detection probability	0	1,00	0,97	0,81	0,44	0,66	0,80
False alarm probability	0,004	1,00	0,50	0,02	0,56	0,94	0,16

Figure 19: Expert-based Comparison of datAcron Results vs. Enriched and Annotated Data

the change speed mainly corresponds to false alarms. We don't have any explanation for this difference".

MSI #19 (Underway) : "We notice that the detection of the underway event is done with a very few false alarms. But compared to the annotated dataset, the detection probability is only 20%".

MSI #28 (RDV) : "The choice was made by the *CER* developers to consider the "rendezvous" as a punctual event, contrary to the experts who annotated more data in case of proximity between the involved vessels. The conclusion is that the two rendezvous situations in the annotated dataset were correctly detected, while the apparent missed detection is very high".

Concluding remarks about the annotation :

The expert reported that results are sometimes very divergent from the annotations (*cf.* Section 5.4.5). In such a case it is difficult to get out numerical evaluations without asking additional questions regarding the subjectivity part of the annotators. The expert recognised the tedious work of annotation that has been done by the other experts (experts A and B in Figure 18). He noticed that similar portions of trajectories could receive different annotations, particularly when the interesting parameter was close to the threshold. In those particular tricky cases, the annotator can take different decisions, depending on his annotation experience, on his "mood" or focus capability. He concluded with few questions: "Can we be 100% confident in the expert truth, what is the best method to evaluate, what are the metrics, how many annotators are required, their level of expertise, etc ...?"

5.7 MSI robustness to reduced veracity

In this experiment, the reference is not some ground truth provided by experts, but the *CER* detections on the non-degraded dataset. Thus, the performances displayed in Figures 20 - 43 in Annex 8.1 are bounded and normalised by the best performance of the MSI detector on the reference dataset. This allows a quantitative analysis of the robustness of the MSI detectors to the lack of data. A proper combination with the (rather qualitative) accuracy assessment by expert provides a more complete assessment of the MSI detectors.

The sensitivity analysis investigates the impact of veracity variations on the performance of the Complex Event Recognition component. In the context of AIS, a typical example of a reduced veracity is the lack of transmitted or received AIS messages. Hence, the following analysis examines the impact of lack of data on the detection behaviour of the Complex Event Recognition component. The sensitivity analysis has been performed on results which were provided by Demokritos using datasets created by CMRE.

5.7.1 Experiment description

The reference dataset used was corresponding to scenarios 2 and 3. From the data degradation methods described in Deliverable [23], the data removal method was used for removing 10, 20 or 30% of the messages of each vessel in the non-degraded dataset. Each data removal was repeated 5 times. For each event results were averaged over the 5 repetitions and the two datasets.

5.7.2 Results

The experimental results are depicted in Annex 8.1 in Figures 20 to 43 for the MSIs "changingSpeed", "gap", "highSpeedNearCoast", "lowSpeed", "movementAbilityAffected", "movingSpeed", "sarCourse", "stopped", "tuggingSpeed", "underWay", "unusualSpeed", "withinArea".

- As expected, for most events detected by *CER*, the larger the data degradation, the lower Recall, Precision and F1 score.
- The total amount of True Positives diminishes stronger between 100% and 90% of data available than between 70% and 90% for 8 out of 12 events. Thus, already a small increase in the lack of data has a large impact on the *CER*.
- The most important observation: For all events False Positives are detected. For "ChangingSpeed" this behaviour is straightforward as the removal of one message does not impact on the reported speed before and after the removed message, hence leads only to a shift in the detection. For other events such as "stopped", "underWay" or "withinArea" the detections were not expected and reveal perspectives for further analysis.
- MovementAbilityAffected: In Figure 29 a large drop of Recall from 1 to 0 and Precision from ca. 0.74 to 0 are depicted for a reduction of ingested data from 80% to 70% of the original dataset. Due to missing True Positive Detections at 70% of data available as shown in 28, both Recall and Precision scores become zero and no F1 score can be calculated. The results require further analysis since the amount of True Positive detections for 80 and 90% are not supposed to be larger than the amount of True Positive detections for 100%.

5.7.3 Remarks

The event "movementAbilityAffected" is excluded from the analysis, as there are more True Positive detections reported for the degraded datasets than for the non-degraded dataset. This results requires further analysis. For the event "tuggingSpeed" no number of detected events in the non-degraded dataset is available. Further, no data is available for the degraded datasets for "loitering" (278), "atAnchorOrMoored" (1935), and "posAground" (1782) with the number of detections in the non-degraded dataset in brackets.

6 Main outcomes

The innovative experimental plan designed and implemented aimed at (1) considering the collaboration aspect of the project where components are developed in parallel and at different speed by different partners, (2) mitigating the risk linked to the late availability of the integrated prototype, (3) minimising the impact of the possible lack of accuracy of the MSIs detection and prediction by the datAcron components on user assessment results, and (4) focusing on operational uses of the datAcron prototype.

Different experimental units were selected for delivering representative results at all levels of system integration. Therefore, the performed decomposition of the evaluation of the datAcron prototype followed the semantic levels of functionality of under-MSI, MSI and scenario levels, involving:

- Maritime data, that are processed by the components;
- datAcron components, that form functionally independent units processing the maritime data;
- Maritime Situational Indicators (MSIs), that represent the result of the data processing displayed to operators;
- Maritime scenarios in which operators are fulfilling domain specific tasks like collision avoidance with the support of the datAcron prototype.

An evaluation space was defined to capture the results in a unified way, with the four dimensions of *big data challenges* (Volume, Velocity, Variety and Veracity), *evaluation criteria* (Timeliness, Scalability, Compression ratio, Classification quality, Clarity, Effectiveness) and *datAcron components*. The results presented in this document and captured in the evaluation space include both the outcome of Task 5.5 and all other maritime related evaluation activities performed under work packages 1, 2, 3 and 4.

For the maritime domain, the datAcron prototype was used as a decision support system for operators fulfilling a maritime surveillance task in which typically immediate interactions with the monitored vessels are required. Therefore, the documented experiments go beyond the analysis of pure computational aspects of datAcron by evaluating other human factor aspects such as situation awareness and the ability of the datAcron prototype to support maritime surveillance operations.

6.1 Performance of datAcron individual components

Major findings on the performance of the *SG*, *CER*, *CEF* and *FLP* components of the datAcron prototype in terms of timeliness, scalability, data compression and accuracy of MSIs detection, forecasting and prediction are:

- Synopses Generator (SG):
  - Strength: For smaller datasets the latency is reduced to 116 ms and to 923 ms for larger datasets. For datasets with higher transmission frequencies an average distance error of ca. 20-75 m is reached for a compression ratio of ca. 75-80%.
  - Weakness: Datasets with lower transmission frequencies yield average distance error of ca. 65-465 m for a compression ratio of ca. 25-74%.

- Complex Event Recognition (CER):
  - Strength: The *CER* recognises the majority of movement patterns of vessels equally well for compressed as for non-compressed data. This promotes the combined application of *SG* and *CER* for the respective MSIs detection. An average recognition time of 1.8 seconds for the tested NARI dataset, which covers a much larger area than a typical zone of surveillance competence, demonstrates the necessary performance for supporting the fulfilment of tasks like collision avoidance.
  - Weakness: The *CER* does not take into account available data sources like weather conditions and ocean conditions, e.g. wind force, currency or wave height, which may have an important impact on the vessels kinematics.
- Future Location Predictor (FLP):
  - Strength: The results on the *FLP* suggest a potential future integration in a component yielding information with a higher level of semantics, like collision alert, time to expected collision or closest point of approach.
  - Weakness: The current maximum prediction horizon of the short-term location prediction of 5 min is too small to stop common vessel types in time, like tankers, bulk carriers and container vessels with maximum speed below 25 knots.
- Complex Event Forecasting (CEF):
  - Strength: The *CEF* delivers forecasts with high prediction rates of 80-98% for NARI and IMISG dataset, both for forecasting vessels entering the harbour and starting fishing. Additionally, the large latency is compensated by larger forecasting horizons which are in average twice as large for IMISG.
  - Weaknesses: The forecasts of the *CEF* cannot be used to define when an event is about to happen, given that *CEF* has a spread which is typically twice as large, as the time from event prediction to event occurrence.

6.2 Expert-based accuracy assessment of MSI detections

The major findings regarding the expert-based accuracy assessment of MSI detections for the *SG* and the *CER* components are:

- During the first period of assessment:
  - The expert highlighted some false detections of stop events (MSI #6) from the *SG* which happened to be either too numerous in a area where stops are not expected (i.e., in the TSS) or in a straight line. Moreover, it was shown that the errors have been propagated to the *CER* which processed outputs of the *SG*.
  - The large majority of underway events (MSI #19) were correctly detected (i.e., compatible with the speed reported by the AIS), while still some inconsistent detections were reported by the *CER*, especially close to ports. Several ships sailing on the maritime route were not detected as underway, while they are clearly matched with the event detection's criteria.
  - The expert analysis of the *SG* outputs allowed to highlight that the extreme speed values were over-represented compared to the original AIS dataset. Indeed, the synopses generator is re-labelling the reported speed, instead of filtering out irrelevant AIS contacts as it was assumed by *CER*, therefore it could not be safely applied as a pre-processing step for event detection.

The feedback provided to the component designers was considered in the improvement of the next versions of the SG and *CER* components.

- During the second period of assessment, only the accuracy of the *CER* component was assessed:
  - MSI #20 (Vessel at anchor or moored) - All detected MSI #20 have a speed consistent with the event (assumed good detection) but some events were probably missed. However, some detected MSI #20 actually correspond to almost stopped fishing vessels working with fish trap or net, which appear to be wrong detection.
  - MSI #6 (Null speed or stopped) - While few MSI #6 detections are slightly over the speed threshold in open seas, detections in the vicinity of ports are consistent with the speed threshold and can be considered as correct detections.
  - MSI #11 (Speed not matching user-defined threshold – set to 4 knots)- Good detection rate (true positive) for vessels having a speed below or above 4 knots. The detection rate of about 10% versus the reference dataset is consistent with compression ratios of the *SG* component.
  - MSI #8 (High speed near coast) - Exhibits generally a good detection rate.
  - MSI #2 (Within an area - JRC fishing area + Natura 2000) - This combination of areas allows to detect possible illegal fishing activities. Most of the events were correctly detected although some missed detections and false detections were observed.
  - MSI #24 (Tugging) - Among three tugging events in the area, the *CER* component correctly detected two events corresponding to (1) a vessel in distress and (2) a regular tugging from an anchor area to the port.
  - MSI #3 (On a maritime route) - 62% of vessels have been correctly assigned to the maritime route they actually followed. This quantitative assessment has been possible thanks to a prior labelling of data by experts. Beyond the result, the labelling method was tested.

