Evaluation of maritime event detection against missing data

Maximilian Zocholl[1], Clément Iphar[1], Manolis Pitsikalis[2], Anne-Laure Jousselme[1], Alexander Artikis[2,3], and Cyril Ray[4]

[1] NATO STO Centre for Maritime Research and Experimentation, La Spezia, Italy
firstname.lastname@cmre.nato.int
[2] NCSR Demokritos, Athens, Greece
manospits@iit.demokritos.gr
[3] University of Piraeus, Greece
a.artikis@unipi.gr
[4] Naval Academy Research Institute, Brest, France
cyril.ray@ecole-navale.fr

Abstract. Detecting and preventing maritime events like collisions or unusual behaviour of vessels are of high importance for maritime safety and security. As the trust of human operators in automated maritime event detection and prediction depends on the quality of the corresponding algorithms, the evaluation methodology becomes a driving force for the future development of maritime event detection and forecasting methods. The main contribution of this article consists in the development of an evaluation methodology and its application to a selected set of maritime event detectors. The approach links a reference dataset, controlled data variations, maritime event detection algorithms with internal parameters, and performance criteria. Among pre-established possible input data variations applied to a reference Automatic Identification System (AIS) dataset, the article focuses on the evaluation of detection accuracy of maritime event detectors implemented with the Event Calculus logical language against variable amounts of missing data, as a frequently observable type of AIS data degradation. Twelve maritime event pattern detectors are evaluated and most of them are found to vary very little in performance while only one detector shows an unexpected strong performance drop giving insights into how to improve the detection method. Results are provided on a real AIS data enriched with specific simulated events.

Keywords: Evaluation methodology · maritime event detection · Event Calculus · datasets creation · missing data · data veracity.

1 Introduction

In the maritime domain Automatic Identification Systems (AIS) installed on vessels transmit periodically position, speed and other information about the vessel [6]. Primarily designed for collision avoidance purposes, AIS data became important for a broad range of detection, prediction and forecasting applications. For the beneficial exploitation of the large amount of AIS data received by base stations and satellites, scalable data processing solutions are required that increase and speed up the maritime situational awareness of human operators.

As operators trust only on antecedent data processing units that yield high data quality, their evaluation requires a meaningful methodological basis. Existing evaluation methods underline the importance of considering imprecise real-world conditions [11]. A comparison of evaluation methods is proposed by [9]. Based on uncertainty represenations, the impact of different imputation strategies on the retrieval performance on incomplete data is characterised in [7]. An evaluation method for classifiers used on classes with strongly varying occurrence probabilities and costs for misclassifications is presented in [10]. While the creation of training and testing datasets is always expensive due to manual labelling work, the creation of maritime training and testing datasets requires subject matter experts and additionally has to rely on assumptions. This is necessary as scarce maritime events are only partially observable via AIS data, which has unknown quality and can be incomplete. The evaluation of computer vision based maritime event detectors using synthetic track generation with position and speed information is proposed by [4]. An evaluation method for maritime event visualisation proposes [14].

The article proposes an evaluation method using missing data mechanisms, as initially classified by [16], and applies them to AIS messages. The main contribution of the presented work is the meaningful combination and articulation of data variation method and evaluation methods and criteria. The article is divided in three parts, firstly maritime situational indicators are presented as a list of maritime events of interest, followed by a description of an implementation of a subset of detectors of these indicators. The second section describes the data variation methodology for the systematic and reproducible creation of testing or training data with known veracity variation, here performed as a reduction of AIS messages for each vessel. The third section presents and discusses the results obtained. Future perspectives and a summary conclude the contribution to the use of AIS data for the evaluation of maritime events detection.

2 Maritime Events

For operators in a maritime surveillance mission maritime events are typically observable only indirectly via a collection of incomplete data which comes from different sources and has different values of veracity. The interpretation of the data in the maritime context results in a maritime situational awareness. As different sources can possibly give different, even contradicting information, the distinction between a type of maritime event, introduced in the following under the notion of maritime situational indicators (MSI) and its subsequently presented implementation is crucial.

2.1 Maritime situational indicators (MSI)

Maritime situational indicators (MSI) are intended to alert an operator observing a maritime situation about important changes, hence to filter information and to drive the operators attention to a specific location in the surveillance area.

Table 1: List of Maritime Situational Indicators.

