Pseudo-synthetic datasets in support to maritime surveillance algorithms assessment

Clément Iphar*, Anne-Laure Jousselme*, Cyril Ray**

*NATO STO CMRE, Viale San Bartolomeo, 400, La Spezia, Italy
{ clement.iphar ; anne-laure.jousselme }@cmre.nato.int
**Naval Academy Research Institute, 29240 Brest Cedex 9, France
cyril.ray@ecole-navale.fr

Abstract. In the maritime domain, the ever-growing availability of data from systems such as the Automatic Identification System (AIS) enables the monitoring of worldwide maritime activities. The processing of huge amounts of spatial and temporal data rises issues linked to Big Data analyses. In particular, this paper focuses on the lack of veracity of data, and specifically on the characterisation of AIS dataset quality. In this paper, we aim at producing datasets either with a known and controlled veracity levels, or with added spatial events. Such quantified variations taking into account the initial quality level of the dataset and the desired level of degradation are performed following the mechanisms enabling data degradation, data improvement or event injection. A library has been developed, enabling the generation of those pseudo-synthetic datasets to be further used as benchmark for the assessment of algorithms solving Maritime Situation Awareness (MSA) issues such as anomaly detection.

1 Motivation

Traditionally, four challenges are associated with Big Data, namely the four V's: Volume, Velocity, Variety and Veracity. The Volume refers to the amount of data to be handled; for instance, it is estimated that companies like *Walmart* collect more than 2.5 petabytes of data per day from their customers and the total amount of data created each day overcomes the exabyte (McAfee and Brynjolfsson, 2012). The Velocity concerns the gathering and processing effectiveness. The Velocity handling ability of gathering and exploiting data is even more important than the Volume handling, as it is more applicative and enables quick searches and thus companies to be more competitive. The Variety challenge covers the fact that data takes several formats, such as images, text messages, sensor data, updates on social media, signals, amongst others (McAfee and Brynjolfsson, 2012). Most of the corresponding data sources have been developed recently, mostly with the rise of digital information. The Veracity challenge is mainly linked to the relation of data to the world. It represents the fact for a datum to be truthful, *i.e.* to correctly depict the world in the way that it is expected to.

Algorithms are developed to process data and provide information as a result. More precisely in the maritime domain, algorithms for anomaly detection (Hadzagic and Jousselme,

2016), route extraction (Pallotta et al., 2013), situational awareness (Morel et al., 2009), trajectory analysis (Andrienko et al., 2016), knowledge discovery (Fernandez Arguedas et al., 2014) or vessel prediction (Vanneschi et al., 2015) are developed, amongst other purposes. Those analyses are mainly based on the AIS (Automatic Identification System), an international standard message-based system introduced in IMO (2004) which is known for having shortcomings on the quality of the data it transmits (Harati-Mokhtari et al., 2007; Iphar et al., 2016). Similar work in data simulation can be found in domains such as social sciences, health sciences (Ping et al., 2017) or bio-surveillance (Lotze et al., 2007).

In this paper, we propose a method for the generation of datasets with controlled veracity levels, so that the algorithms can be tested and challenged in their ability to deal with different levels of data quality. To this end, we first assess the veracity level of a given dataset, considering the nature of the dataset, the nature of its attributes and the dependencies between them.

Once the original veracity level of one dataset assessed, it is possible to adjust the veracity level of any dataset in order for it to fit our requirements. In this respect, the production of datasets involving both real data for which the veracity has been assessed and synthetic data for which the veracity is known enables us to create any dataset with any given veracity level, in accordance with the needs of an algorithm assessment. In the following of this paper, the datasets including both actual and synthetic data will be called pseudo-synthetic datasets.

In order to be able to generate in a convenient and controlled way the desired datasets, a library for pseudo-synthetic dataset generation has been implemented. Its various functions as well as an algorithm assessment based on this library, will be presented in this paper.

The paper is structured as follows: In Section 2, we present the general principles of data generation, including the typology of the data fields, and the concepts around data modification. In section 3, we present the families of functions for the generation of pseudo-synthetic dataset, with their formalisation and parametrisation. In section 4, we present the implementation of a library enabling the generation of such dataset, with the various functions and some application examples.

