University Institute of Lisbon

Department of Information Science and Technology

Maritime Modular Anomaly Detection Framework

Tomás Manuel Cardoso Machado

Dissertation submitted as partial fulfilment of the requirements for the degree of:

Master in Computer Science Engineering

Supervisor

João Carlos Amaro Ferreira, PhD

ISCTE-IUL

Co-Supervisor

Rui Maia

Instituto Superior Técnico

October 2018

Resumo

Detetar anomalias marítimas é uma tarefa extremamente importante para agências marítimas á escala mundial. Com o número de embarcações em mar crescendo exponencial, a necessidade de desenvolver novas rotinas de suporte ás suas atividades e de atualizar as tecnologias existentes é inegável. MARISA, o projeto de Conscientização da Vigilância Integrada Marítima, visa fomentar a colaboração entre 22 organizações governamentais e melhorar as capacidades de reação e tomada de decisões das autoridades marítimas. Este trabalho descreve as nossas contribuições para o desenvolvimento do toolkit global MARISA, que tem como âmbito a deteção de anomalias marítimas. Estas contribuições servem como parte do desenvolvimento da Modular Anomaly Detection Framework (MAD-F), que serve como um data-pipeline completo que transforma dados de embarcações não estruturados em potenciais anomalias, através do uso de métodos eficientes para tal. As anomalias consideradas para este trabalho foram definidas através do projeto MARISA por especialistas marítimos, e permitiram-nos trabalhar em necessidades reais e atuais do sector. As funcionalidades desenvolvidas serão validadas através de exercícios marítimos reias. No estado atual do MAD-F acreditamos que este será capaz de apoiar agências marítimas, e de posteriormente ser integrado nos sistemas dos mesmos.

Palavras-Chave: *Framework* Maritima, Detecção de Anomalias, Dados AIS.

Abstract

Detecting maritime anomalies is an extremely important task for maritime agencies around the globe. With the number of vessels at seas growing exponentially, the need for novel automated methods to support them with their routines and upgrade existing technologies is undeniable. MARISA, the Maritime Integrated Surveillance Awareness project, aims at fostering collaboration between 22 governmental organisations and enhance the reaction and decision-making capabilities of the maritime authorities. This work describes our contributions to the development of MARISA's common toolkit for the detection of maritime anomalies. These efforts, as part of a Masters' dissertation, lead to the development of the Modular Anomaly Detection Framework, MAD-F, a full data pipe-line which applies efficient and reliable routines to raw vessel navigational data in order to output potential maritime vessel anomalies. The anomalies considered for this work were defined by the experts from various maritime institutions, through MARISA, and allowed us to implement solutions given the real needs in the industry. The MAD-F functionalities will be validated through actual real maritime exercises. In its current state, we believe that the MAD-F is able to support maritime agencies and be integrated into their legacy systems.

Keywords: Maritime Framework, Anomaly Detection, AIS data.

Acknowledgements

I would like to acknowledge my supervisors, Rui Maia, and Professor João Ferreira for their constant supervision and assistance. To all the people at INOV-INESC Inovação for providing me a great work environment, especially to Dária and Gonçalo for unarguably supporting my caffeine addiction.

Most importantly, I would like to show my deepest gratitude, to my parents, Ana Maria and Rogério for unconditionally supporting and sponsoring me all throughout my academic journey.

A special "thank you bro", to the family I was fortunate to choose, PhD David Carvalho, Architect Ruben Soares, and Chief Mate Alexandre Bota.

And lastly, I would like to thank my girlfriend Floor, whom I love, for always being there when I fall.

Also, to my sister Mariana, who made the dinner at the time of writing.

Contents

List of Figures

List of Tables

Abbreviations

MARISA	**MAR**itime **I**ntegrated **S**urveillance **A**wareness
AIS	**A**utomatic **I**dentification **S**ystem
VHF	**V**ery **H**igh **F**requency
SOLAS	**S**afety **o**f **L**ife **a**t **S**ea
IMO	**I**nternational **M**aritime **O**rganization
AD	**A**nomaly **D**etection
MAD-F	**M**odular **A**nomaly **D**etection - **F**ramwork
BP	**B**ehavioural **P**oint
AR	**A**nomaly **R**equirement
UN	**U**nited **N**ations
VTS	**V**essel **T**raffic **S**ystem
MMSI	**M**aritime **M**obile **S**ervice **I**dentity
ETA	**E**stimated **T**ime of **A**rrival
SOG	**S**peed **O**ver **G**round
COG	**C**ourse **O**ver **G**round
NMEA	**N**ational **M**arine **E**lectronics **A**ssociation
S-AIS	**S**atellite **- A**utomatic **I**dentification **S**ystem

Chapter 1

Introduction

Approximately 90% of global trade relies on the international shipping industry. Consequently, the ocean is a vital platform for the world economy. Currently, there are approximately $50,000$ merchant ships trading internationally. Given the current demand, this number is bound to increase[1]. Not all such activity is legitimate, with some of it resorting to organised crime and various other illicit schemes that prevail in the maritime domain. Examples of this may be given by piracy, drug trafficking, illegal immigration, arms proliferation and illegal fishing. The definition of maritime safety is a complex endeavour and widely acknowledged as a transnational task [1].

Tracking people and objects within a geographical space has become a ubiquitous challenge. Automatic Identification System (AIS) is an automated tracking system that broadcasts information through very high frequency (VHF) bands, which ultimately assist vessels in navigation. Imposed by the IMO (International Maritime Organisation), every SOLAS (Safety of Life at Sea) vessel must be equipped with such a device. Autonomously broadcast AIS messages contain *kinematic* information such as the ship location, speed, heading, rate of turn, destination and estimated arrival time, as well as *static* information, including the ship name, ID, type, size. AIS messages can be transformed into useful information for maritime traffic manipulations such as vessel path prediction and collision

[1] http://ics-shipping.org/shipping-facts/shipping-and-world-trade

avoidance. For these reasons, the AIS tracking system plays a central role within the development of future autonomous maritime navigation systems [2].

The introduction of AIS in the maritime domain lead to an exponential increase of the volume of vessel trajectory data, making human analysis and evaluation of such data extremely inefficient. Therefore, new effective ways to automatically mine this data are of extreme importance for the future of nautical surveillance. Despite its advancements, mining maritime trajectory data still presents several challenges. Firstly, such data contains uncertainty typical of moving objects. Geo-referenced locations of trajectories constructed by location sensing techniques are prone to spatial uncertainty due to computational error and signal degradation or loss associated with the positioning device. Temporal uncertainty may be generated by different sampling rates and temporal lengths [3]. Secondly, maritime traffic is not constrained to roads - vessels are free to navigate in open waters as long as legal restrictions are observed. These situations hint at the inherent complexity of detecting trajectory anomalies. Nevertheless, vessels tend to be observed travelling in the most economic route, to the advantage of shipping companies. This situation creates a behavioural baseline, from which anomalous behaviour may be inferred. This task reflects the main subject-matter shown in this work.

The definition of anomalous vessel behaviour is of paramount importance and it is given in Section 3.1. Regardless of how such anomalies are construed, a framework capable of dealing with both the detection and identification of anomalous behaviour may be designed. A brief review of the various Anomalous Detection (AD) Frameworks found in the literature, alongside their scope variants, is presented in Chapter 2. Such methods are tailored to different requirements, which are not always synchronised with the ones aimed for the particular needs of this work.

The work that is developed throughout this dissertation is integrated within an ongoing highly-collaborative European project, the MARISA project [2]. Maritime

[2]https://marisaproject.eu

Integrated Surveillance Awareness MARISA is a project funded by European commission under a Horizon 2020 research and innovation programme. The mission of the MARISA project is to enhance the decision making and reaction capabilities of the maritime authorities, by the development of a toolkit. This is achieved within 22 entities working collaboratively towards the current real-world demands of the maritime authorities, which ultimately are the end-user of the toolkit. Such demands were initially presented for the project by the maritime experts and grouped into two major set of activities. The first group of activities, representing also the first stage of the project, focus on the usage of state-of-the-art technologies towards the collaborative development of the MARISA toolkit. The second set of activities is related to the validation of the toolkit capabilities by the execution of trials across different end-user sites. Through the process of meeting with the MARISA end-users, the focus of our current work was defined. The objectives of this present dissertatin were then focused on the first set of the project activities, with a higher emphasis on usage of novel techniques and algorithms to collect and properly process large amounts of heterogeneous data sets for early warning, forensic purposes and illegal act prosecution. Thus, for the sole-purpose of this work and given the context of the MARISA project, a set of specific objectives were defined, and are presented under in Section 1.1.

1.1 Objectives

Taking into account the INOV tasks, for the specific objectives of this dissertation, we are concerned with the task:

of developing a framework to take vast amounts of unstructured vessel data and, upon appropriate meaningful data structuring to be capable of recreating a vessel trajectory storing in a database as well as analysing information that ultimately allows for the detection of what is defined to be an anomaly

The formalised objective is admittedly general and entails many technically distinct challenges, both conceptual and practical. To tackle such difficulties we

subdivided the general objective into smaller objectives, thus breaking down a problem into smaller sub-problems:

- Ingest, pre-process and structure high-throughputs of maritime data.
- Provide procedures to transform spatial vessel data, into sequential data, thus defining vessel trajectory.
- Develop anomaly detection methods, based on the what is to be defined vessel anomalous behaviour, by the competent entities.
- Containerise the solution for the previous objectives into a framework which can be used into different maritime scenarios.

1.2 Outline

Following the introduction, the remainder of this dissertation is organised in six Chapters. A literature review, were questions regarding the maritime domain safety, and used technologies were explored. In the same Chapter we study the previous behaviour analysis frameworks presented in the literature. Methods for trajectory representation, regarding vessel trajectories are also discussed. Following this Chapter, in Chapter 3, we define what is to be considered an *Behavioural Anomaly* for this thesis. Following, we introduce the developed Modular Anomaly Detection Framework (MAD-F), describing the purpose of each module, and the considered data types. Chapter 4 we firstly provide an vessel dataset analysis, which was our initial contact with such domain specific data-types. Further, we explain the development and decisions took trough each modules of the *MAD-F*. Chapter 5 presents the results which were obtained per experiment. Finally Chapter 6, discusses the limitations of the presented work and presents recommendations for future research.

Chapter 2

Literature Review

Objectives for the MARISA project are well defined. In order to achieve the proposed goals and in preparation for this dissertation a vast number of subjects were investigated. An investigation in the following theoretical topics : behaviour analysis, anomaly detection and maritime safety technologies, were the major keywords for this literature review. In Section 2.2.1 an analysis of the principal similar Frameworks found in the Literature, will be presented. A brief introduction to the maritime domain, regarding the Maritime Safety affairs is presented in Section 2.1. In subsection 2.1.1, a description of the AIS technology and its use in the Maritime domain is presented.

2.1 Maritime Safety

Shipping is most likely, the most international task of all Worlds Industries, because of this international nature. It has long been recognised that improving maritime safety, is more effective if it is carried out on a international level, than by individual countries acting unilaterally without any co-ordination, [4].

The UN (United Nations) in 1948, established the International Maritime Organisation (IMO), as the first and principal international organisation devoted to maritime matters.

Since it's creation, the IMO has promoted the adoption of 50 conventions and protocols. The IMO has adopted more than 1,000 codes and recommendations regarding the maritime safety and security. The IMO objectives are easily summarised into their slogan : safe, secure, and efficient shipping on clean oceans.

2.1.1 Automatic identification system (AIS)

While the maritime safety domain is a vast and complex field for this investigation, it is important to focus on the technologies that the maritime domain has presented.

Automatic Identification System (AIS) is used to identify and locate Vessels by electronically exchanging data over high frequency VHF radio bandwidth to, other nearby ships and Vessel Traffic Services (VTS) stations.

The main motivation for the adoption of the AIS was its autonomous ability to identify other Vessels assisting humans with the collision avoidance. It has the ability to detect other equipped Vessel in situations where the radar detection is limited such as around bends, behind hills, and in conditions of restricted visibility by fog, rain, etc [5].

In 2000, the IMO adopted a new requirement for all ships, to carry an automatic identification system (AIS) that automatically provides the Vessel information to coastal authorities and other Vessels.

This regulation was initially imposed for all international ships with 300 gross tonnage or more and for ships with 500 gross tonnage and upwards navigating not international voyages. After 31 of March 2014 all EU fishing Vessels above 15m, are obliged by the European Commission to install an AIS. The ships information sent over the AIS[1] is classified into three main categories, they are presented in Table 2.1.