#	Maritime Situational Indicator	#	Maritime Situational Indicator
1	Close to critical infrastructure	15	No AIS reception
2	Within a given area	16	AIS reception interrupted
3	On a maritime route	17	Change in AIS static information
4	Proximity to other vessels	18	AIS error detection
5	In stationary area	19	Under way
6	Null speed	20	At anchor or moored
7	Change of speed	21	Movement ability affected
8	Mismatch speed area	22	Aground
9	Mismatch speed vessel type	23	Engaged in fishing
10	Mismatch speed vessel history	24	Tugging
11	Mismatch speed user defined value	25	In Search And Rescue (SAR) operation
12	Change of course	26	Loitering
13	Mismatch course vessel destination	27	Dead in water, drifting
14	Mismatch course user defined value	28	Rendez-vous

This list of MSIs is a synthesis of outcomes of workshops gathering user's elicitation and reported in the literature e.g. [1,15]. These MSIs have then been filtered according to their ability to be automatically detected or predicted through the processing of AIS data. For instance, any MSI referring to visual sighting has been excluded from this list. In this paper the implementation of detectors for the MSIs 2, 6, 7, 8, 9, 16, 19, 21 as well as five speed related building blocks are evaluated.

2.2 Complex event recognition with RTEC

The 'Event Calculus for Run-Time reasoning' (RTEC) [3,2] is an open-source Prolog implementation of the Event Calculus [8], designed to compute continuous narrative assimilation queries for pattern matching on data streams. RTEC has a formal, declarative semantics—complex patterns are (locally) stratified logic programs [12]. Moreover, RTEC includes optimisation techniques for efficient pattern matching, such as 'windowing', whereby all input events that took place prior to the current window are discarded/'forgotten'. Details about the reasoning algorithms of RTEC, including a complexity analysis, may be found in [3].

The time model in RTEC is linear and includes integer time-points. An *event description* includes rules that define the event instances with the use of the

Table 2: Main predicates of RTEC. '$F\!=\!V$' denotes that fluent F has value V.

Predicate	**Meaning**
happensAt(E, T)	Event E occurs at time T
holdsAt$(F = V, T)$	The value of fluent F is V at time T
holdsFor$(F = V, I)$	I is the list of the maximal intervals for which $F\!=\!V$ holds continuously
initiatedAt$(F = V, T)$	At time T a period of time for which $F\!=\!V$ is initiated
terminatedAt$(F = V, T)$	At time T a period of time for which $F\!=\!V$ is terminated
union_all(L, I)	I is the list of maximal intervals produced by the union of the lists of maximal intervals of list L
intersect_all(L, I)	I is the list of maximal intervals produced by the intersection of the lists of maximal intervals of list L
relative_complement_all(I', L, I)	I is the list of maximal intervals produced by the relative complement of the list of maximal intervals I' with respect to every list of maximal intervals of list L

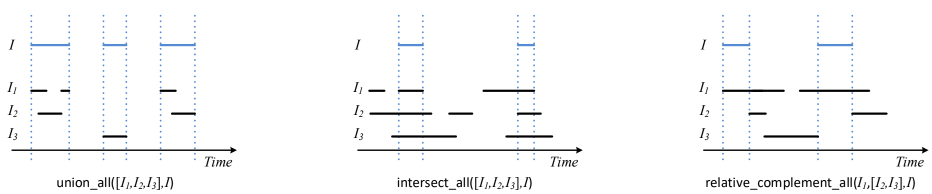


Fig. 1: A visual illustration of the interval manipulation constructs of RTEC. In these examples, there are three input streams, I_1, I_2 and I_3, coloured black. The output of each interval manipulation construct I is coloured light blue.

happensAt predicate, the effects of events on *fluents*—time-varying properties—with the use of the initiatedAt and terminatedAt predicates, and the values of the fluents with the use of the holdsAt and holdsFor predicates. Table 2 summarises the main predicates of RTEC.

Fluents are 'simple' or 'statically determined'. In brief, simple fluents are defined by means of initiatedAt and terminatedAt rules, while statically determined fluents are defined by means of application-dependent holdsFor rules, along with the interval manipulation constructs of RTEC: union_all, intersect_all and relative_complement_all. See Table 2 for a brief explanation of these constructs and Figure 1 for an example visualisation. Complex events/activities are typically durative; thus the task generally is to compute the maximal intervals for which a fluent expressing a complex activity has a particular value continuously.

3 Dataset variations

3.1 Notations

The generation of pseudo-synthetic datasets requires an original AIS dataset to which is applied a series of modifications that will be presented in Section 3.2.

In the remaining of this paper, let us denote by D the set of datasets. As to provide a common frame for the data pseudo-synthesis functions, let us denote by X_n^m a dataset containing n rows and m columns. Original datasets are denoted by $\bar{X}$, and synthetic datasets are denoted by $\widehat{X}$.