2 Principles of data generation

The research described in this paper has been applied on ship information collected through the Automatic Identification System (an embedded device that enable ships to broadcast their position and nominative information via radio communication), prepared together with correlated data aligned in space and time (Ray et al., 2018). The dataset [1] contains four categories of data: navigation data (vessel positions acquired automatically by an AIS receiver), vessel-oriented data (public, official nominative vessel position), geographic data (cartographic, topographic or regulatory context of vessel navigation), and environmental data (weather and ocean data from forecast models and from observations). It covers a time span of six months, from October 1^{st}, 2015 to March 31^{st}, 2016 and provides ship positions over the Celtic sea, the North Atlantic ocean, the English channel, and the Bay of Biscay (France).

1. C. Ray, R. Dréo, E. Camossi, A.-L. Jousselme, Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance (Version 0.1). Data set. Licence CC-BY-NC-SA-4.0. Zenodo. doi.org/10.5281/zenodo.1167594, February 2018

2.1 The Automatic Identification System

The Automatic Identification System (AIS) has been developed as an anti-collision system, as it enables the surrounding stations to receive the messages that a station sends. However, since the inception of fast worldwide communication systems, Vessel Traffic Services (VTS) and shipowners use this system to track, locate and identify the vessels, taking advantage of the cooperative transmission of information that the vessels operate. As defined by IMO (2004), it is mandatory for some vessels (falling under usage and tonnage regulations) to report their AIS information which is classically separated into a static and a dynamic part. The static information contains in particular the Name, the MMSI (Maritime Mobile Service Identity) and IMO (International Maritime Organisation) numbers, call sign, width, length, expected time of arrival, destination, ship type, type of cargo ; whereas the dynamic information, more frequently updated under position report messages, contain MMSI, Latitude, Longitude, Course Over Ground, Speed Over Ground [2], Navigation Status, so all the dynamic information, enabling the receivers to reconstruct the trajectory of the vessel.

All those attributes of AIS messages have various types, as some represent physical values, other are identifiers, text messages or categorical data.

Let X_n^m represent an AIS dataset built of n contacts (messages) with m attributes (or features). The i^{th} contact will be denoted by $\mathbf{x}_i$ and can be seen as a row vector of m columns. The element of this vector corresponding to the j^{th} column will be denoted by x_{ij}. The AIS dataset can thus be seen as a table of heterogeneous values $X_n^m = \{x_{ij}\}_{i=1:n}^{j=1:m}$.

Let J be a set of column indexes, $J \subseteq \{1, \cdots, m\}$. For clarity, columns can be referred by their attribute name, so $J \subseteq \{\text{MMSI}, \cdots, \text{speed}, \cdots, \text{time}\}$. Each of the m columns has an associated attribute, and $\forall j \in \{1, \cdots, m\}$ we define a domain Θ_j in which the attribute takes its values. Let I be a set of row indexes, $I \subseteq \{1, \cdots, n\}$.

2.2 AIS dataset characterisation

Based on this heterogeneous dataset, we set some quantitative and statistical features for the characterisation of the dataset. The dataset is either taken as a whole or divided into subsets. Those partitions can be built on the basis on the identity of the vessel (the national identity number), the vessel type or subtype (e.g. fishing vessel, cargo vessel), the status (considered underway, *i.e.* over 3knots, or not) amongst the list of 12 status, including at anchor, fishing, mooring, etc. Also, subsets of combinations are considered (e.g. data discriminated on identity and Underway status). Each one of those subsets, alongside with the whole dataset, undergoes a series of analysis. Those analyses consider each data field and count the number of *null* entries, of zero entries or of non-valid values for each field. In addition, when applicable to the given field, the mean, the median, the mode, the standard deviation, the skewness and the kurtosis values are computed, enabling the characterisation of the data field within the considered subset.

2.3 Data Generation

In order to add data to the dataset, data augmentation techniques can be followed. Indeed, data augmentation can be defined as the practice of applying techniques to a given dataset in

2. *i.e.* course and speed w.r.t. the North and the seabed, respectively

order to synthetically expand it. In a dataset, data augmentation can be twofold: either by the addition of contacts (lines) to the dataset, or by the addition of attributes (columns) to the dataset. Techniques for the addition of contacts include interpolation, extrapolation, or the use of Gaussian random fields for the propagation of data characteristics to the newly created data. Techniques for the addition of attributes include all labelling techniques (including classification techniques in machine learning). However, most techniques are domain-specific and more particularly type-specific, therefore the beforehand characterisation of the typology of data is of paramount importance.