About the expert-based accuracy assessment methodology, the major findings are:

- The expert-based assessment combining qualitative and quantitative evaluation was very helpful all along the project for the definition or improvements of *SG* and *CER*, related MSIs or associated thresholds;
- The methodology is limited by the subjectivity of the experts and their difficulty to process very large volume of data. It that sense, such assessment method would require additional investigation to further automate and guide the analysis of data;
- The annotation process is a tedious work, rather limited to small amount of data, that requires structured guidelines and thresholds to be efficient.

6.3 Robustness of CER to veracity degradation

The experiment designed to capture and quantify the ability of the *CER* component to cope with missing data allowed for the following conclusions:

- As expected, the accuracy (measured as Recall, Precision and F1 score) of the MSIs detected by the *CER* component decrease when the ratio of missing data increases from 0% to 30%;
- Most of the MSI detectors are quite robust to missing data as a drop of less than 5% in the F1 score is observed (e.g., “WithinArea”, “UnderWay”);

- Some MSI detectors however are sensitive to missing data, some exhibiting large drops of performance (e.g., "Stopped", "MovementAbilityAffected", "AIS gap", "ChangingSpeed");
- For 8 MSIs over 12, the True Positive detections decrease faster between 100% and 90% of ratio of data available than between 70% and 90%, meaning a quite high impact of a small decrease of amount of data on some of the *CER* detectors.

6.4 Situation awareness with datAcron prototype

Major findings regarding the operators effectiveness and the clarity of the datAcron prototype encompass:

- MSIs improve the *prediction* and the *detection* of both collisions and near-distance situations (e.g., tugging, rendez-vous);
- MSIs are not extending the time between the correct prediction of a near distance situation and its occurrence compared to solely displaying raw AIS data;
- MSIs are not changing the *confidence* of users in their prediction and detection of near-distance situations;
- Compared to the exclusive display of raw AIS data the MSIs are changing the *situational awareness* of operators in two ways: Firstly and beneficially, MSIs increase the *perceived information quantity* of situations and operators declare being more focused when MSIs are displayed. Secondly and adversely, MSIs increase also the *perceived complexity and variability* of situations and tend to reduce the *spare mental capacity* of operators;
- The method used for data creation and labelling is tested and found to be suitable for the creation of near-distance situations, as the variations introduced by the method are perceived by the operators to be smaller than the differences induced by the MSIs.

Comparing the situation assessments by operators with support of MSIs either detected by datAcron components (possible false detection) or manually labelled ("true detection"), has led to the following conclusions, with possible limited significance due to the small sample size:

- Both the users' predictions and the detections of events of interest were better when based on MSIs labelled by domain experts than when based on MSIs automatically detected by datAcron components.
- In case of a correct prediction of near-distance situations by experts, the MSIs detected by datAcron components seem to reduce the time span between prediction and occurrence of the respective event (i.e., reaction time);
- Assessing the situation based on manually labelled and datAcron detected MSIs the origin of the MSIs does not seem to have an impact on the confidence level reported by the experts. However, the small number of samples and the scale of confidence records do not allow to draw any further conclusion;
- The availability of interactive functionalities and MSIs detected by datAcron components have a positive effect on the maritime situational awareness. Especially, the information quantity is assessed to be larger, the situations are perceived to be less unstable, less complex and less variable. The operators feel less in the status of arousal, need to concentrate less and their attention is less divided. Further, the experts users estimate in average to have more spare mental capacity.

- All experts understand and interpret the tugging icon correctly, presumably justifiable by its reference to the corresponding COLREGs day shapes of night lights;
- Finally, all operators judge the tugging event displayed in experiment 3 to to be wrong, misleading or not corresponding to the vessel tracks indicated by the dynamic AIS data. In experiment 3, those MSIs were displayed which were detected by datAcron components.

6.5 Conclusions about the methodology

- **The list of MSIs -** The list of MSIs introduced at the beginning of the project played a pivotal role in the datAcron prototype setup to the maritime use case all along the project: On the one hand, it drove the design of datAcron components (especially the *CER*, *CEF* and *LED*) and provided semantics to the *SG*, and on the other hand it drove the design of the experimental plan by clarifying the roles of maritime surveillance experts in the evaluation process (either expert for accuracy assessment or operator for experimenting the prototype).

  Although this list represent a non exhaustive set of indicators of possible interest for operational purposes, it originates from a literature survey of studies themselves recording basic maritime events of interest maritime security and safety. Moreover, it was validated independently by marine officers as relevant indicators.

  The experiments conducted highlighted that although the correct detections of MSIs displayed in real time to the operators neither decreases the operator's instant of collision detection nor increases his confidence in detection, they improve the prediction and the detection of both collisions and near-distance situations (e.g., tugging, rendez-vous), compared to raw AIS positions. Moreover, the MSIs increase the information quantity perceived by the operators who declare being more focused when MSIs are displayed. Finally, MSIs increase the perceived complexity and variability of situations and tend to reduce the spare mental capacity of operators.

- **Accuracy assessment methodology -** Datasets of real events labelled by experts do not exist and the preparation of such datasets was out of the scope of the project. The main reason is the lack of availability for a long period of maritime surveillance experts to perform such a tedious task, that a few of them would agree to do. To overcome this issue while focusing on the processing of real data (compared to simulated ones), we developed some methodologies which allowed to assess the accuracy of the MSIs detected by the datAcron components. Firstly, a maritime surveillance expert validated the detections of the *SG* and *CER* components in qualitative way using GIS software. Secondly, we developed some functions to inject some events of interest in the reference dataset under study. The resulting dataset can be named "pseudo-synthetic" as the events were either real events shifted in time and space to match the area and period of interest or were derived from statistics of the reference dataset. This method ensures (1) to keep the realism very high (compared to purely simulated events) and (2) to produce "ground truthed" dataset. Thirdly and finally, an expert labelled a small sample of AIS positions for a specific MSI, that was then used to compare to the automatic detection.

- **Evaluation space -** In order to structure and capture the individual evaluations of the different datAcron components, performed either at the workpackage level by the components designer or at WP5 level by independent assessments involving maritime surveillance experts, we defined an evaluation space. This evaluation space is a three dimensional space encompasses *big data dimensions*, *evaluation criteria* and *datAcron components*. Within this space, the different evaluation activities have been captured and summarised in a unified way. This space then defined was obviously too large for an exhaustive coverage (i.e.,

each component subset of components assessed along each criterion with each big data challenge). However, the set of evaluation performed happen to cover quite uniformally the evaluation space. For instance, the SI addresses the variety challenge, the *SG* the volume and velocity, the *CER* the veracity.

For the specific purpose of the validation of the datAcron prototype by maritime surveillance operators (i.e., scenario level assessment), the experimental plan has been designed to suit the collaborative aspect of this research project where components were developed and tested independently the different partners under the different workpackages. In particular, the experimental plan mitigated the lack of availability of all datAcron components (and the integrated prototype). To this end, we used a serious gaming methodology to assess the impact of the different information layers (variety challenge) on operators belief assessment, which did not require any numerical support of the datAcron prototype but relied on a board version of it. Also, a proper partitioning of the space fixing some dimensions avoided the double-counting of errors in the assessment of sequential datAcron components (especially the possible lack of accuracy): A proper preparation of the dataset allowed to display correct detections of MSIs to the operators, while the accuracy of the MSIs has been assessed independently. An experiment with MSIs detected by the datAcron components was still performed.

- **Gaming approach -** The gaming approach followed for assessing the impact of information variety on human belief assessment allowed to mitigate the risk of lack of availability of the datAcron prototype to conduct experiments, while decoupling the information content of information layers (context, MSIs, raw AIS) from the outcome of the datAcron components and from the visualisation tool.

6.6 Evaluation perspectives

Given that all experiments at the scenario level shared some pieces of their design, additional conclusions can be drawn by combining the results of the different experiments. Possible research questions that could be answered in the future are:

Based on the data collected in the context of experiment 1, MSIs can be ordered according to their perceived relevance. By combining this finding with the result of experiment 2 it may become verifiable whether MSIs perceived as more relevant are more helpful for predicting and detecting near-distance situations.

By combining and comparing the results of experiment 2 and 3 it may become verifiable how MSI icons impact the interpretation of the situation and to improve MSI icons so that the icons enrich the AIS information in a general sense or with respect to a specific task like collision avoidance.

Further comparison are possible between the results of experiment 3 and 4. Here, it can be investigated how false positive detections of datAcron components, identified in experiment 4, may impact the situational awareness in experiment 3 in order to give guidelines for calibrating the detection thresholds.

7 Conclusions

This deliverable reports the results of Task 5.6 on final evaluation and validation of the datAcron prototype on a maritime use case, and concludes the activities of Work Package 5 on the Maritime Use Case.

While the problem complexity is drastically reduced by the computation of synopses (critical basic events used also to reconstruct AIS trajectories), their semantic counter-parts (named Maritime Situational Indicators) appear to be be relevant to solve some operational problems of maritime surveillance (i.e., collision avoidance).

The accuracy of MSIs detection still exhibits room for improvement, as the automatic detection of MSIs by the synopses generator happens to miss some events or to miss-detected others. However, this accuracy assessment is itself very challenging as it relies on expert manual assessment of the same critical events to which *SG* detections can be compared. The detection of MSIs by experts of maritime surveillance based on raw AIS data is a very tedious task, in some case subjective and very time consuming. The task is furthermore complicated by the freedom of movements of vessels, the lack of veracity of AIS data (partial coverage of receivers, deliberate partial emission of AIS data from vessels). Consequently, a quantitative and precise assessment of MSIs accuracy on real AIS data remains an open research question.

As far as we know, no operational system of maritime surveillance implements such an extensive list of MSI detectors and predictors. In this respect, the datAcron prototype positions itself as a precursor of future maritime surveillance systems.

The experiments run with maritime surveillance experts fulfilling a task of collision avoidance demonstrated the added value of the MSIs to their situation awareness. Specifically, compared to raw AIS data, the MSIs are helpful to draw the user attention and in some cases to discriminate between similar events.