All operations performed on those datasets in this paper are presented by a series of functions $f_k : D \to D$. We denote by x_{i_j} the value of the j^{th} column of the i^{th} entry

3.2 Data variation functions of interest

Three families of functions

Three main families of data pseudo-synthesis functions are distinguished: the data improvement, the degradation and the event injection [5]. Data improvement increases the data veracity level, by the addition of attributes (add columns by labelling operation) or the addition of contacts (add rows that were sent but not received by the antenna). Data degradation lowers the data veracity level, by removing whole data attributes (columns), whole contacts (rows) or adding noise in data (field values are blurred in accordance with a chosen probabilistic model). The removal can be targeted on a trajectory or follow a law (linear decrease with the distance or uniform for instance). Event injection modifies the "story" that the data tells, by contact addition. This addition can consist of either the injection of a spatio-temporal shift of a trajectory already existing in the dataset or the targeted injection of synthetic data directly in the existing fields. The injection of synthetic events consists in the synthesis of specifically designed contacts in order to create maritime events depending on a handful of parameters and providing a high flexibility in the precise modelling of a wanted situation.

Assign function

The assign function is an event injection function having, in our case, the purpose to better fit existing data to a given scenario. This function takes a given data field and changes its value, either by applying a given offset to the previously existing value, or by replacing the value by a fixed value.

The parameters that can be set in this function are: the mode μ (μ_o in case of offset, μ_v in case of fixed value assignment), the value V to be used as assigned value or offset, the data field (column) of application c, the set of messages A (rows) to which this assignment must be performed. The nature of the value V depends on the nature of the data within the field c, as numeric fields must be assigned numeric values and categorical data must be assigned valid categories. The set of messages A which are processed can be a list of id number of messages, a list of MMSI number (unique ship identifier) of interest or a random subsample of such.

$$f_a(\mu, c, A, V): \ \bar{X}_n^m \mapsto \widehat{X}_n^m \tag{1}$$

where $\bar{X}_n^m = X'^m_{Card(A)} \frown X''^m_{n-Card(A)}$ and $\widehat{X}_n^m = \widehat{X}'^m_{Card(A)} \frown X''^m_{n-Card(A)}$, where $\frown$ denotes the append operation.

The operation from $X'^m_{Card(A)}$ to $\widehat{X}'^m_{Card(A)}$ is described as such: $x_{i\in A}^{j=c} \mapsto \widehat{x}_{i\in A}^{j=c}$ where $\forall i \in A$ $\widehat{x}_i^c = V$ if $\mu = \mu_v$ and $\widehat{x}_i^c = x_i^c + V$ if $\mu = \mu_o$.

Event function

The event function is an event creation function, of which the purpose is to synthetise new entries in the dataset in order to generate a given event of interest. Currently, the events available are the collision, the near-collision and the rendezvous, the latter being the voluntary meeting of two vessels at sea defined in three steps: the approach, the co-location (with variable duration) and the separation. The addition of such events, rare in the real world, is crucial in the understanding of maritime behaviours and maritime security.

The main parameters here are the number n and the kind of event E wanted. Then, according to the kind of event, other specific parameters p can be set, for instance for targeting a specific point of real data, choosing the angle of approach, the angle of departure or the time of rendezvousing.

As of today, $E \in \{C, NC, R\}$, respectively standing for Collision, Near-Collision and Rendezvous, with specific parameters associated p_C, p_{NC} and p_R. For instance, p_C is a table of k specific parameters (currently five: targeted point, angle of approach, speed of approach, identity and nature of synthetic vessel) each being a vector of n values, *i.e.* one for each of the events created.

$$f_e(q, E, p): \ \bar{X}_n^m \mapsto \widehat{X}^m_{n+Card(E)} \tag{2}$$

where $\widehat{X}^m_{n+Card(E)} = \bar{X}_n^m \frown \widehat{X}^m_{Card(E)}$, $\widehat{X}^m_{Card(E)}$ standing for the newly created event data. $Card(E)$ is not a fixed number, and it varies accordingly with the number, nature and specific local conditions of the created events.

Remove function

The remove function is a data degradation function, the purpose of which is to decrease the dataset size by removing entire rows of data.