In order to remove data from the dataset, several rules can be followed, and those rules can follow mechanism that lead to the way fields or entries are missing from a dataset. Those mechanisms are the result of the variety of the causes of the fact that data is missing that can be identified. The absence of a piece of information can be the result of a lack of knowledge, a bad reception of data, the fact that an attribute does not applies to some of the entries. In literature, different categories are distinguished, depending on the cause of missingness. Three mechanisms are generally distinguished (Jousselme and Maupin, 2013): Missing At Random (MAR), where the mechanism is conditionally independent of the missing values given the observed variables, Missing Completely At Random (MCAR), a particular case of MAR in which the mechanism is independent from the domain variables and Missing Not At Random (MNAR) if the missing data depends on the unobserved values.

In order to modify data inside the dataset, a series of good practices must be followed. Unless it is done purposely, the global coherence of the data must be respected, particularly in the cases of data shifting, where overlaps with landmasses must be checked, as well as the possible creation of events such as unwanted collision, or behaviours that would depict a contravention to the maritime rules of navigation.

3 Data pseudo-synthesis

3.1 Families of functions

Three main families of data pseudo-synthesis functions can be distinguished: the data degradation, the improvement and the event injection.

Data degradation: Data degradation consists in all kind of data modification that lowers the level of veracity of data. In the case of AIS messages, this lowering can be handled in three different ways: the removal of whole data contacts (one row), the removal of whole data attributes (one column) or the addition of noise in data. In the case of a removal of contacts, the cardinality of the dataset decreases as some entries are removed, on a targeted basis (e.g. removal of a trajectory) or on a random basis (e.g. removal of 10% of the contacts). In the case of a removal of attributes, the cardinality does not vary but the number of columns decreases. The loss of information induced constitutes a data degradation. In the case of the noise addition, neither the cardinality nor the number of attributes are modified, but values of the dataset are blurred and therefore the data degraded. This noise can be applied according to any probabilistic distribution.

Data improvement: Data improvement consists in all kind of data modification that increases the level of veracity of data. In the case of AIS messages, this improvement can be handled in two different ways: the addition of contacts and the addition of attributes. In the case of the addition of contacts, the cardinality of the dataset increases as some entries are added. Those entries are synthetic entries simulating and replacing the messages that would have been received, should the system emission and the antenna reception been perfect. Either all missing contacts or only a subset of them, selected (targeted on one particular trajectory) or random (based on a percentage of addition) can be added. In the case of the addition of attributes, the cardinality does not vary but the number of columns does, as it increases. This addition of information can be the result of, for example, a labelling operation. It adds information to the dataset and therefore constitutes a data improvement.

Event injection: Event injection consists in any modification of the "story" that the data tell, especially by addition of complete trajectories. In the case of AIS messages, this event injection can take two main forms: either the injection of a real event identified from real AIS data and shifted in time and space to match the current dataset or the injection of synthetic data directly in the existing fields. The latter constitutes of a targeted injection by putting either the value of a field to a fixed given value (including the *null* value) or adding an offset value to the existing value. The event injection consists in the synthesis of specific events such as a collision, a near-collision or a rendezvous, located in time and space so that it can model a specific story and creates the corresponding data. Those scenarios depend on a handful of parameters (for instance the angle of approach for the collision, the nearest distance for a near-collision, the time of the meeting for a rendezvous, amongst others), which provide a high flexibility in modelling precisely a high variety of events with synthetised data.

3.2 Formalisation

The formalisation of the data modification processes allows a precise definition of the basic functions that come into play. Indeed, the combination of basic functions is the key point of any sophisticated pseudo-synthetic dataset generation.

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. The original dataset is denoted by $\bar{X}$, and synthetic dataset is denoted by $\tilde{X}$.