[1]http://ec.europa.eu/fisheries/cfp/control/technologies_en

Table 2.1: AIS Information Description

Category	**Description**
Static Information	MMSI - Maritime Mobile Service Identity
	IMO number
	Call sign and name
	Type of ship
	Length and beam
	GPS Antenna location
Sailing Related Information	Draught of ship
	Cargo information
	Destination
	ETA - Estimated Time of Arrival
Dynamic Information	Position of the ship
	UTC - Coordinated Universal Time
	COG - Course Over Ground
	SOG - Speed Over Ground
	Heading
	Navigational Status
	Rate of turn

Each Vessel transmits specific information related to the Vessel itself, the MMSI represents a 9 digit unique ID number, that every Vessel is assign with. Most of the information sent over AIS, is automatically generated by the ships sensors such as the GPS and the compass. Thus minimising the possibility of manipulate this data, although there is still information that is manually inserted by the crew such as the Navigational Status and the Heading.

Ships fitted with AIS are obliged to maintain the AIS in operation at all times. The AIS autonomously broadcast information, every certain time interval, therefore ships ping their AIS information every time interval There are international agreements, that protect the navigational information.

2.2 Behaviour Analysis

Behaviour Analysis, is a vastly researched topic that involves many research fields. A vast number of Frameworks with the main objective of Maritime Behaviour Analysis are proposed in the literature, some of these frameworks are presented in Section 2.2.1.

For this work Vessel behaviour is as considered as a baseline in which abnormal behaviour can be found. This baseline occurs as normal trajectories are various and constant, producing a normalcy model of Vessels dynamics in which Machine Learning Techniques can learn. Anomalies don't necessarily mean that there is something abnormal with the ship Vessel behaviour. That is something hard to imply with only AIS data. Anomalies in the AIS data can represent numerous abnormal events. Some of them that can be illegal, that's why further investigation from maritime authorities is needed.

2.2.1 Similar Frameworks

There are a vast number of frameworks in which Vessel behaviour will be analyse. This will be done with the purpose of anomaly detection which are fully defined as integrated systems. The authors in [6] suggested the framework MT-MAD (Maritime Trajectory Modelling and Anomaly Detection), in which a given set of moving objects, the most frequent movement behaviour are explored, evaluating a level of suspicion hence detecting anomalous behaviour.

The authors in [7], introduced the framework TREAD (Traffic Route Extraction and Anomaly Detection). The framework is proposed in which an Unsupervised Route Extraction is used to create a statistical model of maritime traffic from AIS messages, in order to detect low-likelihood behaviours and predict Vessels future positions.

A framework for Vessel behaviour analysis focusing on Vessel interaction or rendezvous. The proposed framework, is divided into the following three logical

connected phases: Engagement Detection, Scenario Detection and Anomaly Detection. The use of the 3-phase framework serves as a filter to reduce the volume of data that is processed by the sub-sequential phase. Therefore prioritising critical scenarios, that request human intervention [8].

Although accessing the performance of the frameworks, is an ardours task. There is no defined benchmark set where tests can be performed, with labelled samples described as positives or negatives of what are considered anomalies at seas [9].

In [2], a detailed solution for constructing an AIS database, with the potential value for being used as benchmark database for maritime trajectory learning, and efficiency testing of data mining algorithms.

A partition-and-detect trajectory in which trajectories are partitioned into a two-level of granularity achieving high efficiency and high quality trajectory partitions, therefore detecting outlier trajectories using density-based methods [3].

There are numerous studies that show how, Vessels tend to alter their routes in order to achieve safe distances when passing near other Vessel. In [10] a detailed study on Merchant Vessels AIS data, presents how this type Vessels alter their route, when new surface offshore petroleum installations are constructed.

2.3 Trajectories Analysis

Trajectories analysis is a researched field for numerous years. It is researched in areas where moving objects, this objects can be Humans, vehicles, animals, or even natural events such as hurricanes or storms. A survey of trajectory data analysis applications, is presented in [11].

As the volume of positional AIS data exponentially increasing, it is important to find methods in witch raw trajectories data can produce value. This methods that learn with trajectory data can greatly impact the Maritime domain.

Trajectory learning is the process of learning motion-patterns from trajectory data using unsupervised techniques, mainly clustering algorithms [12]. Morris and Trivedi [13], further categorise trajectory learning as a three-step procedure:

1. Trajectory Pre-Processing.
2. Trajectory Clustering.
3. Path Modelling.

In the Maritime domain, as Vessels are free to navigate in open waters, this fact produces a specific level of uncertainty related to Vessel trajectories, there are no standards for Vessel trajectory representation.

A way to discretize a trajectory discovering frequent regions is presented in [6]. Representing the trajectories in a spatial grid in which a cell represents a geographical area with a defined size.

Pallotta, proposed a method that enriches the raw Vessels tracks with a description of the ship movements. This is the raw trajectories are labelled with the Vessel movement type information as 'Stationary' or 'Sailing' [7].

The authors in [3], raw trajectories are partitioned into sub-trajectories, creating a new insight for data analysis, adding the possibility of focused region analysis.

A framework for scene modelling using trajectory dynamics analysis, for the discovery of POIs(Point of Interest) and the learning of AP(Activity Path), [13]. These last representation is quite important for the Maritime domain, as the discovery of new POIs, can indicate the common Vessel destinations (e.g. frequent fishing zones, ports, etc.).

2.4 Time Series

The concept of time series is related to trajectories, as a time series is a set of ordered observations on a a quantitative characteristic of a phenomenon at spaced time period, [14]. Formally, a uni-variate time series xj, is defined as a sequence of real numbers, where n is the length of the series, represented as:

$$xj = \{x(i) \in \mathbb{R} : i = 1, 2, 3, ..., n\}$$

There are numerous applications for time series analysis, one of the main applications, is the use past time series, in order to forecast future values. These applications are used in numerous areas such as economics, engineering and others.

2.4.1 Multivariate Time Series

The AIS data cannot be described as a uni-variate times series, as it is composed by various variables. Therefore AIS data needs to be analysed as a Multivariate Time Series (MTS). For each AIS message, the features can be extracted with the time-stamps that the message was broadcast. A detailed description of the AIS features is found in Section 2.1.1.

A possible representation of a Multivariate Time Series, X is:

$$X = (x_1, x_2, x_3, \cdots, x_m)$$

Where each xj is defined in section 2.4.

$$Xj = \{Xj(i) \in \mathbb{R} : i = 1, 2, 3, ..., n\}\,(j = 1, 2, 3)$$

The analysis and classification of MTS is a arduous task for traditional machine learning algorithms, mainly because these algorithms do not handle well dozens of

variables, [15]. Representing MTS into multiple univariate time series, can create losses in the correlation of these variables, as variables are being processed them independently.

2.4.2 Time Series Clustering

Temporal data mining research, a big emphasis lies on the clustering, and posterior classification of time series data. Time Series Clustering is used to identify in datasets, homogeneous groups where same group object similarity is maximised, and the minimised when not in same group.

The authors in [16], summarise previous work that investigates the clustering of time series applications in various fields, and propose an extensive survey.

The same authors, define a necessity to clustering, when working with unlabelled data. This data can come from various sources including : categorical, numerical, images, spatial, etc.

The main source of data for this work is AIS data, which is a unlabelled multivariate data source. Labelled AIS datasets for anomaly detection are either really expensive, or just not available for the public domain.

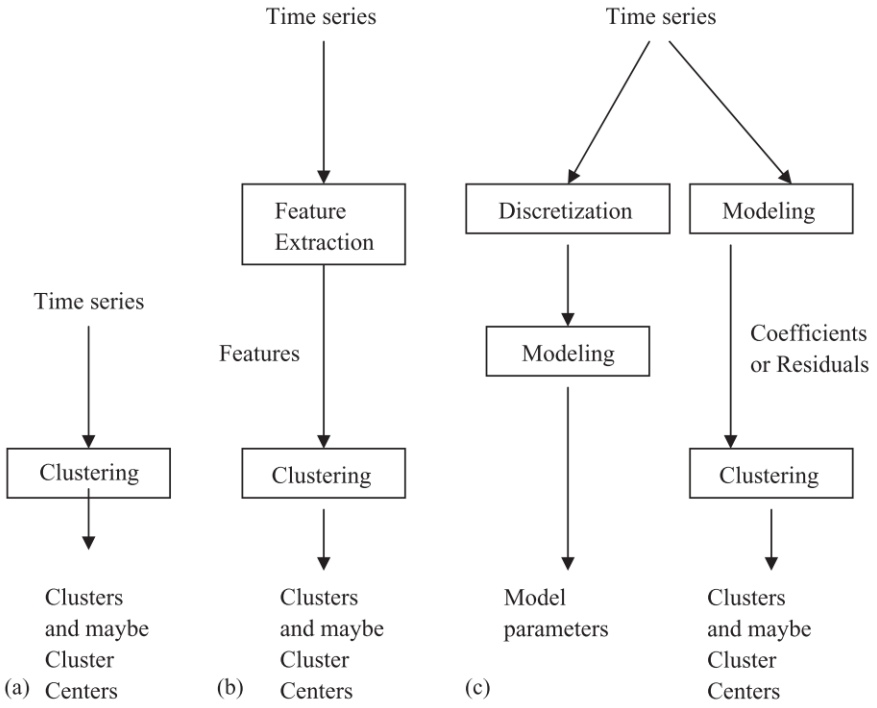

FIGURE 2.1: Three types of time series clustering defined in, [16]

Time Series Clustering can be categorised into three main general approaches, simply described in Figure 2.1, these categories being:

- **Raw-data-based approaches** These approaches work with raw sets of data, normally in the time domain.
- **Route Definition** Several methods of Vessel route definition, are presented in the Literature. Although, at this moment a method was chosen that represents the route as a whole. Therefore no information is lost, a detailed description of the latter is presented in Section 2.3.
- **Model-based approaches** This is a more complex clustering technique, in which, each Time-Series is considered as a statistical model or as a mixture of statistical distributions, thus two time series are considered similar when the models that fit this distributions are similar.

2.4.3 Time Series Classification

Time series classification, is used for numerous purposes, from, the main difference when classifying or clustering Time Series lays in the fact that, classification can occur when a predefined set of classes already exist and the main objective is to classify this data in the different classes, thus in machine learning being considered a Supervised Learning task.

Early work, from 1998, the authors propose p-value hypothesis test, performed for every pair of stationary multivariate time series, [17].

Three main categories of sequence time series classification, are defined by the authors in [18]:

Feature Based Classification A sequence of features is transformed into a feature vector, then convectional classification methods are applied. Feature selection represents is an important task for this method of classification.

Distance Based Classification The distance function that measures the similarity between the time series, induce the quality of the classification overall. A more detailed research on these distances is presented in [19].

Model Based Classification Where models, such as multivariate Gaussian mixture model (GMM) [9], Support Vector Machines (SVM) or Hidden Markov Models (HMM) and other statistical models are used to classify time series.

2.5 Distances Measures

In order to compare classify a time series using distances, the concept of distance, and type of distance must be defined.

A distance is defined as a numerical measurement, that measures how far two objects are from each other. There are a vast number of distances used in computer algorithms. The most commonly used distance measure is the Euclidean

distance, this measurement is a metric distance function, since it obeys to the three fundamentals metric properties: non-negativity, symmetry and triangle inequality [20].

The similarity between two time series, can be calculated by simply summing the ordered point-to-point squared distance between both time series, this is shown in Figure 2.2.

Although, euclidean distance between two time series can only be calculated if, both time series are of equal length, [21]. If two time series are identical, but one is shifted slightly along the time axis, using the Euclidean distance, it may consider the time series very different from each other, [22].

This creates a problem when analysing certain type of time series, as both may not have them same length, or might just be time-shifted, which happens when analysing AIS data. In the literature a few solutions are presented, one of them is using another distance measure.