The parameters to define are the subset of interest α, the nature of data to be removed A (following the same rules that the one from assign function) and the nature N of the removal. This nature N can define the grounds on which data is removed, *i.e.* either totally at random (MCAR) or based on the simulation of a natural process (MAR, such as the distance to the receiving station in the case of a reception simulation), or targeting on the base of some data fields values (MNAR, e.g. remove messages of speed inferior to a given threshold). In our case, $A = A_p$ is a set of messages defined by the percentage p of data not to remove.

$$f_r(\alpha, A_p, N): \ \bar{X}^m_n \mapsto \widehat{X}^m_{n-\lceil \alpha \cdot Card(A_p) \rceil} \tag{3}$$

where $\bar{X}^m_n = \widehat{X}^m_{n-Card(A_p)} {}^\frown X'^m_{Card(A_p)}$, from which $X'^m_{Card(A_p)}$ dataset has been removed. N is user-defined and $Card(A_p)$ is defined with respect to α such as $\forall j \in [1, J]$, L_j is one MMSI in the list L of all MMSI in α, $Card(A_p) = \sum_{j=1}^{J} \lceil p * Card(L_j) \rceil$, p being the parameter defined for A, as described above.

3.3 Dataset generation

Let us denote by X_α the source dataset, consisting of two subsets X_β and X_γ such as $X_\alpha = X_\beta {}^\frown X_\gamma$. X_β and X_γ each consist of 30 minutes of recorded data excerpted from a source AIS dataset [13] described in []. Each dataset contains rows (AIS messages) and columns (parameters). X_β action takes place in the Brest roadstead while X_γ is located in the Four channel, off the Brittany coast, in France.

In this section, let us simplify the notation of datasets by removing the number of rows and columns: $\widehat{X}^m_n = \widehat{X}$. Across this paper, no column modification is made and the value remains constant at $m = 11$. Rows modifications occur at several stages, and the value taken by n is indicated in Figure 3 and Table 3.

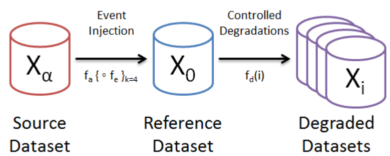


Fig. 2: Datasets Generation Workflow

This source dataset was then enriched by a total of four events, with one rendezvous, one collision and two near-collisions, as well as a series of data assignment on speed values to preexisting data so that the presented scenario is consistent. Les us denote by X_0 the resulting dataset, generated such as:

$$\widehat{X}_0 = \big(f_a(\mu_v, speed, A, V) \circ f_e(2, NC, p_{NC}) \circ f_e(1, R, p_R) \circ f_e(1, C, p_C)\big)(X_\alpha) \tag{4}$$

where A is the subset of interest for specific speed modification (in order to better fit preexisting data to the scenario created by the event injection), V is a vector of those assigned speeds, p_{NC}, p_R and p_C the parameters of the event functions, not explicitly described here and $n_{(\widehat{X}_0)} = n_{(X_\alpha)} + \sum_{i=1}^{4} Card(E_i)$.

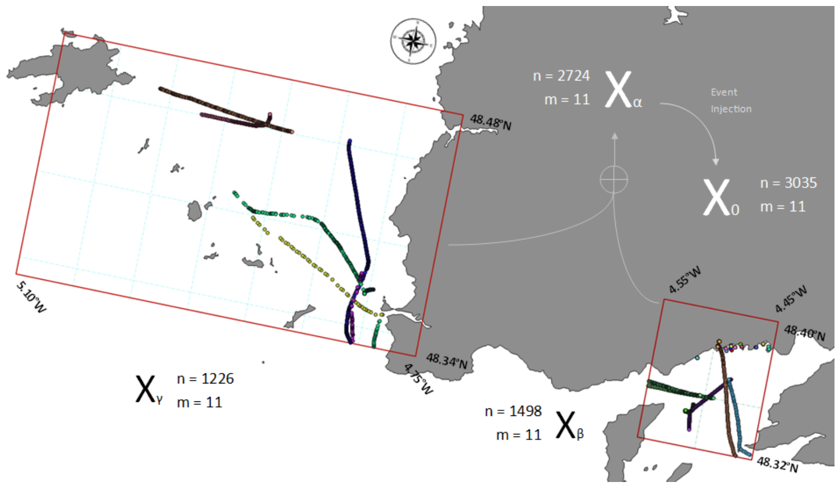


Fig. 3: The $\widehat{X}_0$ reference dataset, with the two boundary boxes of the areas from which X_β and X_γ are excerpted. Each point is a contact, either original ($\in X_\alpha$) or synthetic ($\in \widehat{X}_0 \setminus X_\alpha$). Each colour stands for a unique MMSI value in each box. For each dataset, m and n stand for the number of columns and rows

From now on, $\widehat{X}_0$ will be considered as the reference dataset. The data within this dataset is presented in Figure 3. From this reference dataset, a series of degraded datasets will be generated, using the remove function. Three levels of degradation are performed on data: the removal of 10, 20 and 30% of data from each unique MMSI number. A total of five datasets for each level of degradation are produced. All generated datasets are different because of the random nature of the data removal process random draw of $\lceil p * Card(n_{MMSI}) \rceil$ values amongst the n_{MMSI} messages sent by each unique MMSI.