8 Annex

8.1 Robustness to missing data of the CER

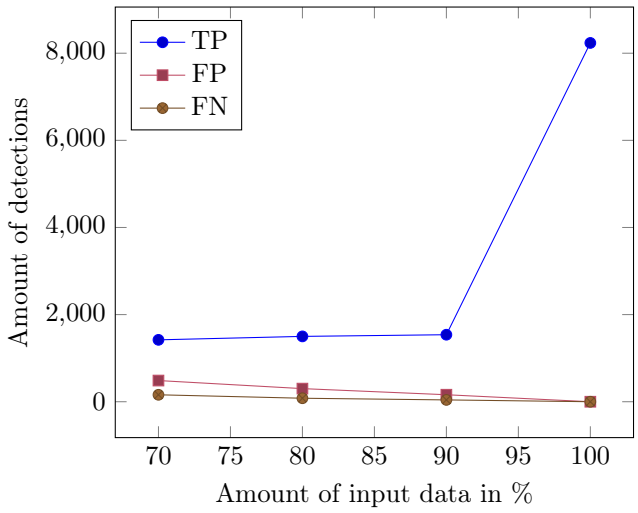


Figure 20: Impact of data degradation on the True Positive, False Positive and False Negative detections of "changingSpeed"

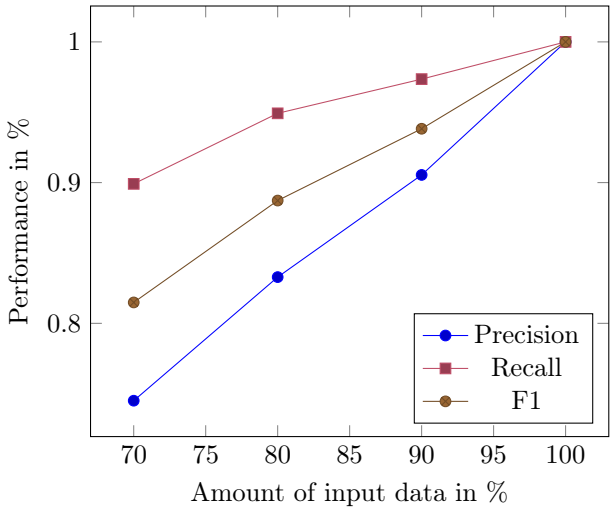


Figure 21: Impact of data degradation on the detection of "ChangingSpeed"

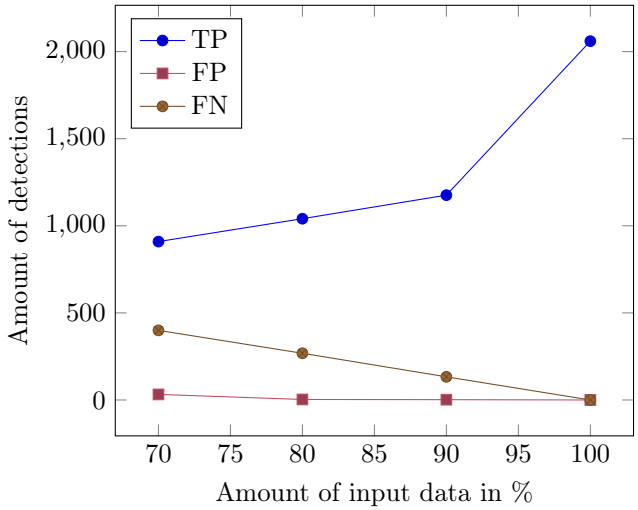


Figure 22: Impact of data degradation on the True Positive, False Positive and False Negative detections of "gap"

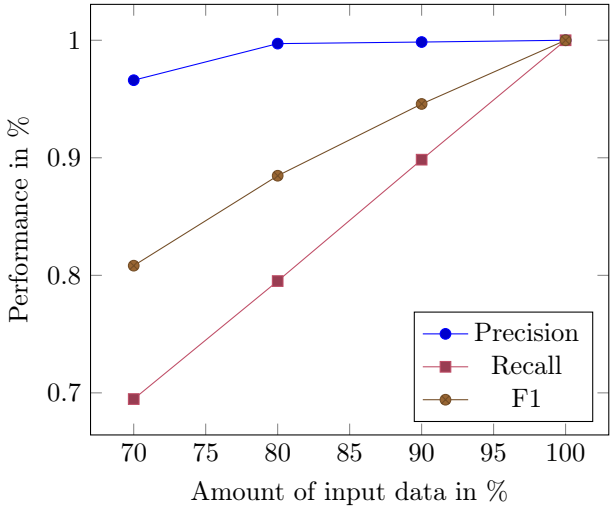


Figure 23: Impact of data degradation on the detection of "gap"

8.2 Capture of results in the evaluation framework

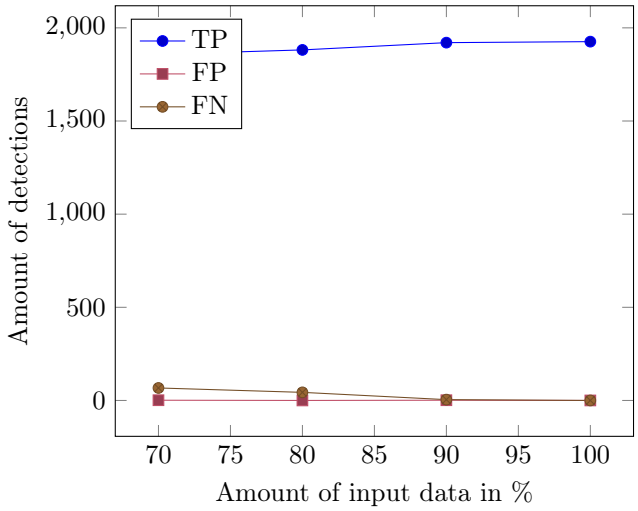


Figure 24: Impact of data degradation on the True Positive, False Positive and False Negative detections of "highSpeedNearCoast"

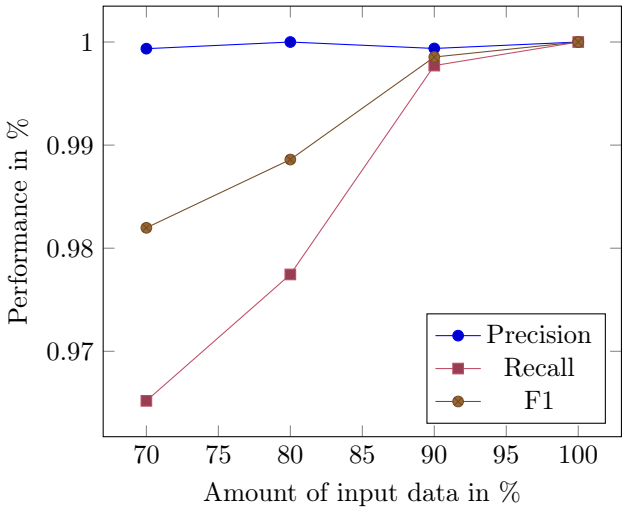


Figure 25: Impact of data degradation on the detection of "highSpeedNearCoast"

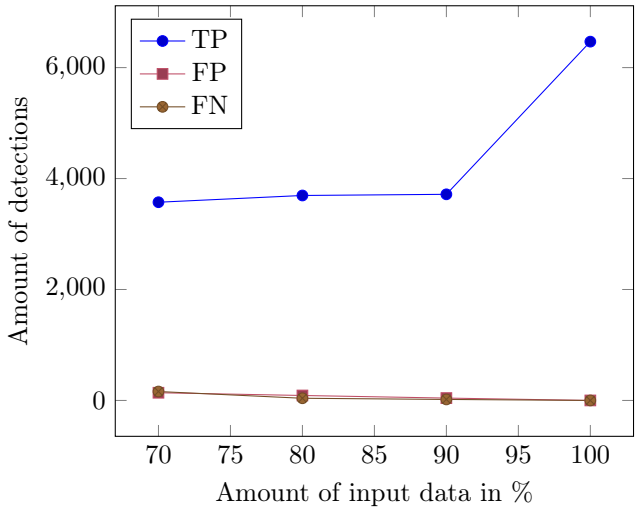


Figure 26: Impact of data degradation on the True Positive, False Positive and False Negative detections of "lowSpeed"

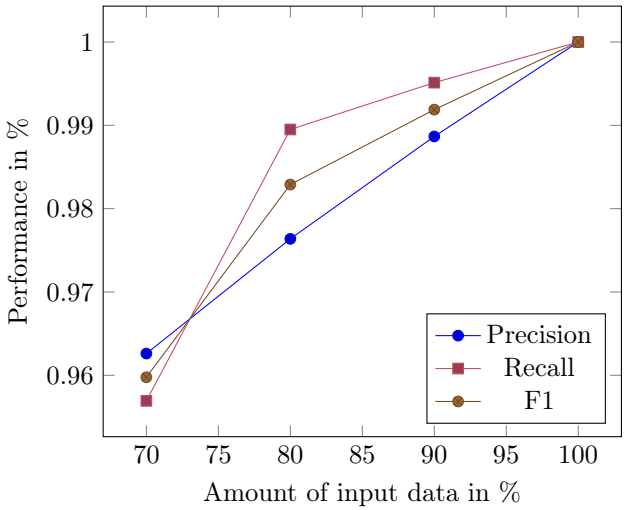


Figure 27: Impact of data degradation on the detection of "lowSpeed"

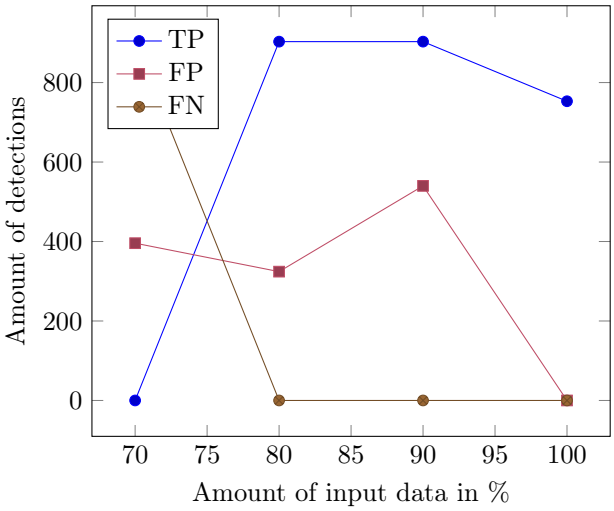


Figure 28: Impact of data degradation on the True Positive, False Positive and False Negative detections of "movementAbilityAffected"