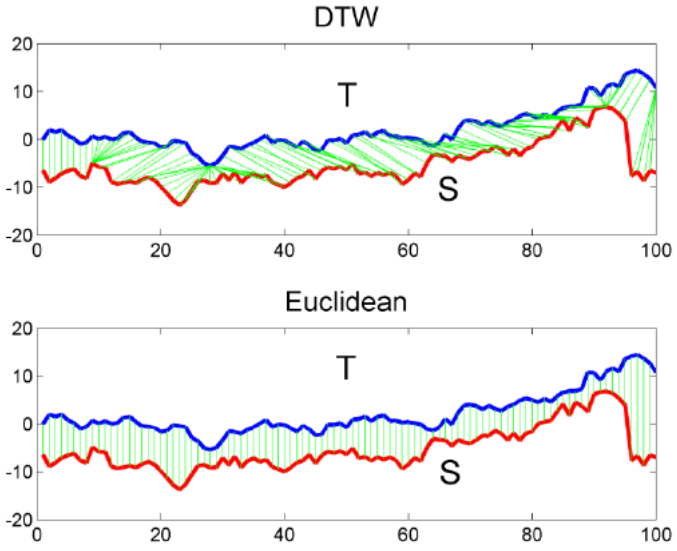

FIGURE 2.2: Difference between DTW distance and Euclidean distance (green lines represent mapping between points of time series T and S), [21]

2.5.1 Dynamic Time Warping (DTW)

As distance measures play an important role for similarity problem, in data mining tasks, Dynamic Time Warping (DTW) is a algorithm that computes the optimal alignment and distance between two time series, [23]. One time series may be "warped" non-linearly by stretching or shrinking it along its time axis. Although computing the DTW between two time series, is quite computationally expensive, as its quadratic time complexity may hamper its use to only small time series, [22].

Chapter 3

Modular Anomaly Detection Framework

In this Chapter, we present the overall description of steps towards the development of the *Modular Anomaly Detection Framework* which is be used throughout this dissertation. A crucial component in this work relied on a technically accurate definition of a *maritime anomaly*. This is generally speaking a challenging task since a data-driven definition is currently lacking or insufficient. A more meaningful solution to this problem was provided by the aid of maritime experts who were engaged in the MARISA project. In particular, members of the Portuguese Navy interacted with us in order to offer the required their exclusive technical insight.

Given their specific input and real-world knowledge of the maritime domain, one can arrive at a well-defined concept of anomaly that can be translated into a precise notion to be used in this Framework. Before we engage in the specific requirements that served as a blueprint for the developed framework, such a definition will be given. This will then be followed by the technical description of such requirements, namely by distinguishing *anomaly* and *data requirements*.

Lastly, a general overview of the proposed Modular Anomaly Detection Framework, which will be referred to as MAD-F from now on, is presented. This is done

in light of Figure 3.1, whose modules are explained individually throughout the following Subsections 3.2.1, 3.2.2, 3.2.3, 3.2.4, 3.2.5 and 3.2.6.

3.1 Anomalies within the MAD-F

An anomaly may have numerous interpretations depending on the context in which it is found. However, it can be generally conceptualised as a subset of data that stands out in some preconceived way when contrasted to the overall dataset. Nowadays, the anomaly detection of vessel behaviour is solely done by human maritime experts. This procedure depends on national security agencies. Within their duties, these agencies are responsible for assuring the coastal surveillance of their territory by assessing possible threats and identifying abnormal behaviour. The current methods employed by these institutions are neither efficient nor scalable and therefore not suitable for the challenges brought by the exponential growth of vessels at seas. This state of affairs creates an ideal situation for the use of data-driven models to assist the maritime experts.

The notion of anomaly just presented is unsatisfactory given both the complexity and purpose of the problem. For the goals of this project, such a technical definition is tailored specifically by the maritime agencies involved in the MARISA project and we therefore refrain from applying our own definitions, which usually stem from abstract statistical data-driven notions.

By having meetings with maritime experts a list of the anomaly requirements was agreed. For this work this list served as not only the concrete anomaly requirements, but also as a guide for the overall implementation of the *MAD-F*. The list of anomaly requirements in shown under in Table 3.1.

TABLE 3.1: MAD-F anomaly requirements, which were defined by maritime officers.

Anomaly Requirement	Provided Description
AR1	Detect Abnormal changes of (more than a configurable value) Direction.
AR2	Detect Abnormal changes of (more than a configurable value) Velocity.
AR3	Detect Vessels disappearance from sensor coverage for more than a configurable Time Period.
AR4	Detect when the observed Vessel Navigational Status is not consistent with the reported Vessel Kinematic features.
AR5	Detect when Vessels report a geographical and time incompatibility.
AR6	Detect when two or more Vessels are approaching close to each other.

As mention previously, requirements for this work were distinguished from anomaly requirements and data requirements. The latter was intrinsic for this work, as the uncertainty of data types and sources when dealing with the maritime field is immense. The problem of having numerous types and sources of data is still aggravated as the maritime domain is also capable to produce enormous workflows of data. Thus, a specific data requirement for this work could be simply specified as:

The developed Framework, must be able to ingest fuse and store different sources of maritime data, while also handling enormous workflows of data in real-time.

3.2 Modular Vessel Anomaly Detection Framework

In order to develop a Framework capable of achieving the requirements defined above in Section 3.1, we propose the Modular Vessel Anomaly Detection Framework. The *MAD-F* is able to ingest data from different feeds of data in real-time while simultaneously constructing a data-base for vessels trajectories in a unsupervised manner. Anomalies are then detected in a offline manner from the saved trajectory data, or online (in real time) addressing the incoming streams of vessel data. The framework was developed to be modular as there are either no inputs or outputs standards for the maritime domains. Thus, by developing a configurable and not static framework, we provide the *MAD-F* with the necessary configuration flexibility, allowing it to be configured for different scenarios or even by different national maritime authorities; or even allowing new *Framework Modules* to be easily integrated in the future.

Thus, by providing a configurable and not static framework, we give the configuration flexibility for the being configured for different scenarios or even different National Maritime Authorities, or even to new Modules being added in the future.

In Figure 3.1 we present the architecture of the *MAD-F* and the following subsections will discuss each of the framework modules: Data Ingestion, Data Pre-processing, Feature Engineering, Trajectory Extraction and both anomaly detection modules : Anomaly Detection Service and Rule Based Anomaly Detection Service.

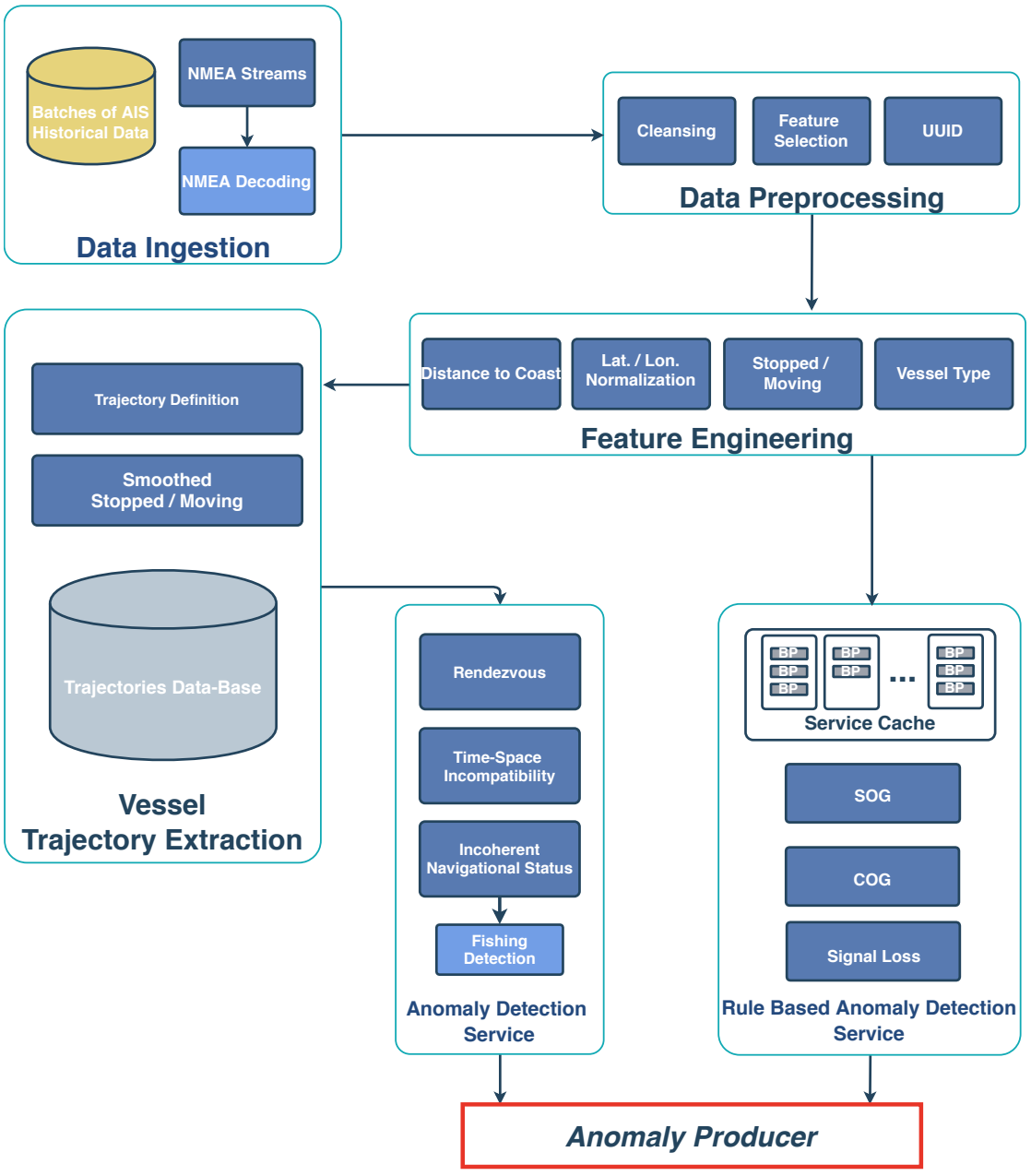

FIGURE 3.1: Proposed architecture for the *MAD-F* Framework

3.2.1 Data Ingestion

Data Ingestion Module, represents the data input for the developed Framework. AIS data was the most representative data type used for this work, as it showcases the actual instantaneous Vessel information. Although, the used AIS data for this work came in two really distinct formats. It either came in *Historical Batches* representing historical sets of data, or real *NMEA AIS Streams* which represent

real, real-time data. For both data formats, the framework is scalable, and able to ingest one or multiple feeds / sources of data simultaneously.

Via the MARISA project, we accessed AIS live feeds from antennas all around the Portuguese coast line. This antennas receive vessels transmissions via AIS up to 20 Nautical Miles of the shore (depending on the weather conditions), and have reception rates up to 30 Messages per minute per vessel. The real live feeds of AIS data, are received via TCP in the NMEA format.

National Marine Electronics Association (NMEA) is a standard communication protocol used by Maritime Sensors such as Accelerometer, Giroscope, GPS receivers, etc. NMEA encapsulates the information from the different Vessel sensors, and broadcasts this information to coastline antennas and nearby Vessels via AIS protocol.

```
b'!AIVDM,1,1,,A,13Ujr<001fOIMLtCOMMQa1@400RV,0*0C\r\n'
b'!AIVDM,1,1,,A,13EjsP?P0MwM<CtE1H<d0?v605hd,0*29\r\n'
b'!AIVDM,1,1,,B,13s=LQgP00wH2BrG@;h@0?v42@1r,0*16\r\n'
b'!AIVDM,1,1,,B,402PeN1v9qf;3OGu@dG@Jf100@1s,0*48\r\n'
b'!AIVDM,1,1,,A,13ebPb0023wKODPCmfq<l:D40@1t,0*61\r\n'
b'!AIVDM,1,1,,A,102CuBhP00OLuCtC>nI00?v40Uhd,0*56\r\n'
b'!AIVDM,1,1,,A,13VjPP?P00wQPEFCVf`12Ov82<2C,0*6C\r\n'
b'!AIVDM,1,1,,A,4028jLQv9qf;3O`:B6Dp8I?00D:9,0*53\r\n'
```

FIGURE 3.2: Snapshot of raw AIS data in NMEA format.

Although the use of real AIS data comes with many challenges, as it is mandatory to decode, sort and store the received data, thus allowing the incoming data to be used as viable source of data. Secondly, as AIS-receiving stations receive the broadcast AIS information from multiple AIS-equipped vessels simultaneously, and the reception range of each AIS-receiving can vary depending on the actual weather conditions and the location of where such station is located. This originates two main problems :

1. Duplication of reception: With the variation of reception ranges from the different AIS-receiving stations, this creates the problem of multiple stations receiving the same vessel broadcast. The duplication of messages is a problem which occurs when handling real NMEA streams, the methods used to solve such problem are presented in Section 4.2.