In order to apply some data modifications to a subset of the dataset, it is necessary to define those subsets, which can be isolated by applying constraints on attributes values. Let us denote by $\Gamma \in \times_j(\Theta_j)$ a set of constraints, by A the subset of rows ($A \subseteq I$) for which the columns complying to this set Γ of constraints, and by $X_{n,A}^m$, the corresponding dataset.

Some operations can be performed on those datasets, as defined by a series of functions $f_k : D \to D$. For instance, J being a set of columns, the removal of the corresponding set from the original dataset is defined by:

$$f_1(J, \bar{X}_n^m) : \qquad \bar{X}_n^m \mapsto X_n^{m\setminus\{J\}} = X_n^{m'} \text{ with } m' = m - Card(J) \tag{1}$$

More particularly, the removal of the speed column will be defined by:

$$f_1(speed, \bar{X}_n^m): \qquad \bar{X}_n^m \mapsto X_n^{m\setminus\{speed\}} \tag{2}$$

The addition of rows, representing for instance the addition of an event of $n' - n > 0$ rows would be defined by:

$$f_2: \qquad \bar{X}_n^m \mapsto X_{n'}^m = \bar{X}_n^m \frown \tilde{X}_{n'-n}^m \tag{3}$$

where $\frown$ denotes the append operation. In addition, let us describe the modification of the values of the data fields in a function f_3 denoted by:

$$\begin{aligned} f_3: \qquad & \forall i,j \in I \times J, x_{ij} \in X_n^m \\ & x_{ij} \mapsto x'_{ij} \end{aligned} \tag{4}$$

where the value x'_{ij} replaces the original value x_{ij} in the dataset.

3.3 Parametrisation

In order to take into consideration all the various parameters of any of the functions, a general formalisation taking into consideration the various parameters that can be set can be proposed. In this section, we propose the example of the addition of noise to data. In this case, the parameters are: the subset $\bar{X}_{n,A}^m$ on which the modification is applied, the percentage p of entries (marked as $\{\}^p$, representing the fact that $p\%$ of the data are concerned, selected by a random draw with uniform distribution), x_m^k and ${x'_m}^k$ which stand for the value x of the k^{th} attribute respectively after and before the noise addition, $N_s^z(x)$ which stands for the application of a noise following the law z and the law-related parameters s (for instance the standard deviation). Then a general definition of the noise addition can be described by:

$$f_n(p, z, s, k): \qquad \bar{X}_{n,A}^m \mapsto {X'_{n,A}}^m : \left\{ {x'_n}^k = N_s^z(x_n^k) \right\}^p \tag{5}$$

where s is denoted by Σ_m^k when a set of user-defined values for noise application is used and $\hat{S}_m^k$ when the values for noise application are computed from the dataset itself. In our process we consider the addition of a normal noise, noting $\omega = \mathcal{N}(x_m^k, \sigma)$ the normal distribution centered on the value with a standard deviation of σ.

As a consequence, such noise addition performed on 50% of a subset $\bar{X}_{n,A}^m$ of all vessels belonging to a vessel V, to which a normal law ω noise is applied, computing its standard deviation from the dataset ($\hat{S}_m^k$) is denoted by:

$$f_n(50\%, \omega, \hat{S}, k): \qquad \bar{X}_{n,\{V\}}^m \mapsto {X'_{n,\{V\}}}^m : \left\{ {x'_n}^k = N_{\hat{S}_n^k}^{\omega}(x_n^k) \right\}^{0.5} \tag{6}$$

3.4 Function composition

In order to create elaborated events, a composition of the basic functions previously defined must be performed. In this section, a simple case is proposed: we want to shift a trajectory in time and space in a dataset. We isolate this trajectory in the subset $\bar{X}_{n,A}^m$, applying the *ad-hoc* restrictions Γ so that only the chosen points are selected (restriction on identity and on time).