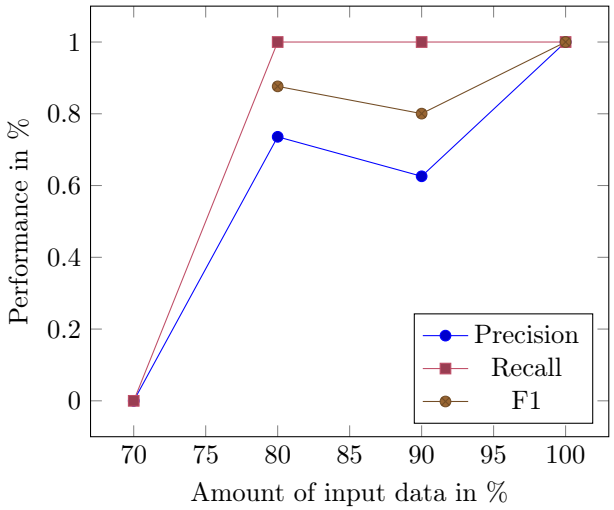


Figure 29: Impact of data degradation on the detection of "movementAbilityAffected"

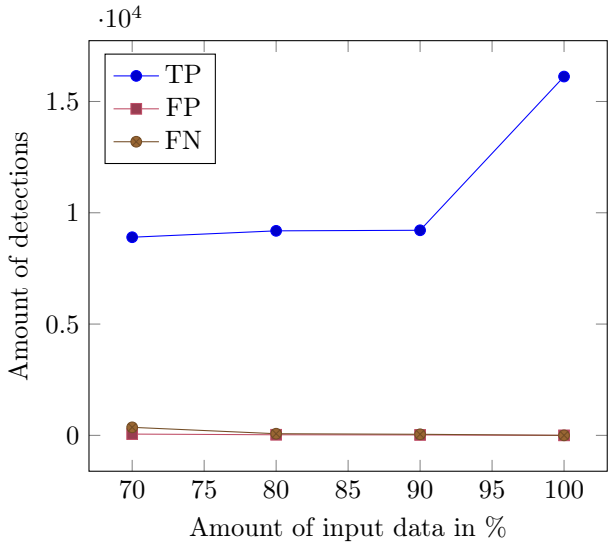


Figure 30: Impact of data degradation on the True Positive, False Positive and False Negative detections of "movingSpeed"

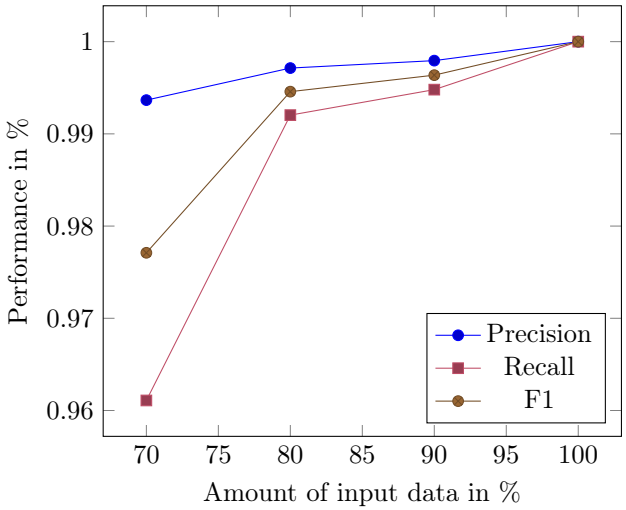


Figure 31: Impact of data degradation on the detection of "movingSpeed"

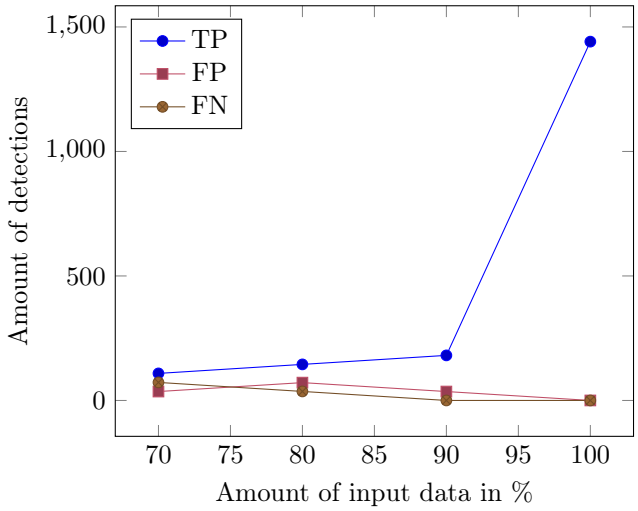


Figure 32: Impact of data degradation on the True Positive, False Positive and False Negative detections of "sarCourse"

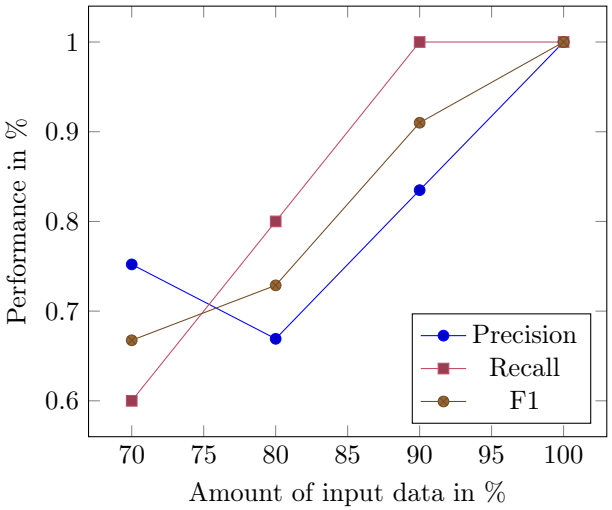


Figure 33: Impact of data degradation on the detection of "sarCourse"

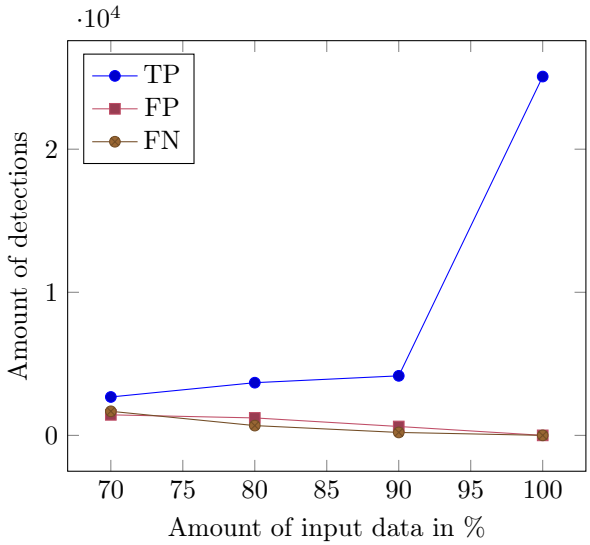


Figure 34: Impact of data degradation on the True Positive, False Positive and False Negative detections of "stopped"

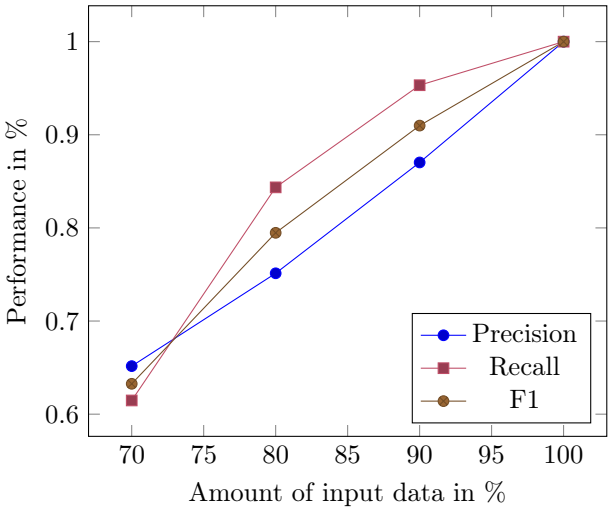


Figure 35: Impact of data degradation on the detection of "stopped"

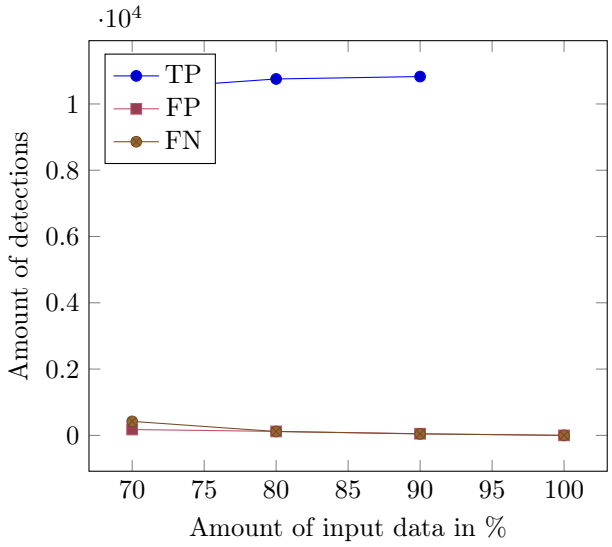


Figure 36: Impact of data degradation on the True Positive, False Positive and False Negative detections of "tuggingSpeed"

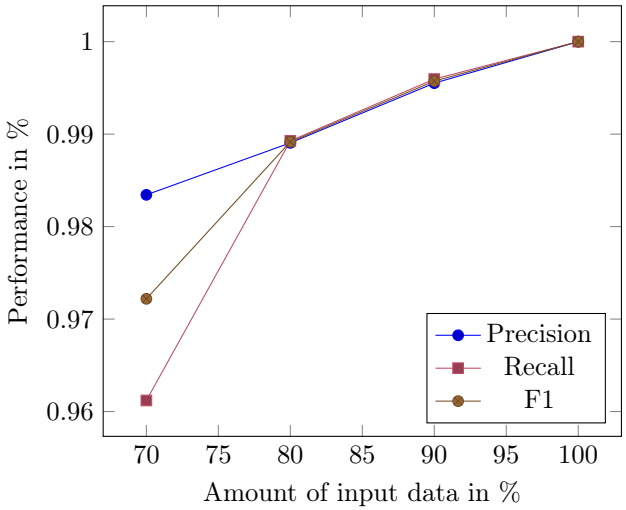


Figure 37: Impact of data degradation on the detection of "tuggingSpeed"