2. Non-reception of broadcast: Similar to the problem presented above the non-reception of by any receiving station can also occur. To address this problem, maritime agencies use satellite AIS (S-AIS). S-AIS solves the problems related to the reception range, but presents another problem with refreshment rates, has the reception of the broadcast are dependent of satellite revisit time [24].

3.2.2 Data Pre-processing

The Data Pre-processing module, is the first step of Data Wrangling in our Framework. The motive for this module is to select, transform, and clean the received data, from the Data Ingestion Module. As described in Section 2.1.1, AIS presents a large amount of different features, which can be used for different problems. Feature Selection represents a important step for this work, as the selection of the "relevant" features directly influences the overall performance of the *MAD-F*, and also the expected results from the anomaly detection task. Such representative task requires pre-conceived knowledge of Vessels dynamics and behaviour, which is only gained with experience in the Maritime Domain. For this work the feature selection was done based on the literature, and also by accessing Maritime Expert Knowledge via the MARISA project.

During the *Pre-processing*, a data-cleaning process is conducted, discarding corrupted data. This is done based on the information that standardises the AIS features, which is further detailed in Section 4.1.

Most importantly, in this module the concept of *Behavioural Point* is defined. *Behavioural Point* which will be referred as BP from now on, for this work represents our normalised representation of the previously selected features. A detailed explanation of this concept is provided in Subsection 4.3.3.

3.2.3 Feature Engineering

Feature Engineering, represents the second step of Data Wrangling in our Framework. During this step, the already pre-defined BPs, in the Data Pre-processing module, are enriched by extrapolating additional features.

Firstly for each BP received by this module, if the Vessel Type is not received in the AIS message, the Vessel Type is either extracted from external vessel static information sources, or it is scrapped from this internet. Secondly, each BP is enriched with by calculating the closest country and respective distance to shore. The same is done to ports, by calculating the distance to the closest port. Also, in this module with the reported kinematic features, the instantaneous move state of the vessels is inferred. Such procedures are further individually explained throughout Subsections 4.4.1, 4.4.2, 4.4.3.

3.2.4 Vessel Trajectory Extraction

Vessel Trajectory Extraction module, handles the definition, storage, updating and inserting of new incoming BPs into defined *Trajectories*. When considering trajectories, the BPs stop being valued as single points in time, and the aggregation of BPs via the vessel identifier throughout time, start representing a vessel trajectory. This allows a more conclusive vessel behaviour analysis based on its past trajectory. Although, in order to analyse trajectories, such concept needs to be defined and represented in a optimal manner. Furthermore, when dealing with real maritime data (and specially when working with real Maritime Authorities) it is extremely important to trace-back/log the data, thus when an a anomaly is generated, knowing which BPs generated which anomalies is possible. In Section 4.5.1 our definition of a vessel trajectory is presented.

3.2.5 Anomaly Detection Service

ADS (Anomaly Detection Service) Module, represents for our Framework the historical, offline anomaly detection module. *ADS* module works offline in effective time, on batches of historical Trajectory Data served from *Trajectory Data-Base* from the *Vessel Trajectory Extraction* module. Access to Trajectory Data, is done by querying the *Trajectory data-base* with a configurable set of parameters, which can be time restrictive(such as the 10 past Hours) and or from a vessel specific set of vessels.

Received trajectory data, is then used to detect: **Time Space Incompatibility**, **Vessels Rendezvous**, and **Incoherent use AIS Navigational Status**. For the latter, we create a sub-method for the which serves as the validation of *Engaged at Fishing Navigational Status* based on vessels types and reported kinematic features. The implemented methodology for the detection of each anomaly is represented in Subsections 4.6.1, 4.6.2, 4.6.3 and 4.6.4 respectively.

3.2.6 Rule Based - Anomaly Detection Service

RB-ADS (Rule Based - Anomaly Detection Service), opposed to the ADS module described above, corresponds to the online, in stream processing anomaly detection module. *RB-ADS* modules works online in near real-time, accessing the stream of already pre-processed BPs from the Feature Engineering Module. In order to the *RB-ADS* be able to perform Anomaly Detection in near real-time, a Queuing Systems for this module was defined. This queue which we named *Service Cache* is further detailed in Section 4.7. The arriving stream of BPs, is are stored in individual Vessel Queues of size N. The individual Queues are then accessed, allowing a real-time calculation of the set of Anomaly which can be defined by Rules. The anomalies validated online trough rules for this work are: **Abnormal change of Velocity**(**AR1**) , the **Abnormal change of Direction**(**AR2**), and the **Vessel Signal Loss**(**AR3**), our approach towards the detection of such anomalies is described in Section 4.7.

Chapter 4

MAD-F Development

In this Chapter, we present the development steps towards the implementation of the *Modular Anomaly Detection Framework* MAD-F. Firstly, we present a list of the technologies used in this work. Then we undertake a initial data analysis from a historical AIS dataset. And finally, the implementation of each MAD-F module is individually explained, providing a detailed clarification of the undertaken approaches.

In order not to develop a fully static framework, a modular development was applied instead. This allows specific modules of the framework to be instanced multiple times with different configuration; or even the possibility of having the new modules added to the framework in the future.

For this end the choice of technologies was done by by emphasising efficiency handling large quantities of data and scalability. Implementation of this Framework was done with the programming language Python, using different specific packages for the different specific tasks. The used packages and their usage will be explained throughout this Chapter. Architecturally wise, the framework was implementing following an somehow layered architecture, similar to the *Lambda Architecture*, which was firstly introduced by the authors in [25]. As so, the chosen data-base for this framework was Apache Cassandra [1], which was essential

[1] http://cassandra.apache.org/

to store the aggregated BPs in a fast and effective way. The detailed usage of the data-base is explained in 4.5. The reception of BPs by the *Trajectory Extraction* module was done using a message queue system Apache Kafka [2]. The same message queuing approach as also implemented for the modules who needed to send and consume messages between them. A detailed explanation of such implementation is provided in the following Sections.

4.1 Data Analysis

In order to gain insight and find the limitations of the AIS data, our initial step towards the implementation of the framework was a the analysis of an historical AIS dataset. The analysed dataset was compiled, and made publicly available by another H2020 European Project[3]. This dataset was chosen, due to the completeness of documentation and description of the actual dataset; which to the extend of our knowledge was the only open-source AIS dataset with such characteristics.

In this Section, we present a data analysis from the dataset [26]. We conducted this data analysis, by firstly providing a general description of the used dataset, and secondly by analysing the overall feature distribution of the each feature in the used dataset. The used dataset, is composed from **18,684,115** AIS messages originated by **4,555** different vessels. The Data-Set covers a period of 6 Months (from 2015-10-01 to 2016-03-31), from a area nearby Brest, France as it is presented under in Figure 4.1.

[2]https://kafka.apache.org/
[3]http://datacron-project.eu

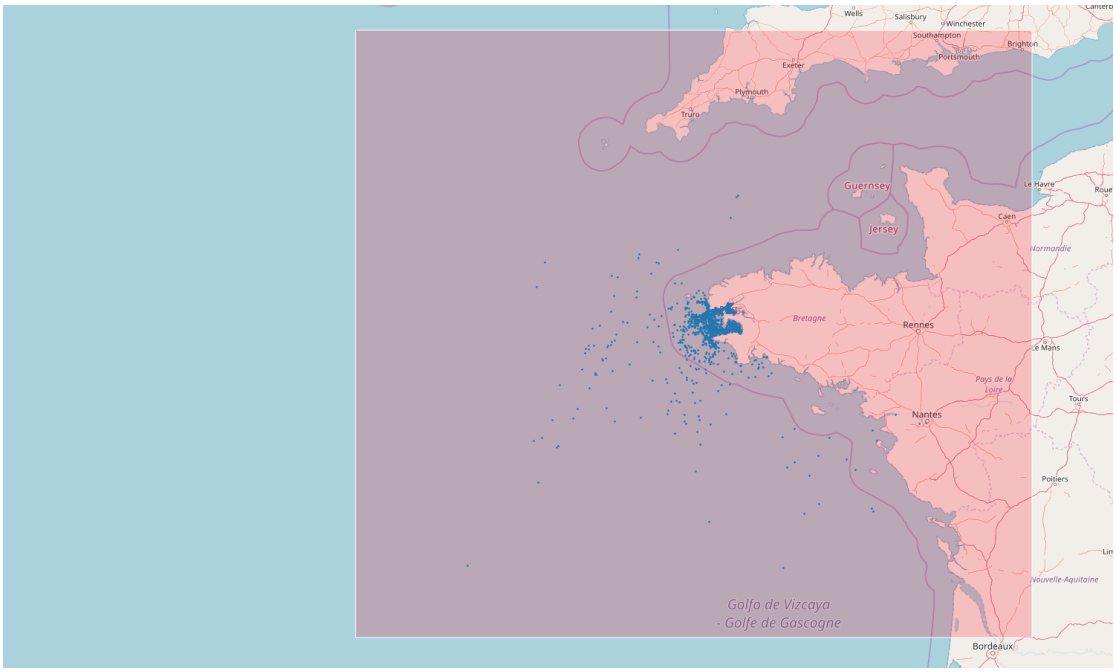

FIGURE 4.1: Area of the dataset represented in the Red, with a sample of 50,000 AIS Positions.

Every AIS message provided in the dataset, is composed by the features that derive from the AIS dynamic information. In Table 4.1, we describe the dataset features by detailing their units and their unit range.

TABLE 4.1: AIS dynamic messages features description.

Feature	Description	Unit	Range
MMSI	Vessel Unique Identifier.		0 to 99999
Status	AIS Navigational Status.		0 to 15
Turn	Rate of turn, right or left.	degrees per minute	0 to 720
SOG	Speed Over Ground.	knots	0 to 111*
COG	Course Over Ground.	degrees	0º to 360º
X	Longitude.	degrees	-180º to +180º
Y	Latitude.	degrees	-90º to 90º
Time	Received Timestamp.	Unix Time	

The dataset not only contains the AIS dynamic information, but also in separate files the related vessel static information of each vessel which has reported

in dataset. By interpolating the MMSI reported in every AIS message, we were able to enrich each AIS dynamic message (or row of the dataset), with the static information related to the vessel which has produced the dynamic message. The vessel's static information contain information of the vessel's actual dimensions and type. The use of information related to the vessels characteristics is used in different types of behavioural analysis. In this work we used the vessel type as an key aggregation indicator, which we better described in Section 4.4.1.

4.2 Data Ingestion

Data Ingestion refers to the model, where the data is input into the Framework. As mentioned in Chapter 3, for this work was assumed that the incoming data would be able to come in two different typologies, either from batches of AIS data or Live NMEA streams.

Historical batches of AIS data (or datasets), are uploaded to this module via .csv files, which then are transformed into DataFrames using the Pandas[4]. For each imported batch of data, the features names must be pointed to the format we present in Section 4.3.3. Although for the NMEA Streams the as the decoding of such streams was needed. The choice of methods to process and decode was not has trivial. As NMEA messages are received in high frequencies, the method to such streams into comprehensible AIS like data, needed to be stable and efficient. In order to achieve this, we used the python library libais[5], which is implemented in the programming language C++, allowing a really efficient decoding of the incoming NMEA messages.

In Chapter 3, we have identified two problems that occur when working with AIS real live. In order to mitigate the duplicated reception, each received messages is tagged with a unique identifier (UUID). For this present work, the created UUID will be done by considering the ID of the vessel(MMSI), and the time the received

[4]https://pandas.pydata.org
[5]https://github.com/schwehr/libais

message was generated by the vessel. Thus, if two same UUID messages are received, we the second messages is discarded, and only the first received messages is considered.

The Framework was developed to be scalable, being able to handle different sources of AIS data, although for the purpose of this work, we limited the used data to two main sources of Data. The Data-Set presented in Section 4.1, and the NMEA feeds made available by the Portuguese Navy.

4.3 Data Pre-processing

Data Pre-processing, represents the module that handles the raw/unprocessed AIS data. This module cleans, transforms and normalises every AIS messages, coming from the Data Ingestion Module. Every AIS message is transformed into our normalised representation of and AIS message, which we defined as a **Behavioural Point**, defined under in Subsection 4.3.3.

4.3.1 Latitude Longitude Normalisation

In order to normalise the reported vessels positions, either from the AIS streams or the used dataset, we defined a set number of decimal cases used. This is done as most of AIS providers only assure a GPS precision of 0.0001 minutes accuracy, but what we found was that some reported positions come with up to 8 decimal cases, which can be caused just from how the dataset files were written. So our normalisation process, we ensured that every vessel position was normalised to a precision on 4 decimal cases. As this represents a global precision error of 11m to 4m, which is shown in Table 4.2.