Then we define three offset functions which take the three values that we want to change: the latitude, the longitude and the time. Let us denote by Δ_{lat}, Δ_{lon} and Δ_{time} those three values that can take positive and negative values as long as their application does not violate AIS specifications. Let us define by:

$$f_\alpha : \quad \bar{X}^m_{n,A} \mapsto {X'_{n,A}}^m : \left\{ {x'_n}^{lat} = x_n^{lat} + \Delta_{lat} \right\} \tag{7}$$

$$f_\beta : \quad \bar{X}^m_{n,A} \mapsto {X'_{n,A}}^m : \left\{ {x'_n}^{lon} = x_n^{lon} + \Delta_{lon} \right\} \tag{8}$$

$$f_\gamma : \quad \bar{X}^m_{n,A} \mapsto {X'_{n,A}}^m : \left\{ {x'_n}^{time} = x_n^{time} + \Delta_{time} \right\} \tag{9}$$

Let us define by $g : D \mapsto D$ as the function operating this data shifting. g will then be denoted by

$$g(\bar{X}) = (f_\gamma \circ f_\beta \circ f_\alpha)(\bar{X}) \tag{10}$$

4 A library for pseudo-synthetic dataset generation

4.1 The purpose of the library

For the implementation of the library, the R programming language was used for its ability to perform large-scale statistical computation, the existence of libraries for database querying, data handling and plotting capabilities.

The library enables the use of either one single function or a series of functions with given parameters so that the resulting dataset matches the expectations in terms of data veracity levels.

The purpose of the library is to enable, with the smallest number of functions, any possible modification to the dataset, either concerning a degradation, an improvement or an event injection. Each of the modifications can fall in two main families: the variations in size (addition or removal of rows or of columns) or variations in values (random or targeted modification of data fields). Figure 1 shows the various functions according to the nature of the data modification.

4.2 The functions

As presented in Figure 1, the library is composed of seven functions, each one linked to one of the three families, and corresponding to a specific action in the process of generation of a pseudo-synthetic dataset with controlled veracity. The main features of each function are also presented in Figure 1 next to each box (e.g. the percentage of data to be removed, in `remove.r`). Some additional parameters, of lesser importance, can be chosen for each of the functions. Those parameters are the basis of the control of the veracity level variations. In this section, the first function is entirely presented, with its description, its algorithm and a figure presenting its effects on a dataset. For the other functions, only their description is presented.

Data addition: The function `add.r` is an improvement function that consists in the addition of rows in the dataset. This function, applied to the subset A defined above, synthesises $k\%$ of the missing data (*i.e.* the dataset is filled by data that should have been received in the first

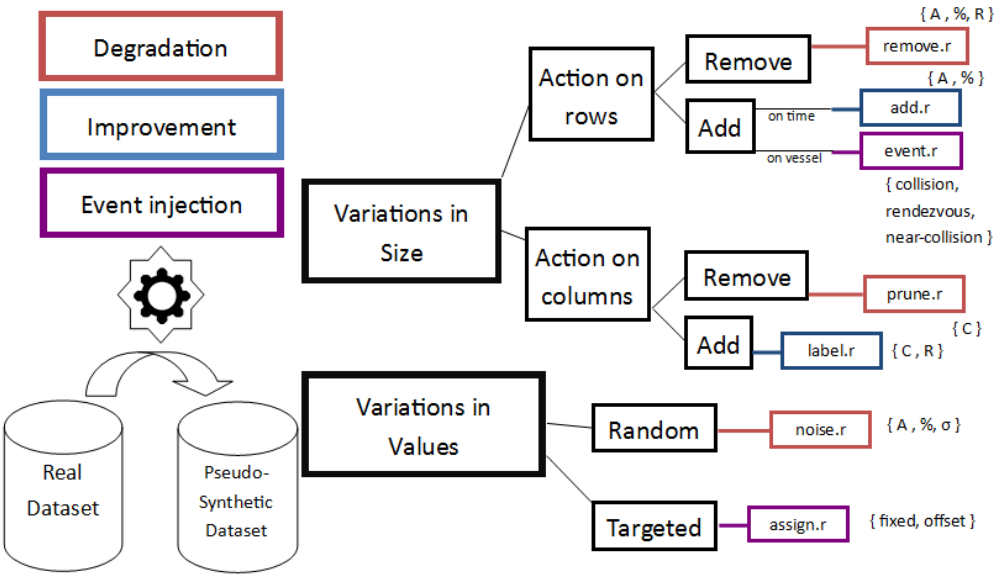

FIG. 1 – *Data variations and associated functions*

place but that was not because of the imperfections of the system). The algorithm 1 presents the implementation of this function. This algorithm requires the percentage of messages to add $k \in [0, 1]$, the list of reporting rates values **r** (the time between two consecutive messages, between 2 seconds and 3 minutes, which changes according to the vessel speed), the total set of initial messages $\bar{X}$, the results of the statistical analysis S, the function f_a that computes the number of missing points between two contacts and the function f_b that computes the values in the fields of the synthetised message, in accordance with the values in S. Figure 2 presents the result of the application of this algorithm on a dataset, where 80% of the missing points are added.