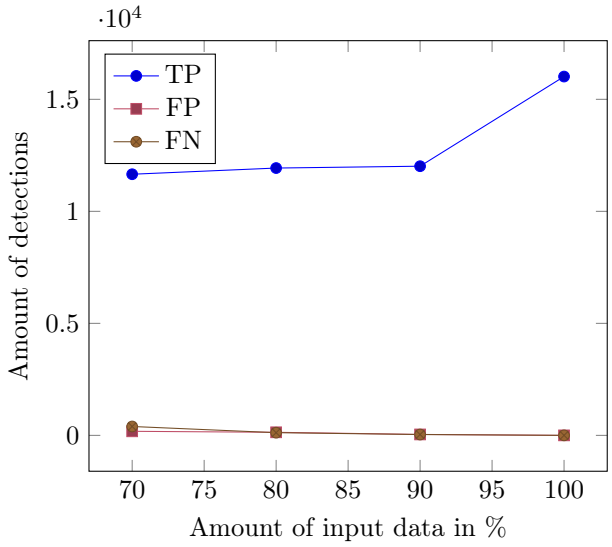


Figure 38: Impact of data degradation on the True Positive, False Positive and False Negative detections of "underWay"

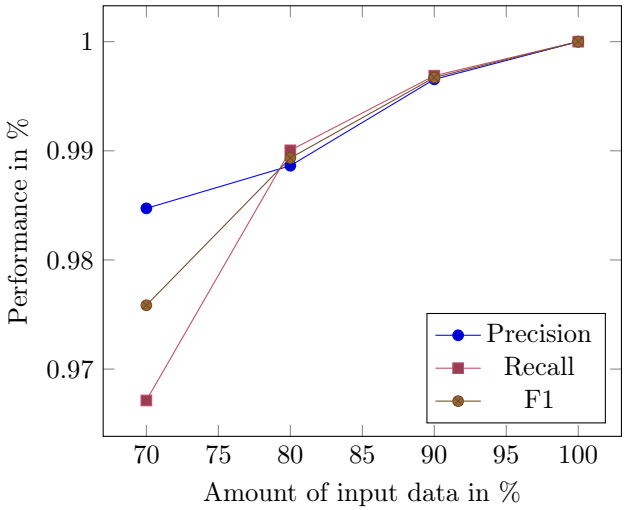


Figure 39: Impact of data degradation on the detection of "underWay"

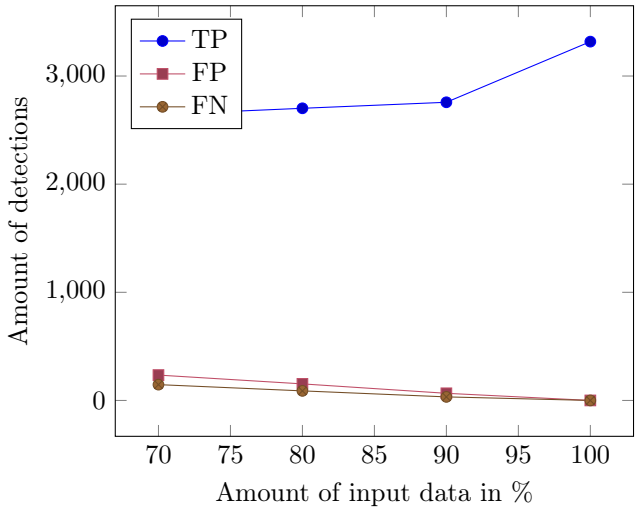


Figure 40: Impact of data degradation on the True Positive, False Positive and False Negative detections of "unusualSpeed"

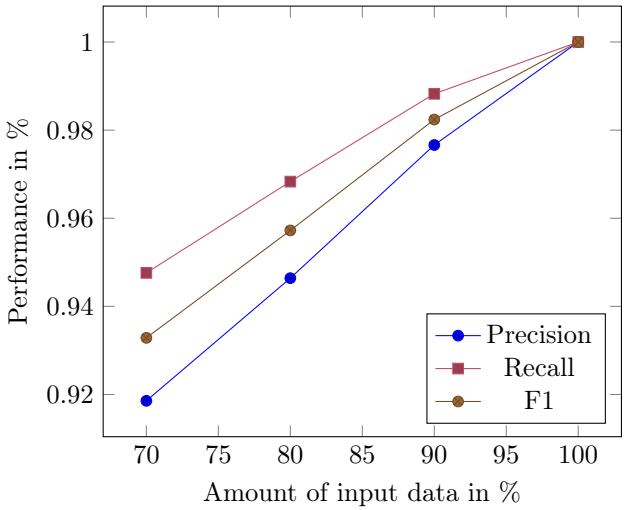


Figure 41: Impact of data degradation on the detection of "unusualSpeed"

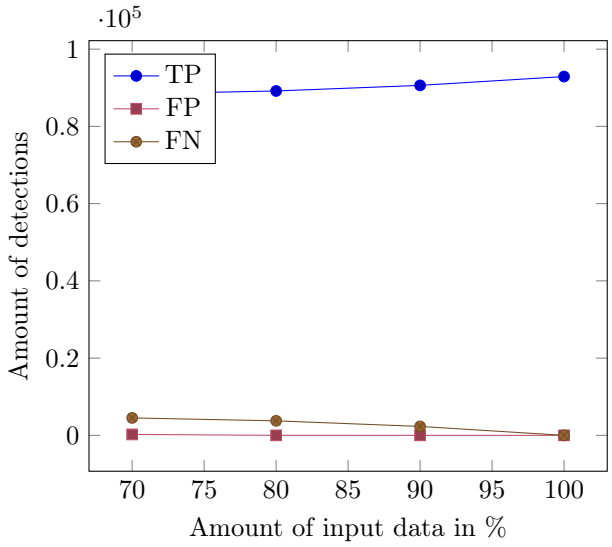


Figure 42: Impact of data degradation on the True Positive, False Positive and False Negative detections of "withinArea"

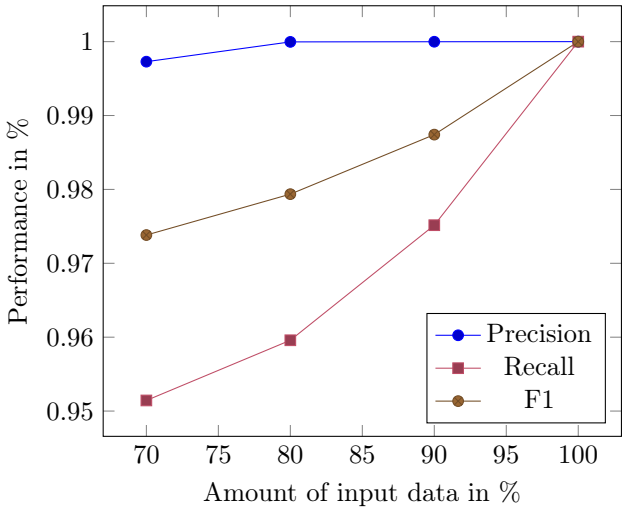


Figure 43: Impact of data degradation on the detection of "withinArea"

Table 31: SG mapping of deliverables to evaluation framework.

	data variation	volume and velocity																									
	dataset	$NARI_{SGv0.7/v0.8}$										$IMISG_{SGv0.7/v0.8}$										$\overline{IMISG}_{SGv0.7}$					
	$\triangle\Theta$	2.5	5						7.5	10		2.5	5						7.5	10		5					
	$\triangle T$	30	10	15	30			60	30	30		30	10	15	30			60	30	30		30					
	# threads/nodes	1	1	1	1	2	4	8	1	1	1	1	1	1	1	2	4	8	1	1	1	1	2	4	8		
Criterion	Compression ratio	D2.1, p.62ff, D2.3, p.86ff: [MSI#6,7,12,16]																				D2.1, p.62ff: [MSI#6,7,12,16]					
	Latency																										
	Throughput																										

Table 32: SG v0.7 - Average over [MSI#6,7,12,16] and slow motion for $NARI_{SGv0.7}$, $IMISG_{SGv0.7}$ and $\overline{IMISG}_{SGv0.7}$ (D2.1).

	data variation	volume and velocity																		
	dataset	$NARI_{SGv0.7}$									$IMISG_{SGv0.7}$									
	$\triangle\Theta$	2.5	5						7.5	10	2.5	5							7.5	10
	$\triangle T$	30	10	15	30			60	30	30	30	10	15	30			60	30	30	
	# threads	1	1	1	1	2	4	8	1	1	1	1	1	1	2	4	8	1	1	1
Criterion	Compr. ratio	0.67	0.74	0.74	0.74	-	-	-	0.75	0.78	0.8	0.73	0.22	0.47	0.73	-	-	-	0.74	0.74	0.74
	Latency	216	-	-	238	-	-	-	-	302	345	2700	-	-	2785	-	-	-	-	2900	2900
	Throughput	15345	-	-	16900	-	-	-	-	15690	14480	19655	-	-	19485	-	-	-	-	20000	20500
	RMSE	7	10	10	18	-	-	-	40	25	36	150	30	70	257	-	-	-	640	335	414

	data variation	volume and velocity								
	dataset	$\overline{IMISG}_{SGv0.7}$								
	delta Theta	2.5	5					7.5	10	
	delta T	30	10	15	30		60	30	30	
	# threads	1	1	1	1	2	4	8	1	1	1
Criterion	Compr. ratio	0.99	-	-	0.99	-	-	-	-	0.99	0.99
	Latency	-	-	-	11400	1800	1285	1285	-	-	-
	Throughput	-	-	-	15035	17145	23760	40900	-	-	-
	RMSE	28	-	-	35	-	-	-	-	79	95

Table 33: SG v0.8 - Average over [MSI#6,7,12,16] and slow motion for $NARI_{SGv0.8}$ and $IMISG_{SGv0.8}$ (D2.3).

	data variation	volume and velocity																		
	dataset	$NARI_{SGv0.8}$									$IMISG_{SGv0.8}$									
	$\triangle\Theta$	2.5	5					7.5	10		2.5	5						7.5	10	
	$\triangle T$	30	10	15	30			60	30	30	30	10	15	30			60	30	30	
	# threads	1	1	1	1	2	4	8	1	1	1	1	1	1	2	4	8	1	1	1
Criterion	Compr. ratio	0.68	0.75	0.75	0.75	-	-	-	0.75	0.775	0.795	0.7	0.25	0.46	0.72	-	-	-	0.74	0.73	0.74
	Latency	-	-	-	116	52	30	16	-	-	-	-	-	-	923	540	186	36	-	-	-
	Throughput	-	-	-	8545	14000	23090	38180	-	-	-	-	-	-	11455	21000	56000	126000	-	-	
	RMSE	11.5	18.5	23	26.5	-	-	-	50.5	51.5	77	163	64	96	276	-	-	-	754	380	466

data variation
delta Theta: threshold for change in heading.
delta T: threshold for gap start.
#threads: number of threads or nodes used for event processing.
Compression ratio.
Latency: time in SG pipeline, measured in milliseconds [5], p.86ff.
Throughput: number of messages processed per second [5], p.86ff.
RMSE: Root mean squared error, based on Haversian distance, measured in meter, including critical points.