TABLE 4.2: Degree precision versus the approximate radius of measured error.

Decimal Places	Degrees	Precision Equator	Precision 45º N/S	Precision 67º N/S
0	1.0	111.3Km	78.7Km	43.5Km
1	0.1	11.3Km	7.8Km	4.4Km
2	0.01	1.13Km	787.1m	435m
3	0.001	111.3m	78.7m	43.5m
4	0.0001	11.3m	7.8m	4.4m
5	0.00001	1.3m	0.7m	0.4m

4.3.2 Data Cleansing

Data Cleaning refers to the process of cleaning the data which is wrongly defined or, has wrong types. When handling with sensor generated data is common that wrong sensor reading can occur. In AIS data, this errors tend to occur as AIS features that are not transmitted at all, or that are transmitted with values that don't correspond to the Feature value range. An example of this is having a Latitude being broadcast with values of 500º. Therefore, we discarded all AIS messages with reported features that were not inside the feature value range. The features value range considered for the proposed framework was the one presented in Section 4.2, Table 4.1, which is similar as the one presented by the authors in [27] as the AIS default feature range.

4.3.3 Behavioural Point

Behavioural Point for this work, is our normalised feature representation of incoming vessel data. A BP is a multidimensional point which is identified by the vessel id who produced the reported message. Therfore a BP_{MMSI} can be represented as:

$$BP_{MMSI} = [t, x, y, SoG, CoG, NS]$$

Where the dimensions of the multidimensional BP represents the features (Time, Longitude, Latitude, Speed Over Ground, Course Over Ground and Navigational Status) respectively. Each BP was correlated to one (one to one) identifier. The used identifier in this work was the Maritime Mobile Service Identity (MMSI). For this work the non replication of the MMSI by different vessel was assumed, this problem was discussed in Section3.2.1. Each BP, as described above is further enriched by extrapolating three additional features, making each BP to be represented as:

$$BP_{MMSI} = [t, x, y, SoG, CoG, NS, VT, DtS, DtP, PN]$$

Where the additional features *VT, DtS, DtP, PN* representing the vessel Type, Distance to Port, Distance to Shore, Port Name. These features are not reported from every AIS messages and need to be extrapolated afterwards. The methods used to extrapolated this features are presented under Section 4.4.

4.4 Feature Engineering

4.4.1 Vessel Type

Vessel Type, is a classification system, where each vessel is categorised by the type of activities it preforms. Classified by a numeric scale from 0 to 99. The first digit represents the general activity category of the vessel, and the combination of the first digit with the second represent the specific activity of the vessel. In Table 4.3 we list all the general vessel categories which are associated with the first digit of the vessel type feature, but also we present the specific vessel categories for the more frequent vessel types occurring on the dataset.

TABLE 4.3: Vessel Type categorisation and most frequent representation.

First Digit	General Category	Relevant Categories	
1	Reserved		
2	Wing In Ground		
3	Special Category	30 - Fishing	30 - 286(6%)
4	High-Speed Craft		
5	Special Category		
6	Passenger		
7	Cargo	70 - Cargo	70 - 1,511(33%) 79 - 273(6%) 71 - 217(5%)
8	Tanker	80 - Tanker	80 - 342(7%)
9	Other		99 - 1,192(26%)

For the used dataset described above in Section 4.1, the Static Vessel Information is available for all the vessel in the dataset. Although, when handling Real-Time NMEA streams or other Batches of Data, the Vessel Static information is not available or broadcast. This, creates a problem of not having the Vessel Type information which is used to query our Trajectory Data-Base. For this we developed a **Web Scrapping application**, described in the following subsection.

4.4.1.1 Vessel Type Scrapper

Web Scrapping is used to extract information from freely available websites. For the sole purpose of this work, we developed an application that would retrieve the Vessel Type information from a "well known vessel traffic webpage". By providing the vessel MMSI to the Vessel Type Scrapper, we retrieve the html webpage data that contains all the static vessel information available on the "well known vessel traffic webpage". From the html data we, striping the html tags, and the non

relevant specific webpage information, we access Vessel Type, as it is presented in Figure 4.2.

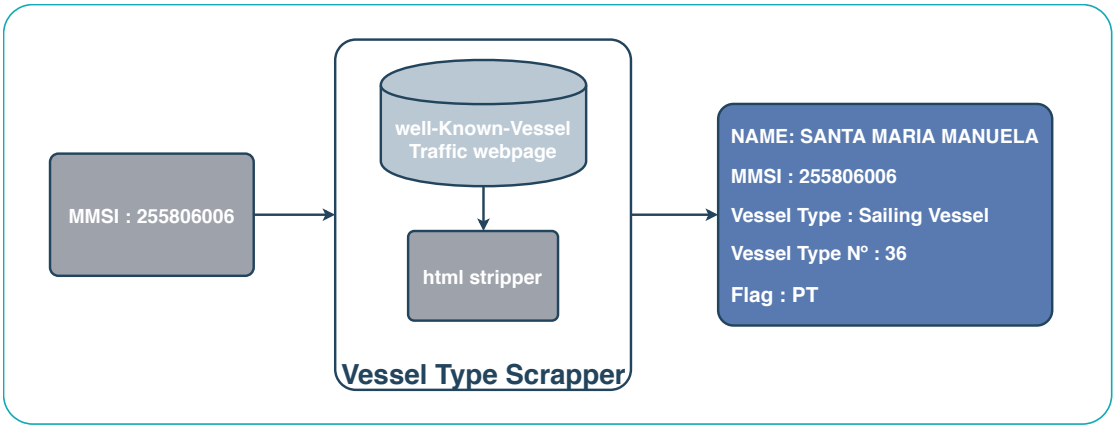

FIGURE 4.2: Example of the Vessel Type Scrapper retrieved information for Vessel MMSI: 255806006

4.4.2 Distance to Coast

Distance to Coast influences, the navigational behaviour for the major part of Vessel Types. In order to enrich the *BPs* which will feed the Anomaly Detection modules, and as the distance to shore is without a doubt a valuable aggregation feature for the maritime domain. We extrapolated the Distance to Shore for every received AIS message.

Although in order to calculate the distance to shore effectively either over historical batches of data or in real time to streams of AIS data, a efficient representation of the coastline is needed. For this we used the ocean coastline data[6]. This representation has mapped Global coastline in a vector of **547,503** points, which is equivalent having a 1:10m Global coastline representation.

The calculation of the closest point was done with a Nearest Neighbour approach, using the Ball Tree algorithm. The choice of this algorithm was done, due to the high volume of data we were using, and the possibility of using the Haversine Distance measures in the already implemented methods from [7].

[6]http://naturalearthdata.com
[7]http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.BallTree.html

Haversine is the most commonly used distance metric in the vessel navigation. As both Latitude(y) and Longitude(x) features are represented in a spherical coordinate system, the use of the most common Euclidean distance is not applicable. Thus we used the Haversine Equation 4.1, represented under.

$$d = 2rsin^{-1}\sqrt{sin^2(\frac{lat_{p_2} - lat_{p_1}}{2}) + cos(lat_{p_1})cos(lat_{p_2})sin^2(\frac{lon_{p_2} - lon_{p_1}}{2})} \quad (4.1)$$

Where d takes as input (p_1, p_2), and it calculates the haversine the 2 point represented as $p1(lat_1, lon_1)$ and $p2(lat_2, lon_2)$. r represents the approximate radius of the Earth which for this work we considered **6,367Km**.

4.4.2.1 Distance to Port

Distance to Port, to the maritime scenario, and more specifically maritime international trade, represents an additional feature which is of great importance. The Estimation of Time of Arrival presents itself as a necessity for container terminals, as this terminals base operational decisions on such estimation. The estimation of time of arrival, and the prediction of the arrival port based on past vessel trajectory information, are two tasks which use the distance to port feature for such purpose, [28, 29].

This being said, we enriched each *BehaviourPoint* by calculating the actual nearest port, and the distance to it. To achieve this, we used a similar approach as explained above in Subsection 4.4.2.

Although, getting a list of every port was not trivial, as there are numerous ports around the World, and such information is not centralised nor normalised. We accessed the detailed information of the World Port Indexes in [8]. The World Port Index data was in a GIS(Geographic Information System) shapefile format,

[8] http://msi.nga.mil/MSISiteContent/StaticFiles/NAV_PUBS/WPI

which is common format for the Maritime Domain, but not usable in our Framework. Therefore, we firstly normalised the data format using the Python package dbfread[9], and then stored the normalised port data in our data-base. For each of the **3,865** ports we extracted the respective Port position, Country, and Name. In Figure 4.3 we present the port position over the Iberian coast in Orange.

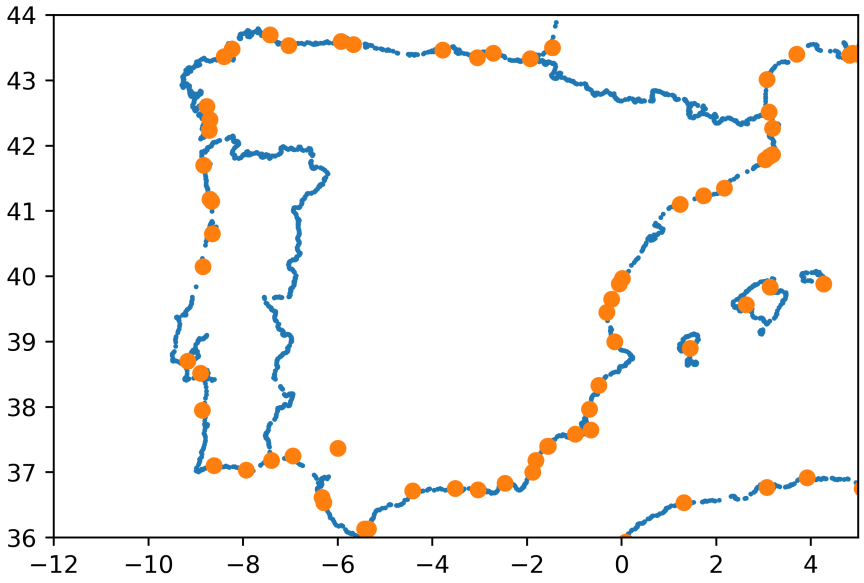

FIGURE 4.3: Iberian Ports(in orange), with the considered coastal Points(in Blue)

4.4.3 Stopped/Moving

Enriching the reported *BPs* by determining if at this point in time a vessel was in fact moving or stopped represents an overall information gain over the whole vessel trajectory. Such information can be used for the understanding of the normal vessels behaviour, or the detection of global points of interest. Thus, in order to gain such information, we used two different method. The first one was a point based approach, where we infer if whether a vessel is stopped or moving based on the its last report, this method is described under in this Subsection.

[9]http://dbfread.readthedocs.io

The second approach involves the use of a vessels past trajectory information, we present this approach further in this Chapter in Section 4.5.2.

Rule Based Approach: This approach is vastly used in the literature, as it is the simplest way to characterise the stopping of a vessel, based solely on the vessels reported speed or as reported by the AIS the Speed Over Ground (SOG). Thus, a BP which has a reported speed under a certain defined threshold Δ is considered as stopped and the opposite are considered moving. As it is shown in equation 4.2, where BP_n represents actual Behavioural Point we want extrapolate the stopped or moving feature.

$$kinematicstatus(p_n) = \begin{cases} BP_n.SOG > \Delta; & Moving \\ BP_n.SOG \leq \Delta; & Stopped \end{cases} \tag{4.2}$$

The most commonly used Δ value found in literature was 0.5 knots. This approach despite fitting most of the vessels behaviours, for the some types of fishing vessels it does not fit such behaviours. This occurs as some fishing activities, require the vessel to be drastically slow down for short periods of time. In Figure 4.4, we present a fishing vessel trajectory, where the points represented in blue were to be considered as *Stopped* if a $\Delta = 0.5$ was to be considered.