Table 3: List of degraded datasets

name	i	X	p	k	m	n	name	i	X	p	k	m	n	name	i	X	p	k	m	n
$\widehat{X}_1$	1	$\widehat{X}_0$	90	1	11	2729	$\widehat{X}_6$	6	$\widehat{X}_0$	80	1	11	2428	$\widehat{X}_{11}$	11	$\widehat{X}_0$	70	1	11	2125
$\widehat{X}_2$	2	$\widehat{X}_0$	90	2	11	2729	$\widehat{X}_7$	7	$\widehat{X}_0$	80	2	11	2428	$\widehat{X}_{12}$	12	$\widehat{X}_0$	70	2	11	2125
$\widehat{X}_3$	3	$\widehat{X}_0$	90	3	11	2729	$\widehat{X}_8$	8	$\widehat{X}_0$	80	3	11	2428	$\widehat{X}_{13}$	13	$\widehat{X}_0$	70	3	11	2125
$\widehat{X}_4$	4	$\widehat{X}_0$	90	4	11	2729	$\widehat{X}_9$	9	$\widehat{X}_0$	80	4	11	2428	$\widehat{X}_{14}$	14	$\widehat{X}_0$	70	4	11	2125
$\widehat{X}_5$	5	$\widehat{X}_0$	90	5	11	2729	$\widehat{X}_{10}$	10	$\widehat{X}_0$	80	5	11	2428	$\widehat{X}_{15}$	15	$\widehat{X}_0$	70	5	11	2125

The list of generated datasets $\widehat{X}_i$ is presented in Table 3 (in which p is the removal rate, k the identifier of the dataset linked to the p value, m the number of columns and n the number of lines), $\forall i \in [\![1, 15]\!]$, i referring to the corresponding line in Table 3, the synthetic datasets, computed using the function described in Section 3.2 are:

$$\widehat{X}_i = f_r(source_i, A_{p_i}, MCAR) \tag{5}$$

As a consequence, $\forall i \in [\![1, 5]\!]$, $\widehat{X}_i$ are the dataset with a rate of removal of 10%, 20% $\forall i \in [\![6.10]\!]$ and 30% $\forall i \in [\![11, 15]\!]$. Those datasets, alongside with $\widehat{X}_0$, are used in the assessments, the results of which are presented in Section 4.

4 Evaluation of MSI detection with data removal

The proposed evaluation method depicted in Figure 4 includes the following elements and their exemplary instantiations: The workflow starts from a **Real world phenomenon** that triggers data variations. Here, the variation of AIS reception probability is chosen. The reception probability diminishes non-linearly with increasing distances between transmitter and receiver. In order to make the results applicable to different AIS receivers, the data is varied linearly. The **Varied Data** is the outcome of the process described in 3, specified by three data removal rates. As an example for a **Detector** RTEC is chosen, as presented in 2.2. The output of the detector is then evaluated with the **Evaluation Criteria** and measures. As one of many possible instantiations, the metrics used for measuring the accuracy are the number of true positives (TP), false positives (FP) and the derived metrics recall, equal to the TPs per number of all actual positives (P) and precision, equal to the TPs per number of all assumed positives which are the sum of TPs and FPs, and F1-score. The variation of these metrics relative to the amount of removed data is depicted and possible reasons for the observed differences in the variations between different MSI detections are discussed. The **Interpretation** of the evaluation of the linear data variation for a specific receiver requires the chaining of a specific receiver function f_r : Distance $\rightarrow$ Reception Probability and the detector function f_d : Reception Probability $\rightarrow$ Accuracy as $f_d(f_r)$. Similarly, the interpretation is meant to contain application specific cost functions for a weighting of different types of misclassification.