Table 34: CER mapping of deliverables to evaluation framework.

	data variation	volume								variety/veracity																	
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	D3.2, p.10ff:																									
	Average recognition time (sec)	[MSI#2,6,8,19,26,28]								D3.4, p.28ff:																	
	Recall									[MSI#2,6,8,9,19-28]																	
	Precision																										
	F1																										

Table 35: CER - Average over [MSI#2,6,8,19,26,28] and [MSI#2,5,6,8,9,19-28].

	data variation	volume								variety/veracity																	
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput	0.3	0.4	0.5	1.48	0.21	0.28	0.3	0.35																				
	Avg. rec. time	10	8	6	2	22	16	15	12	0.1	0.2	0.6	1.8	-	80	180	250	495	-	0.2	0.7	1.85	5	-	180	270	450	900	-
	Recall									-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Precision									-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	F1									-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 36: CER - [MSI#2]: Within a given area.

	data variation	volume																													
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24		
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																						
	Average recognition time (sec)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Recall													1					1					(1)							(1)
	Precision													1					1					(1)							(1)
	F1													1					1					(1)							(1)

Table 37: CER - [MSI#4]: Proximity to other vessels.

	data variation	volume																		variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																				
	Average recognition time (sec)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Recall													-					-					-					-
	Precision													-					-					-					-
	F1													-					-					-					-

Table 38: CER - [MSI#5]: In stationary area.

	data variation	volume																				variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24		
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																						
	Average recognition time (sec)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Recall													-					-					-					-		
	Precision													-					-					-					-		
	F1													-					-					-					-		

Table 39: CER - [MSI#6]: Null speed (stopped).

	data variation	volume																		variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																				
	Average recognition time (sec)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Recall													1					1					(1)					(1)
	Precision													1					1					(1)					(1)
	F1													1					1					(1)					(1)

Table 40: CER - [MSI#8]: Mismatch speed area, here high speed near coast (highSpeedNC).

	data variation	volume																				variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24		
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																						
	Average recognition time (sec)	-	-	-	-	-	-	-	-	0.09	0.13	0.22	0.36	-	-	-	-	-	-	0.2	0.36	0.6	1.2	-	-	-	-	-	-		
	Recall													0.911					-					(1)					-		
	Precision													1					-					(1)					-		
	F1													0.953					-					(1)					-		

Table 41: CER - [MSI#9]: Mismatch speed vessel type (vesselIST).

	data variation	volume																		variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																				
	Average recognition time (sec)	-	-	-	-	-	-	-	-	0.09	0.13	0.22	0.44	-	-	-	-	-	-	0.27	0.4	0.71	0.13	-	-	-	-	-	-
	Recall													0.936					0.973					(1)					(1)
	Precision													0.938					0.975					(1)					(1)
	F1													0.937					0.974					(1)					(1)

Table 42: CER - [MSI#19]: Under way.

	data variation	volume																		variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																				
	Average recognition time (sec)	-	-	-	-	-	-	-	-	0.09	0.13	0.22	0.36	-	-	-	-	-	-	0.18	0.35	0.58	1.1	-	-	-	-	-	-
	Recall													0.995					0.992					(1)					(1)
	Precision													0.999					0.996					(1)					(1)
	F1													0.997					0.994					(1)					(1)

Table 43: CER - [MSI#20]: At anchor or moored. Avg. recognition time as minimum of anchored and moored.

	data variation	volume																		variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																				
	Average recognition time (sec)	-	-	-	-	-	-	-	-	0.09	0.13	0.22	0.31	-	-	-	-	-	-	0.22	0.36	0.62	1.11	-	-	-	-	-	-
	Recall													1					1					(1)					(1)
	Precision													1					1					(1)					(1)
	F1													1					1					(1)					(1)

Table 44: CER - [MSI#21]: Movement ability affected (maa).

	data variation	volume																		variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																				
	Average recognition time (sec)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Recall													0.989					0.716					(1)					(1)
	Precision													0.987					0.986					(1)					(1)
	F1													0.988					0.830					(1)					(1)

Table 45: CER - [MSI#22]: Aground.

	data variation	volume																		variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																				
	Average recognition time (sec)	-	-	-	-	-	-	-	-	0.09	0.13	0.18	0.36	-	-	-	-	-	-	0.22	0.36	0.62	1.2	-	-	-	-	-	-
	Recall													1					-					(1)					-
	Precision													1					-					(1)					-
	F1													1					-					(1)					-

Table 46: CER - [MSI#23]: Engaged in fishing, here trawling.

	data variation	volume																				variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24		
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																						
	Average recognition time (sec)	-	-	-	-	-	-	-	-	0.09	0.18	0.36	0.71	-	-	-	-	-	-	0.22	0.4	0.8	1.56	-	-	-	-	-	-		
	Recall													0.997					0.996					(1)					(1)		
	Precision													0.992					0.953					(1)					(1)		
	F1													0.994					0.974					(1)					(1)		

Table 47: CER - [MSI#24]: Tugging.

	data variation	volume																		variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																				
	Average recognition time (sec)	-	-	-	-	-	-	-	-	0.09	0.13	0.22	0.44	-	-	-	-	-	-	0.22	0.4	0.98	3.29	-	-	-	-	-	-
	Recall													0.879					0.881					(1)					(1)
	Precision													0.983					0.993					(1)					(1)
	F1													0.928					0.934					(1)					(1)

Table 48: CER - [MSI#25]: In SAR operation (inSAR).

	data variation	volume																		variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																				
	Average recognition time (sec)	-	-	-	-	-	-	-	-	0.09	0.13	0.31	0.8	-	-	-	-	-	-	0.22	0.36	0.71	1.51	-	-	-	-	-	-
	Recall													0.999					0.998					(1)					(1)
	Precision													0.998					0.993					(1)					(1)
	F1													0.999					0.993					(1)					(1)

Table 49: CER - [MSI#26]: Loitering.

	data variation	volume																		variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																				
	Average recognition time (sec)	-	-	-	-	-	-	-	-	0.09	0.13	0.22	0.36	-	-	-	-	-	-	0.22	0.4	0.58	1.11	-	-	-	-	-	-
	Recall													1					1					(1)					(1)
	Precision													1					1					(1)					(1)
	F1													1					1					(1)					(1)

Table 50: CER - [MSI#27]: Dead in water, drifting (adrift).

	data variation	volume																				variety/veracity									
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24		
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																						
	Average recognition time (sec)	-	-	-	-	-	-	-	-	0.09	0.13	0.22	0.36	-	-	-	-	-	-	0.22	0.4	0.58	1.11	-	-	-	-	-	-		
	Recall													0.895					0.949					(1)					(1)		
	Precision													0.847					0.406					(1)					(1)		
	F1													0.870					0.568					(1)					(1)		

Table 51: CER - [MSI#28]: Rendez-vous.

	data variation	volume								variety/veracity																	
	dataset	$NARI_0$				Greekseas				$NARI_1$					$IMISG_1$					$NARI_2$					$IMISG_2$				
	number of cores	1	2	4	8	1	2	4	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	window size (h)	1	1	1	1	1	1	1	1	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24	2	4	8	16	24
Criterion(meas.)	Throughput(thousand events/sec)	-	-	-	-	-	-	-	-																				
	Average recognition time (sec)	-	-	-	-	-	-	-	-	0.09	0.13	0.22	0.4	-	-	-	-	-	-	0.22	0.4	0.67	1.11	-	-	-	-	-	-
	Recall													1					1					(1)					(1)
	Precision													1					1					(1)					(1)
	F1													1					1					(1)					(1)

Table 52: CEF mapping of deliverables to evaluation framework.

	data variation	-																										
	dataset	Greekseas																										
	prediction threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Markov-chain order	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
Criterion	Precision Spread Distance	D3.2, p.43																										

Table 53: CEF mapping of deliverables to evaluation framework (continued).

	data variation	volume																									
	dataset	$NARI_1$																									
	prediction threshold	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
	Markov-chain order	0	0	0	0	0	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4
	feature use	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Criterion	Precision Spread Distance	D3.5, p.18: Approaching a Port																									

	data variation	volume																									
	dataset	$NARI_1$																									
	prediction threshold	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
	Markov-chain order	0	0	0	0	0	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4
	feature use	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Criterion	Precision Spread Distance	D3.5, p.18: Fishing																									

Table 54: Future Location Prediction (FLP) for short- and long-term prediction.

	data variation	volume													
	dataset	$NARI_0$						$NARI_1$				$NARI_0$			
	prediction	10s	20s	40s	1min20s	2min40s	5min	1h30min	3h	6h	12h	1h30min	3h	6h	12h
Crit.	Median RMSE (m)	1	2	5	10	17	36	8450	15500	12680	12680	8450	14080	12680	12680
	Average (m)	5	10	20	40	100	280	-	-	-	-	-	-	-	-
	Max RMSE (m)	7	14	28	72	197	540	54900	101400	177460	260560	50700	116900	198590	257750