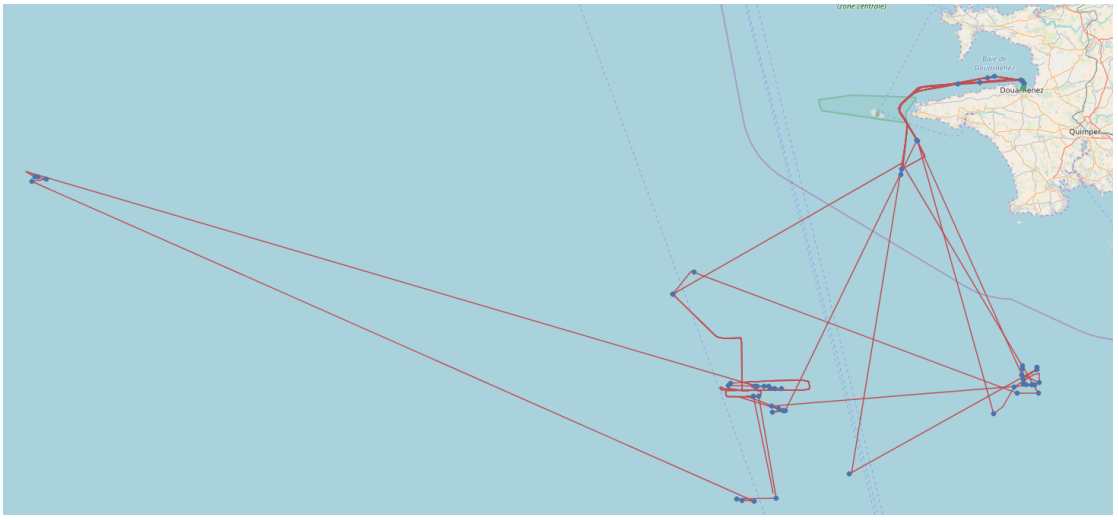

FIGURE 4.4: Fishing vessel (MMSI: 228858000) trajectory. Where the Blue points represent the Stopped points on the overall trajectory.

4.5 Trajectory Extraction

In this Section we present our interpretation and the definition of what was for this work considered as a vessel trajectory.

4.5.1 Trajectory Definition

Representing a trajectory in a optimal way, can become a difficulty task in the maritime domain. Currently there are a vast number of solutions described in the literature. They, represent a trajectory differently, depending on the type of problem.

Our approach to represent a maritime trajectory, was to consider a trajectory as a whole. This is, as vessels are obliged to broadcast their AIS information in a semi-continuous rates. By normalising each broadcast into the defined **Behavioural Point**(BP). We can aggregate each BP based on the BPs_{MMSI} vessel identifier which is the vessel MMSI. Thus the aggregation of BPs_{MMSI} represents for us a trajectory, which is represented as:

$$TR_{MMSI} = BP_{MMSI_1}, BP_{MMSI_2}, BP_{MMSI_3}, BP_{MMSI_4}, \cdots, BP_{MMSI_n}$$

Every trajectory is then sorted, and kept sorted based on the timestamp of each BP_{MMSI}. The representation of the BPs over a time allows us to consider a each trajectory (TR_{MMSI}) as a multivariate time-series. Each trajectory, can be then defined as a group of N time-series. Where N represents the number of features considered for the BPs definition.

Nevertheless, what was considered as most relevant, for our definition of a trajectory was the effectiveness, and scalability of such representation. This is, the effective adding of new BPs to a trajectory, and the accessing of historical trajectories in effective time. We achieved this by implementing the data-base in Apache Cassandra. From such we defined a set of pre-defined queries to which allowed the effective access to a whole or partial trajectory, in near real time.

In Figure 4.5 we represent an example of the partial vessel trajectory which was plotted over a map.

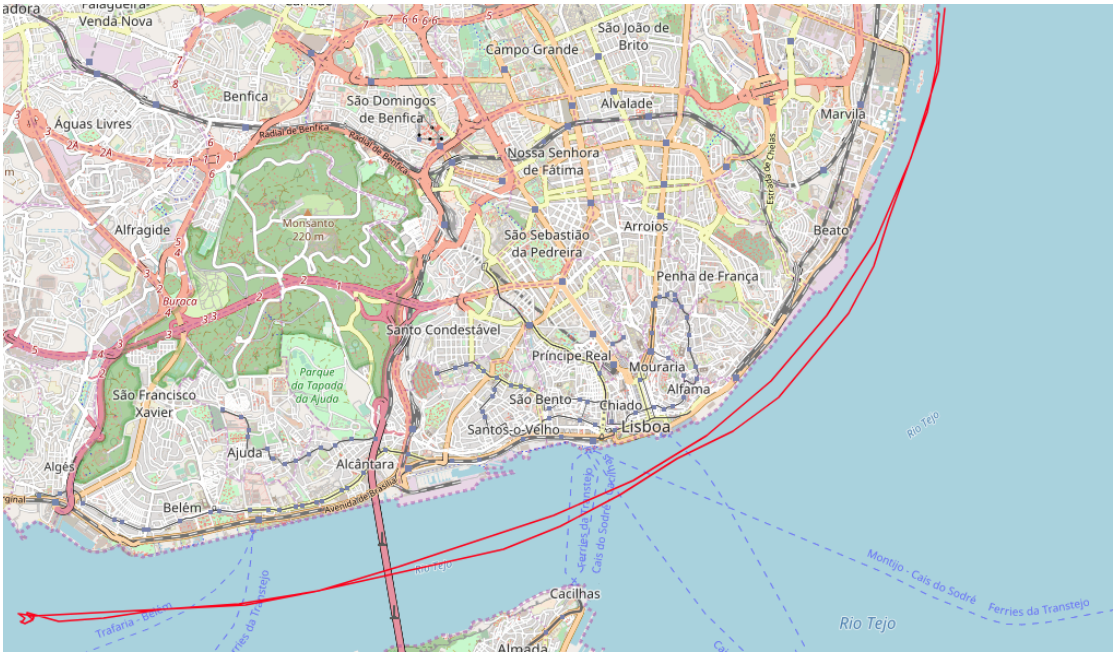

FIGURE 4.5: Trajectory snapshot(2017-11-05 10:22 to 2017-11-05 22:42) from Vessel MMSI: 255806006

The same trajectory plotted Figure 4.5, can be also represented as a multivariate time-series, as it is represented in Figure 4.6. By just considering the four most relevant kinematic features of a BP, the positional features (where x represents the Longitude, and y represents the Latitude) and the speed and course features.

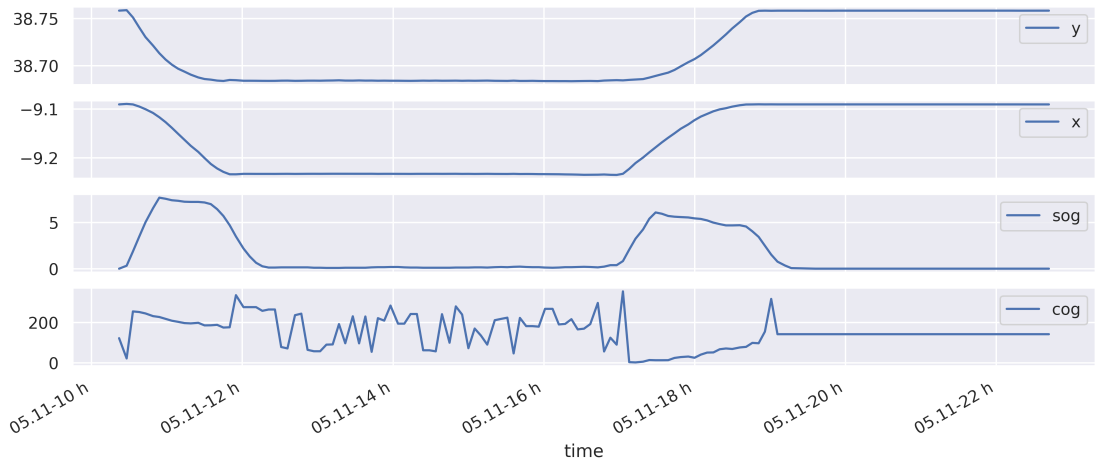

FIGURE 4.6: Trajectory represented in Figure 4.5, presented as a multivariate time-series.

4.5.2 Smoothed Stopped / Moving

In order to resolve the problem presented in Section 4.4.3, where the rule based stopped/moving approach had problems when dealing with some type of fishing activity trajectories. And also as a trajectory could be overseen as a multivariate time-series. We used a commonly used time-series analysis technique, *Rolling Mean*. By smoothing the vessels *SoG* time-series, based on the previous configurable W BPs, where W represent the window size considered. We smooth the random or abrupt variations in the observed speed features, which will in the end better describe the kinematic movement behaviour presented by these fishing vessels. The configurable W, allows the end-used of this framework, to configure the smoothness of the over the reported speed feature. This ultimately leads to a better representation of the vessel kinematics, which will be used for the anomaly detection methods presented in Subsection 4.6.1, 4.6.2 and 4.6.4.

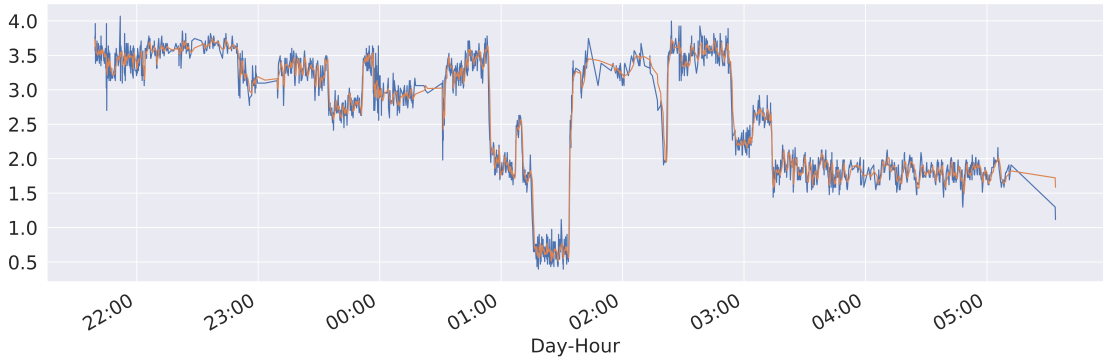

FIGURE 4.7: Snapshot of Trajectory represented in Figure 4.4 SOG feature presented as a time-series.

4.6 Anomaly Detection Service

4.6.1 Time-Space Incompatibility

Time Space incompatible corresponds to an anomalous or incoherent situation where the reported actual vessels position is not compatible if compared with previous reported positions, and vessels kinematics. The detection of this situation,

is also represented as an Anomaly Requirement(AR_4), Section 3.1.

In order to detect this incoherence's, we developed a method that takes as input an historical vessels trajectory TR_{MMSI}, and for each Behavioural Point ${BP_{MMSI}}^{T-1}$ we estimate the vessels position at instance ${BP_{MMSI}}^{T}$. The estimation is done by assuming that a vessels movement can be represented in a *Linear Motion*. As vessels tend to move in the most economical way, the Vessels travelled distance, was calculated, using the formula:

$$Distance = Velocity \cdot \Delta Time \tag{4.3}$$

Where $\Delta Time$ represents the actual time shift from point $(T-1)$ to (T). The $Velocity$ represents the BP_{SOG} feature, which is reported in knots. The $Velocity$ is firstly converted to m/s. By calculating the Equation 4.3 for each BP^T based on the reported Position of the previous BP^{T-1}, and assuming a vessel tend to move in a somewhat linear motion, we can predict that vessel should be in a distance radius of D for the next BP^T, as it is shown in Figure 4.8.

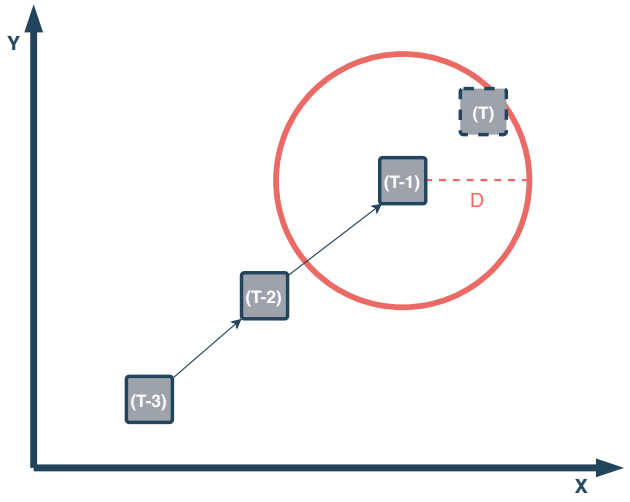

FIGURE 4.8: Linear estimation based on the previous reported BP.

By defining a configurable *Distance Factor Threshold* dft, representing a factor that would be multiplied by D, is possible to deduct that, if the a vessel at Time BP^T is at a distance superior than $(D.dft)$, it is considered at a incoherent position. Therefore it is reported as anomalous.