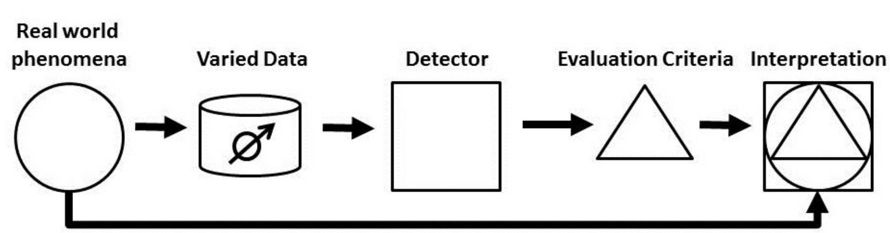


Fig. 4: Evaluation workflow

4.1 Variations of MSI detections on data removal

In this experiment, the evaluation reference is provided by RTEC detections on the dataset X_0. This allows for a quantitative analysis of the variation in accuracy of MSI detectors to the lack of data. It does not state that the detections on the non-degraded dataset are correct in the first place. In order to evaluate the accuracy of detections on X_0, an expert is suitable to provide a more complete assessment of the MSI detectors.

The sensitivity analysis investigates the impact of veracity variations on the performance of RTEC. 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 data removal on the performance of RTEC.

Experiment description

From the data degradation methods described in section 3.2, the data removal method is used for removing randomly 10, 20 and 30% of the messages of each vessel from the non-degraded dataset. Thus, random 90, 80 and 70% of the original dataset are kept and processed by RTEC. The data removal process was repeated 5 times leading to 5 different datasets for each data removal rate. The same datasets were used as input for all event detections. As a selection of MSIs presented in 2.1 MSIs 2, 6, 7, 8, 9, 16, 19, 21 as well as five speed related building blocks were evaluated. For each event results are shown as average, minimum and maximum of the 5 different datasets.

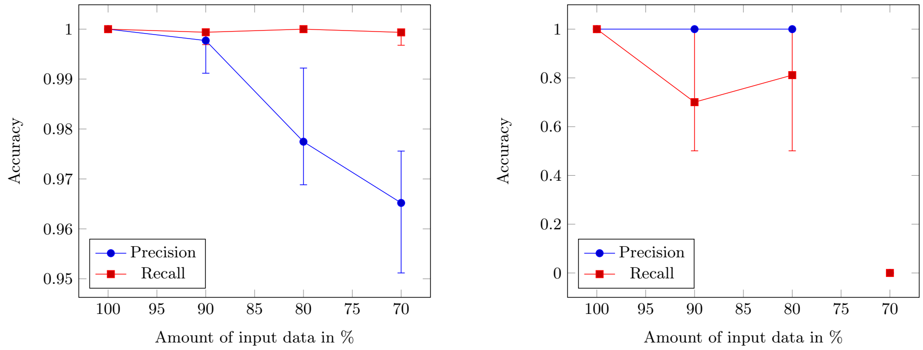
Discussion of results

As expected, for most events detected RTEC, the larger the data degradation, the lower Recall, Precision and F1. Remarkably is that most of the events detected show only relatively small variation in performance with respect to the large amounts of removed data, as shown in Table 4. The range of performance variations is shown in Figure 5a with the smallest variations by the event “high speed near coast” and the largest variations of “movement ability affected” in Figure 5b.

In the group of events with small performance variations a reduction of 30% leads to a reduction of less than 3% of the F1-score for “movingSpeed”, “tuggingSpeed”, “underway” and “withinArea”, ca. 4% for “lowSpeed” and ca. 7% for “unusualSpeed”. A behaviour which is still mitigating the effect of data removal is observed for “changingSpeed” and “gap” for which a data reduction of 30% leads only to a reduction of less than 20% in F1 score.

A linear and strong variation of performance due to the data removal is shown by the event “stopped”, for which 30% of data removal lead to ca. 37% of reduction in F1-score; Similar for "SAR course".

The largest variation in performance is observed for the event “movement ability affected”. Here, the removal of 20% of the data leads to an indulgent behaviour while the removal of additional 10% of the data leads to a factual collapse of Recall and Precision, implied by the fact that no TP detections are



(a) "high speed near coast" shows small variations

(b) "movement ability affected" shows large variations

Fig. 5: Impact of data removal on accuracy measured by mean, minimum and maximum precision and recall over 5 datasets.

made in the 5 differently degraded datasets. The primary reason for the strong variability in performance is due to the fact that only one event is detected in X_0. But MSIs with strong variations in performance do not necessarily have a small number of detections in the non-degraded dataset X_0. The accuracy of the "stopped" event detector is also varying strongly while the number of detections in X_0 reaches 4,364 detections. This is a larger number of detections than other MSIs which vary much less, e.g. "high speed near coast" with 1,925 detections in X_0. Hence, the variability of the performance is not necessarily uniquely dependent on the number of samples.