8.3 Data recorded for situation description and confidence in Experiment 2

	scenario 1		scenario 2		scenario 3		scenario 4	
Min	CoNoMSI	RdvNoMSI	RdvMSI1	NcMSI1	CoMSI	NcMSI2	RdvMSI2	NcMSI3
0.5	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-
1.5	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-
2.5	-	-	nds†	-	-	-	-	-
3	-	nds†	nds	-	-	-	-	-
3.5	-	nds	c5/rv5	-	-	-	-	-
4	-	**c4**	c5/rv5	-	-	-	nds†	-
4.5	-	-	**c5/rv5**	-	-	nds	-	-
5	-	-	*rv**	-	-	-	**nds***	-
5.5	-	-	-	-	-	-	t	-
6	-	-	-	nds	-	nds	-	nds
6.5	-	-	*noc*	cqs†	-	-	-	-
7	-*	*noc5,nc/rv4*	-	-	cqs†	-	-	-
7.5	-	*c/nc/rv*	-	**nc,noc5**	-*	*nc5*†	-	*nc*†
8	-	-*	-	-	-	-	-	-
8.5	-	-	-	-*	-	-	*not*	-*
9	-	-	-	-	-	-*	-	-
9.5	-	-	-	-	-	-	-	-
10	-	-	*rv5*	-	*noc5*	-	*t5*	*nc*
Prediction	FN	TP	TP	TP	TP	(TP)	TP	(TP)
Detection	FN	TP	TP	TP	FN	TP	FN,FP	TP
Time † to *	-	5	2.5	2	0.5	1.5	1	1
Pred.Conf.	-	rv4	rv5	nc5	-	nc5	-	-
Det.Conf.	-	-	rv5	-	noc5	-	t5	-

Table 55: Recorded Data Expert 1. The *-sign indicates the actual occurrence of the event. † indicates the earliest occurrence of a situational description rated as TP not withdrawn before the occurrence of the event. A **bold** acronym indicates the reported occurrence of an event. Acronyms in standard text indicate predictions before and acronyms in *italic* post-event assessment after the reported occurrence of an event. Numbers indicate the confidence level between 1 (low) and 5 (high), if available. Events are classified in: near-distance situation (nds), close-quarter situation (cqs), near-collision (nc), collision (c), rendez-vous (rv), tugging (t) and no collision (noc), etc. Comma (,) and slash (/) correspond to the logical OR and AND operators respectively. Minus (-) indicates no reported assessment or an assessment without relation to nds. FN refers to the event in the situational dataset, FP refers to the expert user prediction or detection, TP refers to both.

	scenario 1		scenario 2		scenario 3		scenario 4	
Min	CoNoMSI	RdvNoMSI	RdvMSI1	NcMSI1	CoMSI	NcMSI2	RdvMSI2	NcMSI3
0.5	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-
1.5	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	cqs	-
2.5	-	cqs	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-
3.5	-	-	nds†	-	-	-	-	-
4	-	-	-	-	-	-	**c**	-
4.5	-	-	-	-	-	-	**t**	-
5	-	-	*c**	-	-	-	-*	-
5.5	-	nc5	*c5*	-	-	-	-	-
6	-	-	*c3*	-	-	cqs†	-	-
6.5	-	**nc/c**	-	nds†	-	cqs	-	-
7	-*	-	-	-	cqs†	-	-	-
7.5	-	*noc*	-	nds	-*	-	-	-
8	-	-*	-	*nc*	*cqs*	-	-	-
8.5	-	-	-	-*	-	-	*c*	-*
9	-	-	-	-	-	-*	-	-
9.5	-	-	-	-	-	-	*c3*	-
10	noc5	*nc/c4-5*	*nc5*	*nc5*	*c5*	*nc5*	-	-
Prediction	FN	FN,FP	TP	(TP)	TP	TP	FN,FP	FN
Detection	FN	FN,FP	FN,FP	TP	TP	TP	FN,FP	FN
Time † to *	-	-	1.5	2	0.5	3	-	-
Pred.Conf.	-	nc5	-	-	-	-	-	-
Det.Conf.	noc5	cqs4-5	nc5	nc5	c5	nc5	c3	-

Table 56: Recorded Data Expert 2. The *-sign indicates the actual occurrence of the event. † indicates the earliest occurrence of a situational description rated as TP not withdrawn before the occurrence of the event. A **bold** acronym indicates the reported occurrence of an event. Acronyms in standard text indicate predictions before and acronyms in *italic* post-event assessment after the reported occurrence of an event. Numbers indicate the confidence level between 1 (low) and 5 (high), if available. Events are classified in: near-distance situation (nds), close-quarter situation (cqs), near-collision (nc), collision (c), rendez-vous (rv), tugging (t) and no collision (noc), etc. Comma (,) and slash (/) correspond to the logical OR and AND operators respectively. Minus (-) indicates no reported assessment or an assessment without relation to nds. FN refers to the event in the situational dataset, FP refers to the expert user prediction or detection, TP refers to both.

	scenario 1		scenario 2		scenario 3		scenario 4	
Min	CoNoMSI	RdvNoMSI	RdvMSI1	NcMSI1	CoMSI	NcMSI2	RdvMSI2	NcMSI3
0.5	-	-	-	-	-	-	-	-
1	nc	-	-	-	-	-	-	-
1.5	-	-	-	-	-	-	-	-
2	-	nds	-	-	-	-	cqs	-
2.5	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-
3.5	-	-	nds	-	-	-	-	-
4	c3/cqs†	-	-	-	c	-	c5	-
4.5	-	-	cqs	-	noc	cqs†	**t**	-
5	-	*c*	**c5***	cqs†	-	-	-*	-
5.5	-	-	*c*	-	-	cqs	*nt*	nds
6	-	-	-	-	-	-	-	-
6.5	-	-	-	-	-	-	-	*nc4*†
7	**c5***	*c/nc*	-	cqs/c	c4†	-	*nt*	-
7.5	-	-	-	-	-*	cqs	-	-
8	-	-*	-	-	-	-	-	-
8.5	-	-	-	-*	*c4*	-	*nt*	-*
9	-	-	*noc*	-	-	**nc5***	-	-
9.5	-	-	-	-	-	-	-	-
10	*c/noc*	*rv4/c3*	*noc5*	*rv5*	-	-	*nc4*	*nc4*
Prediction	TP	FN,FP	FN,FP	TP	TP	TP	FN,FP	(TP)
Detection	TP	TP	TP	FN,FP	TP	TP	FN,FP	TP
Time † to *	3	-	-	3.5	0.5	4.5	-	2
Pred.Conf.	-	-	-	-	c4	-	c5	nds4
Det.Conf.	c5	rv4	noc5	rv5	c4	nc5	nc4	nc4

Table 57: Recorded Data Expert 3. The *-sign indicates the actual occurrence of the event. † indicates the earliest occurrence of a situational description rated as TP not withdrawn before the occurrence of the event. A **bold** acronym indicates the reported occurrence of an event. Acronyms in standard text indicate predictions before and acronyms in *italic* post-event assessment after the reported occurrence of an event. Numbers indicate the confidence level between 1 (low) and 5 (high), if available. Events are classified in: near-distance situation (nds), close-quarter situation (cqs), near-collision (nc), collision (c), rendez-vous (rv), tugging (t) and no collision (noc), etc. Comma (,) and slash (/) correspond to the logical OR and AND operators respectively. Minus (-) indicates no reported assessment or an assessment without relation to nds. FN refers to the event in the situational dataset, FP refers to the expert user prediction or detection, TP refers to both.

8.4 Data recorded for maritime situational awareness

Dimension of situational awareness cp. [25]	exp. 2			exp. 3
	sc.1	sc.2	sc.3	$\overline{sc.2}$
Instability of Situation: How changeable is the situation? Is the situation highly unstable and likely to change suddenly (High) or is it very stable and straightforward (Low)?	1-2	4-5	6-7	1-2
Complexity of Situation: How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)?	1-2	3-4	5-6	1-2
Variability of Situation: How many variables are changing within the situation? Are there a large number of factors varying (High) or are there very few variables changing (Low)	2-3	3-4	5-6	2-3
Arousal: How aroused are you in the situation? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)?	6-7	5-6	5-6	4-5
Concentration of Attention: How much are you concentrating on the situation? Are you concentrating on many aspects of the situation (High) or focused on only one (Low)?	2-3	4-5	5-6	5-6
Division of Attention: How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (High) or focused on only one (Low)?	2-3	4-5	5-6	5-6
Spare Mental Capacity: How much metal capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (High) or nothing to spare at all (Low)?	6-7	5-6	5-6	5-6
Information Quantity: How much information have you gained about the situation? Have you received and understood a great deal of knowledge (High) or very little (Low)?	2-3	4-5	5-6	5-6
Familiarity with Situation: How familiar are you with the situation? Do you have a great deal of relevant experience (High) or is it a new situation (Low)?	6-7	6-7	6-7	6-7

Table 58: Situational awareness data recorded on experiment 2 and 3 for expert 1. Situational awareness rating cp. [25]: 1-Low, 7-High.

Dimension of situational awareness cp. [25]	exp. 2			exp. 3
	sc.1	sc.2	sc.3	$\overline{sc.2}$
Instability of Situation: How changeable is the situation? Is the situation highly unstable and likely to change suddenly (High) or is it very stable and straightforward (Low)?	5	6	6	6
Complexity of Situation: How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)?	2	4	5	2
Variability of Situation: How many variables are changing within the situation? Are there a large number of factors varying (High) or are there very few variables changing (Low)	2	5	5	2
Arousal: How aroused are you in the situation? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)?	6	6	6	6
Concentration of Attention: How much are you concentrating on the situation? Are you concentrating on many aspects of the situation (High) or focused on only one (Low)?	2	3	5	2
Division of Attention: How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (High) or focused on only one (Low)?	2	4	4	2
Spare Mental Capacity: How much metal capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (High) or nothing to spare at all (Low)?	7	4-5	6	5
Information Quantity: How much information have you gained about the situation? Have you received and understood a great deal of knowledge (High) or very little (Low)?	3	5	5	3
Familiarity with Situation: How familiar are you with the situation? Do you have a great deal of relevant experience (High) or is it a new situation (Low)?	6	6	6	6

Table 59: Situational awareness data recorded on experiment 2 and 3 for expert 2. Situational awareness rating cp. [25]: 1-Low, 7-High.

Dimension of situational awareness cp. [25]	exp. 2			exp. 3
	sc.1	sc.2	sc.3	$\overline{sc.2}$
Instability of Situation: How changeable is the situation? Is the situation highly unstable and likely to change suddenly (High) or is it very stable and straightforward (Low)?	5	5	6	6
Complexity of Situation: How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)?	4	5	5	6
Variability of Situation: How many variables are changing within the situation? Are there a large number of factors varying (High) or are there very few variables changing (Low)	3	5	6	7
Arousal: How aroused are you in the situation? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)?	4	4	5	4
Concentration of Attention: How much are you concentrating on the situation? Are you concentrating on many aspects of the situation (High) or focused on only one (Low)?	4	5	6	3
Division of Attention: How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (High) or focused on only one (Low)?	5	4	6	4
Spare Mental Capacity: How much metal capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (High) or nothing to spare at all (Low)?	5	5	5	5
Information Quantity: How much information have you gained about the situation? Have you received and understood a great deal of knowledge (High) or very little (Low)?	2	4	5	7
Familiarity with Situation: How familiar are you with the situation? Do you have a great deal of relevant experience (High) or is it a new situation (Low)?	7	5	6	6

Table 60: Situational awareness data recorded on experiment 2 and 3 for expert 3. Situational awareness rating cp. [25]: 1-Low, 7-High.