The emphasis of this Section was on the detection of Time-Space Incompatibly, which is depend on the level of error acceptance is achieved using the methods above. Although by assuming that vessels have a huge inertia, making them unable to perform quick changes of speed and direction, the authors in [30] present a *Linear Estimation Algorithm*. As the reported CoG represents the direction of movement, it is possible to based on Equation 4.3, to estimating the position of the Vessel, opposed to the distance from previous position. This is done by firstly calculating the Latitude and Longitude *shift* based on the BP^{T-1}, using Equation 4.4.

$$\begin{aligned} \Delta X &= Distance \cdot sin(COG \cdot \pi/180) \\ \Delta Y &= Distance \cdot cos(COG \cdot \pi/180) \end{aligned} \tag{4.4}$$

Where ΔX and ΔY represent the Longitude and Latitude features shift respectively. *Distance* represents the distance which is calculated using the Equation 4.3. Finally, the estimated coordinates of the vessel are:

$$\begin{aligned} X' &= X + \Delta X \\ Y' &= Y + \Delta Y \end{aligned} \tag{4.5}$$

4.6.2 Navigational Status Validation

AIS Navigational Status describes the vessel current activity based on a set static set of defined status, as shown in Table 4.4.

Table 4.4: AIS Navigational Status enumeration.

Navigational Status Value	Description
0	under way using engine
1	at anchor
2	not under command
3	restricted manoeuvrability
4	constrained by draught
5	moored
6	aground
7	engaged in fishing
8	under way sailing
9 - 14	reserved for future use
15	Default

The Navigational Status requires to be manually set, and constantly updated (according to the current vessel activity), by the vessel crew members. This creates the problem of relying on Human action to update the actual vessel navigational status, which is prone to errors. The use of the wrong navigational status being considered an Anomaly represented as (AR_4) in Section 3.1, our approach towards the detection of such Anomaly, started by firstly gaining insight of each navigational status, and their usage at seas. By accessing maritime knowledge via Maritime Experts, we enriched our previous description of each navigational status, by classifying the appropriate *Stopped or Moving Label* to each status. Maritime Experts based this classification, on the expected kinematics of each navigational status provided the following Table 4.5.

TABLE 4.5: Expert stopped or moving label over the AIS navigational status.

Expert Label	Navigational Status Number
Stopped	1, 5, 6
Moving	0, 7*, 8
Non-Quantifiable	2, 3, 4, 15

7* (Engaged at Fishing) represents a special navigational status which cannot be validated with a stopped or moving analysis. Our efforts to validate this specific status are presented under in Subsection 4.6.3.

The actual navigational status validation, is done as a *point based comparison*. By comparing the previous BPs enriched feature *Smoothed Stopped or Moving* (Section 4.5.2) with the stopped/moving label Maritime Experts has defined for each Navigational Status. An example for this validation could be: If a BP has been received with the Navigational Status *0 - under way using engine*, but the reported Kinematics describe it as *Stopped*, which for this Status should be *Moving*.

4.6.3 Fishing Activity Detection

Based on the Navigational Status Validation presented above we decided to enrich this methods with the detection of a special navigational status, the fishing activity (Navigational Status - 7 - Engaged in Fishing).

Fishing is a activity that generates huge profits for the global maritime lobby. This activity being so profitable and competitive between fishing companies, generates a problem for the Maritime Authorities. To avoid informing other fishing vessel of lucrative "fishing spots", fishing vessels try to hide their location as most as possible. This behaviour is anomalous when AIS is turned off. Fishing vessels are more prone to undermine the fishing competition, leading to Illegal Unreported and Unregulated fishing [10](IUU). Linked to IUU is the depletion of fish stocks, as well as the destruction of marine habitats and therefore putting honest fishers at

[10]http://fao.org/iuu-fishing/en/

an unfair disadvantage and thus weakening coastal communities, particularly in underdeveloped countries [31].

A vessel fishing activity is commonly defined as the period of time by which a vessel has fishing gear in the water. Since at the time we didn't have access to any Maritime Expert Classified datasets, we focused on the validation of this particularly navigational status (7 - engaged in fishing) by analysing the kinematic features capable of inferring whether the vessel is currently fishing or not. The main characteristic that identifies the fishing activity is the fast variation of direction together with a change in the speed. This can be seen as a generalisation as there are multiple different fishing, each different one having its own specific kinematic behaviour. Nevertheless, we claim that a fishing behaviour may be reasonably assumed to be highly dependent on speed variations. Specifically speed variations allow the fishing activity to be described by two main behavioural patterns. The first one is the *high speed behaviour*, typical of a vessel steaming at normal cruising speed from a fishing spot to another. The other one, the *low speed behaviour* is represented by the speed when vessels tend to drop the fishing gear in the water, or preform other manoeuvres which can be related to the fishing activity itself. Based on the work presented by the authors in [32, 33, 34], this double speed behavioural profile may be represented as a bi-modal distribution of speeds. Assuming that the speed profiles are characterised by only two speed modes, it is reasonable to apply Expectation Maximisation Gaussian Mix Models in order to estimate two distribution parameters, namely the respective mean and standard deviation of each mode, and to assign the observations to one of these behavioural profiles. Such bi-modal Gaussian distributions may be appreciated in Figure 4.9, where a typical histogram of the distribution of the reported speed is plotted.

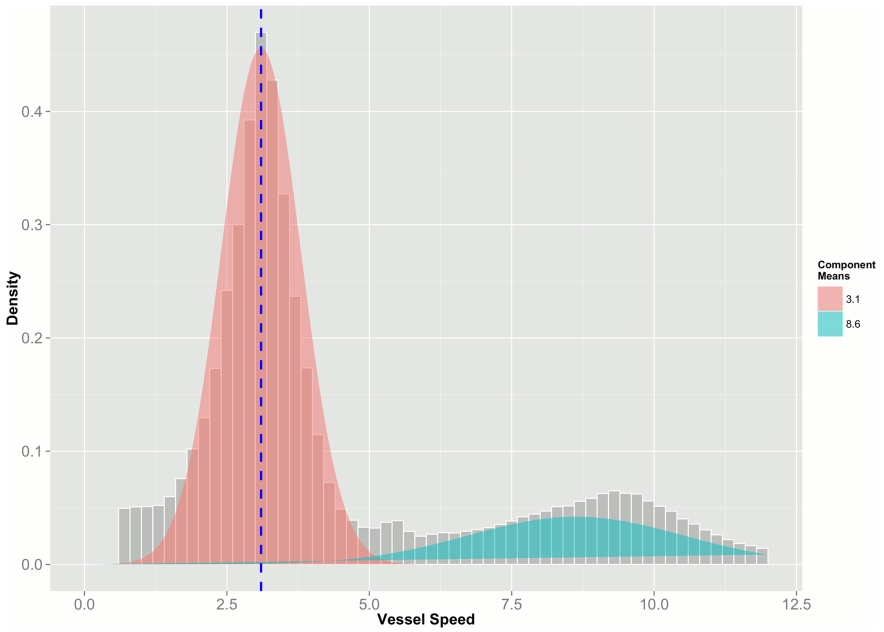

FIGURE 4.9: Example of speed profile for one vessel fitted for a bi-modal distribution, from [33].

From the presented dataset in Section 4.1 we fitted a Gaussian mix model to our data. This was done based on the approach presented by the authors in [33], where we used the already developed methods of [11]. From the already processed dataset, we filtered only the *BPs* that were type Fishing Vessels (Vessel Type 30). This reduced the number of considered *BPs* to approximately 3 Million. To avoid irregular vessel movement, and analyse the vessel movement patterns that could induce the vessel fishing activity, we filtered the *BPs* that would be a distance of more than 2 Nautical Mile from shore, leading to a sub dataset of 436, 043 possible Fishing *BPs*. The results and the discussion from the usage of this module are presented in Subsection 5.3.3.1.

4.6.4 Vessel Rendezvous

Another anomaly requirement *AR6* which was defined by the MARISA project, was the development of services, able to detect when two or more vessels are approaching close to each other. The detection of this anomaly is complex, as it

[11] http://scikit-learn.org/stable/modules/mixture.html

can occur in multiple scenarios. Although, for this current work we focused on the detection rendezvous, from huge batches of historical data, in a effective way.

Rendezvous occurs when two vessels meet allowing for the transfer of cargo, fuel, provisions, fish catch, crew or gear from one vessel to another. When transshipping takes place far from port, it can allow fishing vessels to avoid scrutiny at port and conceal suspicious activities like illegal fishing. But most alarming this practice leads to other nefarious activity, ranging from smuggling to human trafficking, [35].

Nevertheless, the concept of rendezvous is still quite complex to formalise by Maritime Officers, as there numerous legislation. Thus, for the purpose of this work, and because the emphasis is on the detection of possible rendezvous, a simplification of this vessel interaction is assumed, therefore:

Vessel Rendezvous, is then defined for this work as, the interception or closeness of two or more vessels, in a configurable time period.

In order to detect the rendezvous occurrences from multiple vessels, each single trajectory is partitioned into **t** time-groups e.g. a time-group of 5min. Thus for each rendezvous analysis the maximum number of comparisons are the number of **t** is defined by the trajectory with the oldest stored BPs.

After all trajectories are grouped into N time groups of size t, for each time group, if two or more vessels have reported in the same time-group, the Haversine distance (Formula 4.1) between every combination of two vessels is calculated. If the any **C2** calculated distance is smaller than **d**, an rendezvous anomaly is generated for those two vessels. **d** represents a configurable distance threshold for a rendezvous occurrence.

The method was implemented in such way, that the scale of approximations made could be controlled by the input configurations. But also, allow the input configurations defined the dimensionality of the problem. As when defining the time-groups size **t**, this in fact defines the number of group validations will be calculated, but also determines the granularity of the detection.

Figure 4.10(Left), shows two different vessel trajectories. While is obvious that the routes are similar in a positional way, they occur at different times, as is can be see in the Figure 4.10(Right).

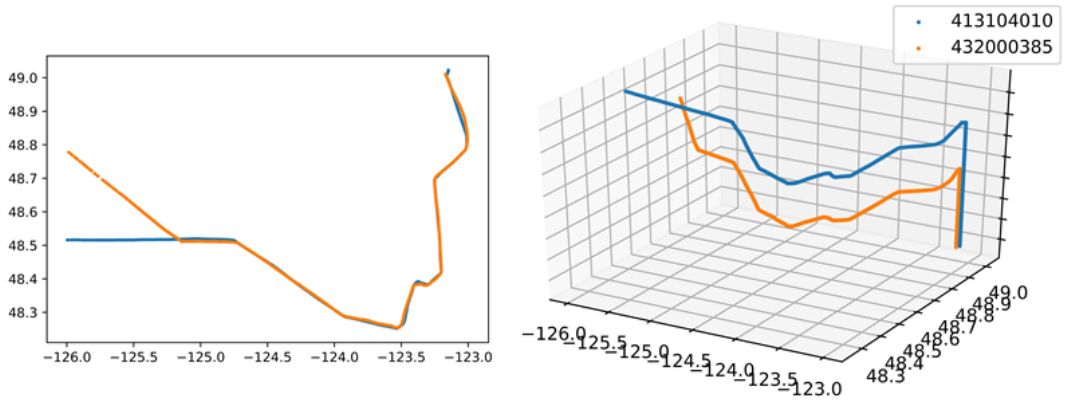

FIGURE 4.10: Routes of MMSI: 413104010 and 432000385; axes representing (lat.,long.)[Left] and (lat.,long.,time)[Right]

4.7 Rule Based Anomaly Detection Service

A Rule is defined as something that can, at least in the way we approach them, be expressed as an if-then sentence, [36] Anomaly Detection based on the definition of rules, is extremely used in the literature, as it represents an effective way to detect Anomalies at seas. Although this in only viable if and only if the rules are defined by Subject Matter Experts (SMEs) [37, 38].

Rule Based Anomaly Detection Service, was developed to detect Anomalies that can be codified into a rule or a set of rules, in real-time. Our approach to detect Anomalies in real-time was by storing the N last Behavioural Points for each Vessel in a Service Cache, working like a first in first out (FIFO) queue. N is a configurable Value, which represents the limit of messages stored in Service Cache for each Vessel, we provide an intuitively way to reduce the hardware requirements to run this service in Real-Time.