Data removal: The function `remove.r` is a degradation function that consists in the removal of rows in the dataset. This function, applied to the subset of rows A defined above, consists of the removal of $k\%$ of the data, following a set of rules R. This set of rules 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 of some data fields (MNAR).

Event addition: The function `event.r` is an event injection function that consists in the addition of rows in the dataset. Contrarily to `add.r`, this function synthesises chosen events in the dataset. The three events are collision (with the parameters such as the number of collision, the targeted location, the angles of approach, the nature of the vessel, the speed of the vessel), near-collision (with the parameters such as the number of near-collision, the targeted location,

Algorithm 1 Add $k\%$ of missing data

Require: $k, \mathbf{r}, \bar{X}, S, f_a, f_b$
  for all $i \in \bar{X}$ **do**
    if $\exists j \in \bar{X} : x_i^{(MMSI)} = x_j^{(MMSI)}$ & $\Delta_{i,j}(t) = \min_{\forall v \in \bar{X}} \Delta_{i,v}(t)$ & $\Delta_{i,j}(t) > 0$ **then**
      q $\leftarrow f_a(\mathbf{x}_i, \mathbf{x}_j, \mathbf{r})$ {if next message exists, compute number of missing}
      if $q \geq 1$ **then**
        for z from 1 to q **do**
          if random(0,1) $\leq k$ **then**
            $\mathbf{x}_z = f_b(\mathbf{x}_i, S)$ {if random draw favorable, values are computed}
            $\bar{X} \leftarrow \mathbf{x}_z$ {new entry pushed in semi-authentic dataset}
          end if
        end for
      end if
    end if
  end for

the angle of approach, the distance of nearest approach) and rendezvous (with the parameters such as the number of rendez-vous, their location, the angle of approach and departure of the synthetised vessel, the duration of the rendezvous and the length of the decelerating and accelerating phases).

Prune dataset: The function `prune.r` is a degradation function that consists in the removal of columns in the dataset. It is applied on a determined set of columns C.

Label dataset: The function `label.r` is an improvement function that consists in the addition of columns in the dataset. This function takes a set C of columns to be added and fills the data within the fields of the newly created columns according to a set of rules R (e.g. classification).

Noise addition: The function `noise.r` is a degradation function that consists in the replacement of the value of data fields of a subset of rows A of the dataset defined as above, by the application of a noise to $k\%$ of the corresponding entries. The noise can follow various laws according to the type of data field, however today only a normal noise is applied to numeric physical values, with a selected set of standard deviations σ, different for each of the data attributes and assign either with a fixed user-defined value or extracted from a statistical analysis of the whole dataset or any of its subsets as presented in Section 2.2.

Field value assignment: The function `assign.r` is an event injection function that consists in the replacement of the value of data fields in a targeted way. In this case, a targeted data field (or a set of targeted data fields) will be set to a given value. This value can either be the *null* value, any fixed value or the current value of the field to which is added a fixed offset value.

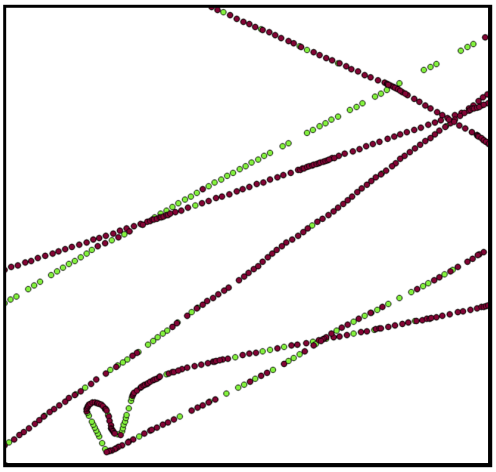

FIG. 2 – *Application of a row addition process with* $k = 0.8$. *Purple points: original data* ($\bar{X}$). *Green points: new points* ($\tilde{X}$). *Resulting dataset contains both* ($X = \bar{X} \frown \tilde{X}$)