8.5 Agenda of the final experiments week

datAcron WP5-WP4 experiments

La Spezia, 5 - 9 November 2018

AGENDA

Monday, 5 November 2018

9:30 - 17:30	**CMRE** Data alignment (join of components' outputs: LED, SI, SG, CER) Preparation for scenario evaluation Rehearsal and Briefing with the experts

Tuesday, 6 November 2018

9:30 - 12:30	**CMRE, NARI, FRHF** Joint meeting for finalization of experiments' objectives and setting up including assessment metrics			**CMRE** Preparation for scenario evaluation
12:30 -13:30	Lunch break			
13:30 -17:30	**CMRE, FRHF, NARI** Prototype setup including agenda, questionnaire, evaluation dataset to be ingested and displayed, informed consent, confirm thresholds, setting up of capturing tools, exp2 vs. exp5	**CMRE** Preparation for scenario evaluation	**CMRE, FRHF** Visualization of icons **FRHF** IVA/Viz development	**NARI, CMRE** Exp 4: template, objectives, expectations, reporting Exp4: data integration: LED, SG, CER outputs vs. raw, enriched

Wednesday, 7 November 2018

9:30 - 12:30	**CMRE, NARI** Prototype setup: setting up of capturing tools, evaluation dataset to be	**FRHF** IVA/Viz development and eye tracking software setting and tuning	**NARI, CMRE** Exp4: data integration: LED, SG, CER outputs vs. raw, enriched Exp 4: MSI evaluation : LED, SG, CER outputs vs. raw, enriched	**CMRE** Exp 1 (Variety Game, including Exp 2 & 3 training)

	ingested and displayed.			
12:30-13:30	Lunch break			
13:30-15:00	**FRHF, CMRE** IVA/Viz development and eye tracking software setting and tuning Video preparation for Exp2, Exp 3 setup (Evaluation of IVA/Viz functionalities and Eye tracking)	**NARI, CMRE** Exp 4: MSI evaluation : LED, SG, CER outputs vs. raw, enriched	**CMRE** Exp 1 (Variety Game, including Exp 2 & 3 training)	
15:00 - 17:00				**CMRE, NARI, FRHF** Datacron telecon
17:00 - 17:30	**CMRE, NARI, FRHF** Finalization of prototype setup, report on Exp4, rehearsal of Exp 2, 3, 5			

Thursday, 8 November 2018

9:30 - 10:00	**CMRE, NARI, FRHF, Cadets** **Rehearsal** Exp 2: Maritime Surveillance & Eye tracking experiments	**CMRE** Exp 1 (Variety Game)	**NARI** Exp4: data integration: LED, SG, CER outputs vs. raw, enriched
10:30 - 10:45	Break		
10:45 - 11:30	**CMRE, NARI, FRHF, Expert 1** Exp 3: IVA/Viz experiments and Exp5: datacron (computed) MSIs	**CMRE** Exp 1 (Variety Game)	**NARI** Exp4: data integration: LED, SG, CER outputs vs. raw, enriched
11:30 - 11:45	Break		
11:45 - 12:15	**CMRE, NARI, FRHF, Expert 1** Exp 2: Maritime Surveillance & Eye tracking experiments	**CMRE** Exp 1 (Variety Game)	**NARI** Exp4: data integration: LED, SG, CER outputs vs. raw, enriched
12:45 - 13:30	Lunch break		

13:30 - 14:15	**CMRE, NARI, FRHF, Expert 2** Exp 3: IVA/Viz experiments and Exp5: datacron (computed) MSIs	**CMRE** Exp 1 (Variety Game)	**NARI** Exp 4: MSI evaluation : LED, SG, CER outputs vs. raw, enriched
14:15 - 14:45	Break		
14:45 –15:30	**CMRE, NARI, FRHF, Expert 2** Exp 2: Maritime Surveillance & Eye tracking experiments	**CMRE** Exp 1 (Variety Game)	**NARI** Exp 4: MSI evaluation : LED, SG, CER outputs vs. raw, enriched
15:30 - 16:00	Break		
16:15 - 17:00	**CMRE, NARI, FRHF** Discussion on preliminary results - Result analysis - First reporting Wrap-up		

Friday, 9 November 2018

9:30 - 10:30	**CMRE, NARI, FRHF, Expert 3** Exp 2: Maritime Surveillance & Eye tracking experiments	**CMRE** Exp 1 (Variety Game)	**NARI, CMRE** Exp 4: MSI evaluation : LED, SG, CER outputs vs. raw, enriched
10:30 - 11:00	Break		
11:00 - 11:30	**CMRE, NARI, FRHF, Expert 3** Exp 3: IVA/Viz experiments & Exp5: datacron (computed) MSIs	**CMRE** Exp 1 (Variety Game)	**NARI, CMRE** Exp 4: MSI evaluation : LED, SG, CER outputs vs. raw, enriched
11:30 - 12:30	**CMRE, NARI, FRHF** Final discussion Roles and responsibilities until M36 - Result analysis - Reporting Wrap-up		
12:30 -	Lunch break		

13:30	
14:00 - 15:00	**CMRE** Exp 1 (Variety Game)

References

[1] Convention on the international regulations for preventing collisions at sea (COLREGs). Standard, International Maritime Organization, 1972.

[2] Elias Alevizos, Alexander Artikis, Georgios Paliouras, Evangelos Michelioudakis, Ehab Qadah, Michael Mock, and Georg Fuchs. Complex event forecasting, H2020 datacron d3.4, 2018.

[3] Elias Alevizos, Ioannis Kontopoulos, Efthimios Tsilionis, Alexander Artikis, Michael Mock, and Georgios Paliouras. Robust complex event recognition (interim), H2020 datacron d3.2, 2017.

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

[5] Eva Chondrodima, Harris Georgiou, Kostas Patroumpas, Nikos Pelekis, Petros Petrou, Stylianos Sideridis, Dimitris Spirelis, Panagiotis Tampakis, and Yannis Theodoridis. Cross-streaming, real-time detection of moving object trajectories (final), H2020 datacron d2.3, 2018.

[6] Eva Chondrodima, Harris Georgiou, Nikos Pelekis, Petros Petrou, Stylianos Sideridis, Panagiotis Tampakis, Yannis Theodoridis, Anne-Laure Jousselme, and Clément Iphar. Big data analytics for time critical mobility forecasting, H2020 datacron d2.5, 2018.

[7] Francesca de Rosa, Anne-Laure Jousselme, and Alessandro De Gloria. A reliability game for source factors and situational awareness experimentation. *International Journal of Serious Games*, 5(2):45–64, Jun. 2018.

[8] Harris Georgiou, Sophia Karagiorgou, Kostas Patroumpas, Nikos Pelekis, Petros Petrou, Stylianos Sideridis, Eugenia Stoufi, and Yannis Theodoridis. Cross-streaming, real-time detection of moving object trajectories (interim), H2020 datacron d2.1, 2017.

[9] Harris Georgiou, Petros Petrou, Panagiotis Tampakis, Stylianos Sideridis, Eva Chondrodima, Nikos Pelekis, and Yannis Theodoridis. Short- and long-term prediction of routes online (final), H2020 datacron d2.4, 2018.

[10] S. A. Harvald. Factors affecting the stopping ability of ships. *International Shipbuilding Progress*, 23(260):106–121, 1976.

[11] Helmut Hilgert. Defining the close-quarters situation at sea. *The Journal of Navigation*, 36(3):454–461, 1983.

[12] Clément Iphar, Cyril Ray, Maximilian Zocholl, Anne-Laure Jousselme, Richard Dréo, and Elena Camossi. Maritime datacron prototype set up, H2020 datacron d5.5, 2018.

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

[14] Anne-Laure Jousselme, Giuliana Pallotta, and John Locke. Risk game: Capturing impact of information quality on human belief assessment and decision making. *International Journal of Serious Games*, 5(4):23–44, Dec. 2018.

[15] 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.

[16] Nikos Katzouris, Evangelos Michelioudakis, Elias Alevizos, Alexander Artikis, and Georgios Paliouras. Adaptive complex event recognition (interim), H2020 datacron d3.1, 2017.

[17] Nikos Katzouris, Evangelos Michelioudakis, Alexander Artikis, and Georgios Paliouras. Adaptive complex event recognition (final), H2020 datacron d3.3, 2018.

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

[19] Cyril Ray, Richard Dréo, 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.

[20] Cyril Ray, Richard Dréo, Elena Camossi, and Anne-Laure Jousselme. Heterogeneous integrated dataset for maritime intelligence, surveillance, and reconnaissance (version 0.1) [data set], 2018.

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

[22] Cyril Ray, Richard Dréo, Elena Camossi, Anne-Laure Jousselme, and Clément Iphar. Heterogeneous integrated dataset for maritime intelligence, surveillance, and reconnaissance (under evaluation). *Data in brief*, 2018.

[23] Cyril Ray, Clément Iphar, Richard Dréo, Waldo Kleynhans, Elena Camossi, Anne-Laure Jousselme, Maximilian Zocholl, Ernie Batty, Quentin Roche, and Arnaud Metzger. Maritime data preparation and curation (final), H2020 datacron d5.4, 2018.

[24] Cyril Ray, Clément 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.

[25] RM Taylor. Situational awareness rating technique (sart): The development of a tool for aircrew systems design. In *Situational Awareness*, pages 111–128. Routledge, 2017.

[26] Efthimis Tsilionis, Manolis Pitsikalis, Alexander Artikis, and Georgios Paliouras. Robust complex event recognition (final), H2020 datacron d3.4, 2018.

[27] George Vouros, Christos Doulkeridis, Nikos Pelekis, Elias Alevizos, Georg Fuchs, and Gennady Andrienko. datAcron component-scenario mapping. https://docs.google.com/spreadsheets/d/1n61_6RJmDfrTeCB71O3yS9TX_6UQ2fd_qFbXW-01RSU/edit#gid=0. Accessed: 2018-12-20.