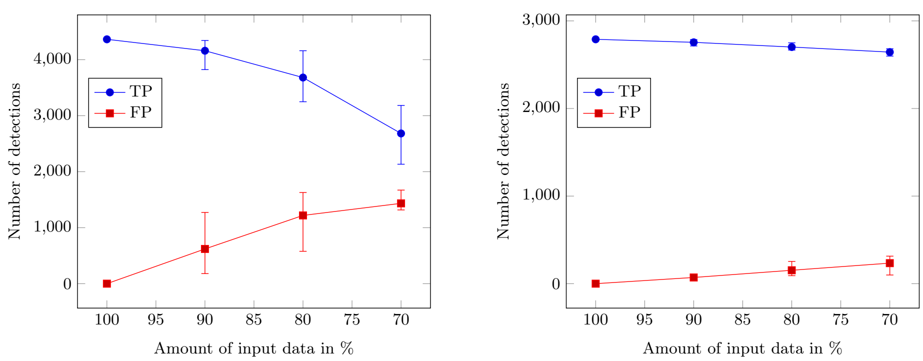
Table 4: Mean values for Precision, Recall and F1-score for detection of maritime event patterns on 90, 80 and 70% of data from the reference dataset X_0.

		90%			**80%**			**70%**		
MSI#	**Pattern**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
7	changing speed	0.974	0.908	0.939	0.949	0.836	0.888	0.899	0.748	0.815
16	gap	0.899	0.999	0.946	0.795	0.997	0.885	0.695	0.969	0.809
8	high speed nc	0.998	0.999	0.999	0.977	1	0.9886	0.965	0.999	0.982
11	low speed	0.995	0.989	0.992	0.99	0.976	0.983	0.957	0.963	0.96
21	maa	1	0.700	0.8	1	0.811	0.876	0	0	-
19a	movingSpeed	0.995	0.998	0.996	0.992	0.997	0.995	0.961	0.994	0.976
25a	SAR Course	1	0.901	0.934	0.8	0.751	0.835	0.6	0.834	0.89
6	stopped	0.953	0.878	0.911	0.844	0.755	0.795	0.612	0.65	0.63
24a	tugging speed	0.996	0.996	0.996	0.989	0.989	0.989	0.961	0.983	0.971
19	under way	0.997	0.997	0.997	0.99	0.989	0.989	0.967	0.985	0.975
9	unusual speed	0.988	0.975	0.981	0.968	0.947	0.957	0.948	0.919	0.933
2	within area	0.975	1.0	0.987	0.96	1	0.979	0.951	0.997	0.974

As the data removal process is aleatoric, the repeated removal of the event of interest due to the removal of the same AIS messages is highly unlikely for “movement ability affected”. This may indicate that the design of the detection pattern is also depending on a higher level of data availability and veracity than other detectors.

An obvious difference between this detector with strongly varying accuracy and the other detectors that could provide an explanation is the fact that “movement ability affected” is a ‘statically determined’ fluent while all other detectors are ‘simple’ fluents, as introduced in section 2.2. This could point to a weakness in the concept of interval based detectors, given that they typically require a minimum length, defined by a temporal threshold, in order to be detected. An argument against this hypothesis as unique cause is the fact that the number of TP and FP vary differently. While the number of TP decreases the number of FP increases from an average of 324.2 to 395.8 FP from 80 to 70% . Another reason for the drop of “movement ability affected” detections might be found in ”gap”, as this event terminates all other events.

Also surprising is the difference between “unusual speed” and “stopped”. Despite their conceptual similarity, the variation of detection accuracy of “unusual speed” in 6b is significantly smaller than the variations for the “stopped” event depicted in Figure 6a vary significantly stronger.



(a) “stopped” (b) “unusual speed”

Fig. 6: Two speed related events with different accuracy variations measured by mean, minimum and maximum True Positive (TP) and False Positive (FP).

For all events FPs are detected. For some events this behaviour is intuitive. As an example, the event “changingSpeed” is based on the reported or calculated speed of a vessel and a changing speed can be represented simplified as a unit step function where a number of messages imply a low speed followed by a number of messages implying a higher speed. The removal of the message on the limit between the two speed levels does not impact on the reported speed before and after the removed message, hence it can be assumed that the data removal leads only to a shift in the detection, creating a FP as the new limit between the two speed levels does not correspond to the old limit. For other events such as "stopped" depicted in Figure 6a and similarly for "underway" or "withinArea" FP detections are not expected as the removal of one AIS message is not changing the speed of another AIS message to zero or the removal of one AIS message inside a specific area is not impacting on the position of other AIS messages. These observations reveal perspectives for further analysis.