4.3 Experiments

The generation of pseudo-synthetic datasets in support to algorithms assessment has been experimented in the context of maritime surveillance. A real AIS-based dataset has been prepared with generated maritime events and presented to naval officers. The purpose of the experiment was to assess the capability of maritime experts to properly understand the scenario presented to them and more specifically to assess the evolution of this understanding with the evolution of the veracity of the dataset. Such an objective obviously requires to synthetise datasets with a controlled veracity level, including multiple spatial events.

Based on preliminary exchanges with experts we retained a collision avoidance use case. In that context, the aim of a naval officer is to prevent and avoid a collision involving vessels. Scenarios have been sketched considering four similar maritime situations on which an expert can be mistaken: collision, tugging, near-collision and rendez-vous.

Three of these maritime events have been subsequently combined in three datasets of 30 minutes, each was created using the library functions, correctly parametrised so that it matches the story decided. In total, 2 collision events, 2 near-collision events and 2 rendez-vous events were generated over the 3 datasets, as well as 20 shifted trajectories and some targeted data assignments. These events have been organised in three scenarios presented to experts:

— Scenario 1: include a collision (minute 20) and a rendez-vous (minute 24)
— Scenario 2: include a rendez-vous (minute 15) and near-collision (minute 25)
— Scenario 3: include a near-collision (minute 23) and a collision (minute 26)

Figure 3 illustrates trajectories that have been created for these experiments, according to the stories of scenarios 2 and 3.

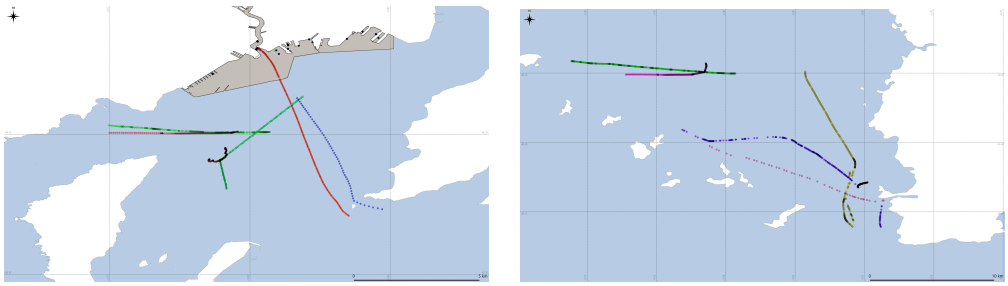

FIG. 3 – *Maritime situation of Scenario 2 (left) and 3 (right)*

The experiments have been organised in a timeframe of one week, involving four maritime domain experts and two naval cadets. The use case scenarios were presented to the experts and they were invited to comment on their analysis of the maritime situations presented to them. Their comments and all the results of this experiment are described in Zocholl et al. (2018) [3].

5 Conclusions

In this paper, the general principles of data generation where presented, with specific methods for AIS data degradation, data improvement and event injection with known and controlled veracity level variations. The generation of datasets composed of both original AIS data and synthetic AIS data, called pseudo-synthetic datasets, is presented under the form of data modification functions that aim at keeping control over veracity. A library has been developed which implements these methods enabling the generation of those pseudo-synthetic datasets with controlled veracity levels. Those pseudo-synthetic datasets are meant to be used as benchmark data for the assessment of MSA algorithms.

Acknowledgement

This work is supported by project datAcron (H2020-ICT-2015), which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N^o 687591.

References

Andrienko, G., N. Andrienko, C. Claramunt, G. Fuchs, and C. Ray (2016). Visual analysis of vessel traffic safety by extracting events and orchestrating interactive filters. In M. Vespe and F. Mazzarella (Eds.), *Proceedings of the Maritime Knowledge Discovery and Anomaly Detection Workshop*, JRC Conference and Workshop Reports, pp. 44–47.

3. available at http://datacron-project.eu/

Fernandez Arguedas, V., G. Pallotta, and M. Vespe (2014). Automatic generation of geographical networks for maritime traffic surveillance. In *Proceedings of the 17th International Conference on Information Fusion*. ISIF.

Hadzagic, M. and A.-L. Jousselme (2016). Contextual anomalous destination detection for maritime surveillance. In M. Vespe and F. Mazzarella (Eds.), *Proceedings of the Maritime Knowledge Discovery and Anomaly Detection Workshop*, JRC Conference and Workshop Reports, pp. 62–65.

Harati-Mokhtari, A., A. Wall, P. Brooks, and J. Wang (2007). Automatic Identification System (AIS): A Human Factors Approach. *Journal of Navigation*.

IMO (2004). International Convention for the Safety of Life at Sea. Technical report, IMO.

Iphar, C., A. Napoli, C. Ray, E. Alincourt, and D. Brosset (2016). Risk Analysis of falsified Automatic Identification System for the improvement of maritime traffic safety. In T. B. Lesley Walls, Matthew Revie (Ed.), *Proceedings of the ESREL 2016 Conference*, pp. 606–613. Taylor & Francis.

Jousselme, A.-L. and P. Maupin (2013). Uncertainty representations for information retrieval with missing data. In ISIF (Ed.), *Proceedings of the 16th international conference on Information Fusion*.

Lotze, T., G. Shmueli, and I. Yahav (2007). Simulating multivariate syndromic time series and outbreak signatures. Robert H. Smith School Research Paper No. RHS-06-054.

McAfee, A. and E. Brynjolfsson (2012). Big data: the management revolution. *Harvard Business Review 90*, 60–66.

Morel, M., A. Napoli, A. Littaye, M.-P. Gleizes, V. Bazin, B. Alhadef, C. Scapel, B. Leroy, J. Lebrevelec, and D. Dejardin (2009). *Surveillance et contrôle des activités des navires en mer*. CNRS Editions.

Pallotta, G., M. Vespe, and K. Bryan (2013). Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. *Entropy 15*(6), 2218–2245.

Ping, H., J. Stoyanovich, and B. Howe (2017). Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the SSDBM'17 conference*.

Ray, C., R. Dréo, E. Camossi, A.-L. Jousselme, and C. Iphar (2018). Heterogeneous integrated dataset for maritime intelligence, surveillance, and reconnaissance. *Data In Brief , Accepted for publication*.

Vanneschi, L., M. Castelli, E. Costa, A. Re, H. Vaz, V. Lobo, and P. Urbano (2015). Improving Maritime Awareness with Semantic Genetic Programming and Linear Scaling: Prediction of Vessels Position Based on AIS Data. In A. M. Mora and G. Squillero (Eds.), *Applications of Evolutionary Computation*, Volume 9028 of *Lecture Notes in Computer Science*, pp. 732–744. Cham: Springer International Publishing.

Zocholl, M., A.-L. Jousselme, R. Dréo, C. Ray, C. Iphar, F. de Rosa, E. Camossi, G. Keraudren, and F. Rozé (2018). Maritime final validation, H2020 datacron d5.6.

Résumé

La surveillance des activités maritimes mondiales est rendue possible par l'accessibilité toujours plus importante de données issues de systèmes tels que le Système d'Identification Automatique (AIS). L'analyse a grande échelle de données spatio-temporelles a pour conséquence l'apparition des problématiques liées au traitement des mégadonnées. En particulier, cet article porte sur les carences en véracité de la donnée, et plus particulièrement sur la description de la qualité d'un jeu de données AIS. Le but de cet article est de proposer une méthode permettant la génération de jeux de données ayant une véracité connue et contrôlée, ou comprenant des événements additionnels. Ces variations quantifiées, prenant en compte le niveau de qualité initial du jeu de données et le niveau de dégradation désiré, sont réalisées selon des mécanismes permettant la dégradation ou l'amélioration de la donnée, ou l'ajout d'événements aux données. Une bibliothèque de code a été réalisée, permettant la génération de jeux de données pseudo-synthétiques, afin qu'ils puissent être utilisés en tant que jeux de données de référence pour l'évaluation d'algorithmes liés à la connaissance de la situation maritime, tels que des algorithmes de détection d'anomalies maritimes.