It seems possible, that events which occur less often in the dataset give a smaller variation of the response to data removal as the small number of 5 repetitions under represents unlikely events. The development of a specialized data removal method which is capable of removing targeted AIS messages known to be P detections in the reference dataset would allow for a better comparison of response behaviour of differently distributed events. While this approach would improve the comparability between events with similar detectors, e.g. "low speed" and "stopped", it potentially creates new inequalities between time point and interval based detectors. Another potential direction for future works is the development of a detector- or task-related metric. On the example of "changing speed" it becomes clear that a removal of an AIS message leads to a shift in the detection, thus both a FN and a FP. From an operational point of view this double-penalized behaviour might be still preferable in many cases before a missed detection which is not shifted in time and which does not create a FP.

5 Conclusions

The article summarises the development of an evaluation method that uses replicable data variation methods for the creation of pseudo-synthetic AIS datasets. For each vessel in the original dataset these degraded datasets contain a known and reduced number of AIS messages, which makes the detection of maritime events more difficult. The degraded datasets are then ingested into RTEC, an implementation of Event Calculus for maritime event pattern detection.
The detections obtained on degraded datasets are compared for each AIS message with the detections on the original dataset in order to derive True and False Positives as well as Recall and Precision for 90, 80 and 70% of the original dataset. The major part of the 12 evaluated maritime event patterns show a robust or even very robust behaviour, given that a data removal of 30% results only in a performance reduction of less than 20% for 9 pattern and for 6 of those only to a reduction of less than 4%. The performance reduction of 2 pattern is slightly over-proportional with respect to the amount of data removed. Only for 1 pattern the reduction of 30% leads to a complete non-detectability.

6 Acknowledgement

This work was supported by project datAcron, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 687591. The authors wish to thank the NATO Allied Command Transformation (NATO-ACT) for supporting the CMRE project on Data Knowledge and Operational Effectiveness (DKOE).

References

1. Andler, S., Fredin, M., Gustafsson, F., va Laere, J., Nilsson, M., Svenson, P.: SMARTracIn - a concept for spoof resistant tracking of vessel and detection of adverse intentions. In: SPIE Defense, Security, and Sensing. Orlando, FL (2009)
2. Artikis, A., Sergot, M.J.: Executable specification of open multi-agent systems. Logic Journal of the IGPL **18**(1), 31–65 (2010)
3. Artikis, A., Sergot, M.J., Paliouras, G.: An event calculus for event recognition. IEEE Trans. Knowl. Data Eng. **27**(4), 895–908 (2015)
4. Auslander, B., Gupta, K.M., Aha, D.W.: A comparative evaluation of anomaly detection algorithms for maritime video surveillance. In: Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense X. vol. 8019, p. 801907. SPIE (2011)
5. Iphar, C., Jousselme, A.L., Ray, C.: Pseudo-synthetic datasets in support to maritime surveillance algorithms assessment. In: proceedings of the VERITA Workshop, 19ieme Journées Francophones Extraction et Gestion des Connaissances (EGC) 2019 (Jan 2019)
6. ITU: Technical characteristics for an automatic identification system using time-division multiple access in the vhf maritime mobile band (2010)
7. Jousselme, A.L., Maupin, P.: Comparison of uncertainty representations for missing data in information retrieval. In: Proceedings of the 16th international conference on information fusion. pp. 1902–1909. IEEE (2013)
8. Kowalski, R.A., Sergot, M.J.: A logic-based calculus of events. New Generation Comput. **4**(1), 67–95 (1986)
9. Lavesson, N., Davidsson, P.: Evaluating learning algorithms and classifiers. International Journal of Intelligent Information and Database Systems **1**(1), 37–52 (2007)
10. Margineantu, D.D., Dietterich, T.G., et al.: Bootstrap methods for the cost-sensitive evaluation of classifiers (2000)
11. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Machine learning **42**(3), 203–231 (2001)
12. Przymusinski, T.: On the declarative semantics of stratified deductive databases and logic programs. In: Foundations of Deductive Databases and Logic Programming. Morgan (1987)
13. Ray, C., Dréo, R., Camossi, E., Jousselme, A.L., Iphar, C.: Heterogeneous integrated dataset for maritime intelligence, surveillance, and reconnaissance. Data In Brief, (Accepted for publication, In Press) (2019). https://doi.org/10.1016/j.dib.2019.104141
14. Riveiro, M., Falkman, G.: Supporting the analytical reasoning process in maritime anomaly detection: Evaluation and experimental design. In: 2010 14th International Conference Information Visualisation. pp. 170–178. IEEE (2010)
15. Roy, J., Davenport, M.: Exploitation of maritime domain ontologies for anomaly detection and threat analysis. In: Proc. of the IEEE int. Waterside Security Conference (WSS) (2010)
16. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)