When the Service Cache has stored N BPs for a certain Vessel the Rule Based Anomaly Service is called for this Vessel, as demonstrated in Figure 4.11, in Vessel MMSI n case. If a Vessel queue has N BPs, and was the Rule Based Anomaly

Service was already called of this messages, all Vessel Service Cache being FIFO, we discard the oldest BPs, thus the Vessel Queue after is full for the first time it as always N.

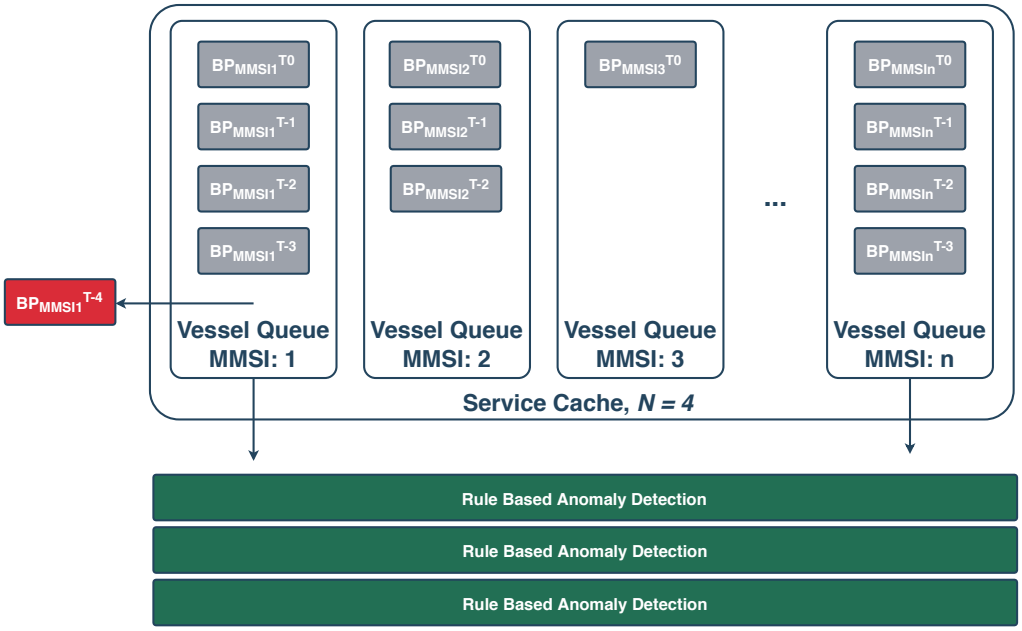

FIGURE 4.11: Demonstration of a possible cases for RB-ADS Service Cache coordinator.

Rule-Based Anomaly Detection Service, allows a detection of Anomalies based on little previous knowledge of each Vessel Trajectory. Despite the Service could be configured with any Rule that could be written for either Temporal, Spatial or Features based Rules, for example:

If Vessel MMSI: X in Zone: Y Stopped for more than M minutes then Report as Anomaly.

For this work we decided to focus on the creation of configurable Rules that could detect the Anomalies presented in Section 3.1, as were the Anomalies that would be Validated by Maritime Officers, which we present in Subsections Under.

4.7.1 Speed

Speed represents a Set of configurable Rules that were implemented to detect the Anomaly, (AR_2), defined in Section 3.1.

Our first approach for the detection of an *Abnormal Change of Velocity* was, calculating the *SoG* difference from the latest received, BP^T with the previous BPs^{T-1}, stored in the Vessel Queue. If the difference is bigger than a defined *SoG Treshold*, then an Anomaly is generated with the BPs stored in the Message Queue, and the configurations that generated this Anomaly, which could be represented as:

if $abs({BP_{MMSIn}}^T.SoG - {BP_{MMSIn}}^{T-1}.SoG) > SpeedThreshhold$
then *Anomalous*.

Although, calculating the *SoG* difference was only viable if $N = 2$ was considered as the size of the Vessel Queue. When considering more than two BPs the difference is not representative of the actual *Abnormal Change of Speed*, in order to mitigate this we created a new configuration, representing the operation it should be done in this case, which for this anomaly we considered the Average Difference, the Max Difference.

4.7.2 Course

Similar the **Speed** defined above, Course represents a Set of configurable Rules implemented for the Anomaly detection, of (AR_1), which is the *Detection of Abnormal change of Direction*. Our approach for the detection of (AR_1) was quite similar to the Detection abnormal change of Velocity. Although, we noticed depending how Vessels are Moored at port[12] Vessels tend to swing, due to the Sea Currents, or just from the movement of other Vessels moving in Ports. This creates abrupt changes of CoG, which are not representative of the Anomaly Requirement

[12]http://marineinsight.com/marine-navigation/mooring-methods-ships/

(AR_1). In order to mitigate this problem, we defined other configurable condition, Minimum SoG Threshold. Thus, the representation of the configurable Set of rules for the detection of rule for the detection of (AR_1), is :

if $abs({BP_{MMSIn}}^{T}.CoG - {BP_{MMSIn}}^{T-1}.CoG) > CourseThreshhold$
and ${BP_{MMSIn}}^{T}.SoG > minSpeedThreshhold$
then *Anomalous*.

4.7.3 AIS Signal Loss

The *disappearance from sensor coverage for more than a configurable Time Period*, (AR_3), from a data stand point the is represented as the loss of signal, or in other words, the non reception of AIS messages from this Vessel for more than M Minutes.

In this work, we detect the loss of signal from Vessel, by analysing when did a certain Vessel transmitted for the last time. This is done in real time, by the RB-ADS by one of two ways: the A priori way or the posteriori way.

First, as we store the Last N BPs for each Vessel that we received AIS messages from, in a the respective Message Queue from the Service Cache. By calculating the difference between the last received ${BP_{MMSI}}^{T}$ to the ${BP_{MMSI}}^{T} - 1$, we can know what was the elapsed time. Therefore, if this elapsed Time is bigger than a configurable Time M, this is reported as Anomaly. This method is considered a posteriori, as we are waiting for a new message to generate a Signal Loss Anomaly.

The a priori way, is when a Signal Loss anomaly is generated with out the reception of the a new message of a certain Vessel. Having the latest BP for each Vessel stored in the Vessel Queue, if more than M minutes have passed without receiving a Message for this Vessel an anomaly is generated. Both methods of detection represent the actual Signal Loss from a Data Stand point, and depending on the situation both can generate value to the End-Users.

Chapter 5

MAD-F Evaluation

In this Chapter, we present a group of experiments on the MAD-F capabilities. Firstly, we undertake data ingestion and storage performance test, which is followed by a exploratory analysis of the gathered data. Secondly we validate each of the ADS modules, starting with the RB-ADS Experiment, where the data that was collected from the previous Experiment is injected in this module. This was achieved with the development of a simulator. After an Experiment over the ADS is conducted, where for each of the anomalies that are detected we provide and exploratory analysis of the results. The *validation* of the results presented in this Chapter, can only truly be done by Maritime Officers. What is to be called as anomalies in this work must not be interpreted as an actual maritime illegality, but only as a possible anomaly, which needs always to be validated by Maritime Officer. The real validation of the developed MAD-F will be done by the project end-users, the Maritime Officers and Experts. MARISA being an highly collaborative and undergoing project, such validation at the time of writing this dissertation were still to occur. The validation of the MAD-F by the project end-users is described under in Section 5.4. All the Experiments under in this Chapter were conducted on a Desktop PC using a Intel Core I5-7600k CPU with 16Gb of RAM.

5.1 Data Ingestion Experiment

Data Ingestion Experiment refers to the Experiment where we assessed the performance of the Data-Ingestion capability of MAD-F. In order to achieve this, we provided a real NMEA feed as input to our the Data Ingestion Module. The NMEA feed was provided by the Portuguese Navy via the MARISA project, and this specific feed aggregated messages from multiple antennas around Portugal.

With the provided feed, we allowed the MAD-F to be executed for FIVE STRAIGHT DAYS, thus ingesting pre-processing and wrangling the NMEA feed into Behavioural Points. As for this experiment we used a real NMEA feed, the messages were firstly decoded into a readable format, and only then after the whole pre-processing was done, the BPs were stored in the Trajectory Extraction Cassandra Database.

From the 5 days of data acquiring, we acquired from a total of $2,259,615$ BPs from $5,563$ different Vessels. As the provided feed did not broadcast any vessel static information, from the vessel that generated each message. The vessel static information namely the vessel type and country of origin, were scrapped from the internet using the developed *Vessel Type Scrapper* which we presented in Section 4.4.1. From the $5,563$ vessels, 6 of them were not considered for this Experiment. The MMSI of this vessels was either not found or their MMSI was representative for more than one Vessel. The latter, represents an abnormal situation which could be denominated Spoofing, as represented by the authors in [39] an in [1]. This is a occurring problem when handling AIS data, and will be discussed in the future work. In Figure 5.1, we present the vessel type distribution from the acquired BPs.

[1] http://globalfishingwatch.org/data/spoofing-one-identity-shared-by-multiple-vessels

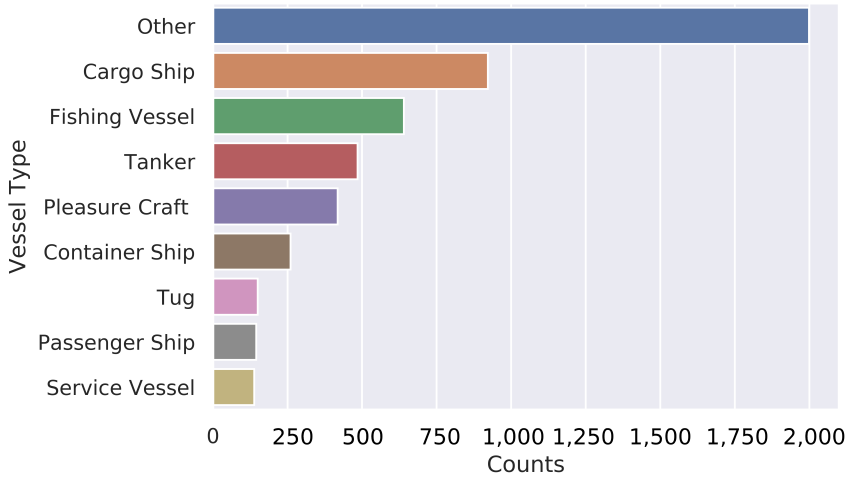

FIGURE 5.1: Vessel Type distribution of 5,157 Vessels.

Another feature calculated while transforming data into *BPs* was the point-based closest distance to coast and ports, which was done to every received NMEA AIS message received. In order to validate if this calculation were in fact being accurate, we analysed based on the received *BPs* which countries were the closest and the respective ports. In Table 5.1 we present the distribution of closest countries.

TABLE 5.1: Most Frequent Closest Countries Counts.

Country	Counts	Counts(%)
Spain	1,306,436	58%
Portugal	663,841	29%
Morocco	259,776	11%
Gibraltar	26,885	1%
France	2,516	0.1%

In Table 5.2 we present a post validation of the closest Ports for every received message. This was done as a way to analyse if based on the closest country the most closest ports seems plausible, as an individual message validation would be impossible. Thus, under we present a subset of top most frequent closest ports from the 69 total possible Ports found in the data.

Table 5.2: Most Frequent Closest Ports Counts.

Port Name	Counts	Counts(%)
Lisboa	142,464	6.3%
Villa Garcia De Arosa	112.254	5%
Europa Point(Gibraltar)	109,273	4.8%
Lagos	107358	4.7%
Las Palmas	106,116	3.5%
Cadiz	79,577	3.5%
La Corunha	78,929	3.5%
Malaga	65,227	2.9%
Vigo	64,759	2.9%
Faro	62,925	2.9%

In Figure 5.2 we display all the 2.2 Million BPs into a density plot, in order to analyse the positional occurrences from the transmitted messages. As scattering Millions of points is computationally heavy, if regular plotting packages were to be used, this would not be possible with the Hardware specifications presented in the beginning of this Chapter. Thus the density plots presented in this Chapter were done using the [2] package, which is optimised huge datasets.

What we found by analysing Figure 5.2 it is extremely likely that were in fact received by the Portuguese Navy antennas. This explains the reception of messages near the Madeira and Azores islands. What is possible also to analyse, is that nearby the Portuguese coastal line a few lines of high density traffic show up. These lines represent the navigational lanes, and when vessels navigate in this lanes, these tend to have a more standardised behaviour. Such fact can be explored for other methods of AD, as it is presented by the authors in [40] and [41]. The acknowledgement of this lanes for AD will be endorsed in Future Work.

[2] http://vaex.astro.rug.nl

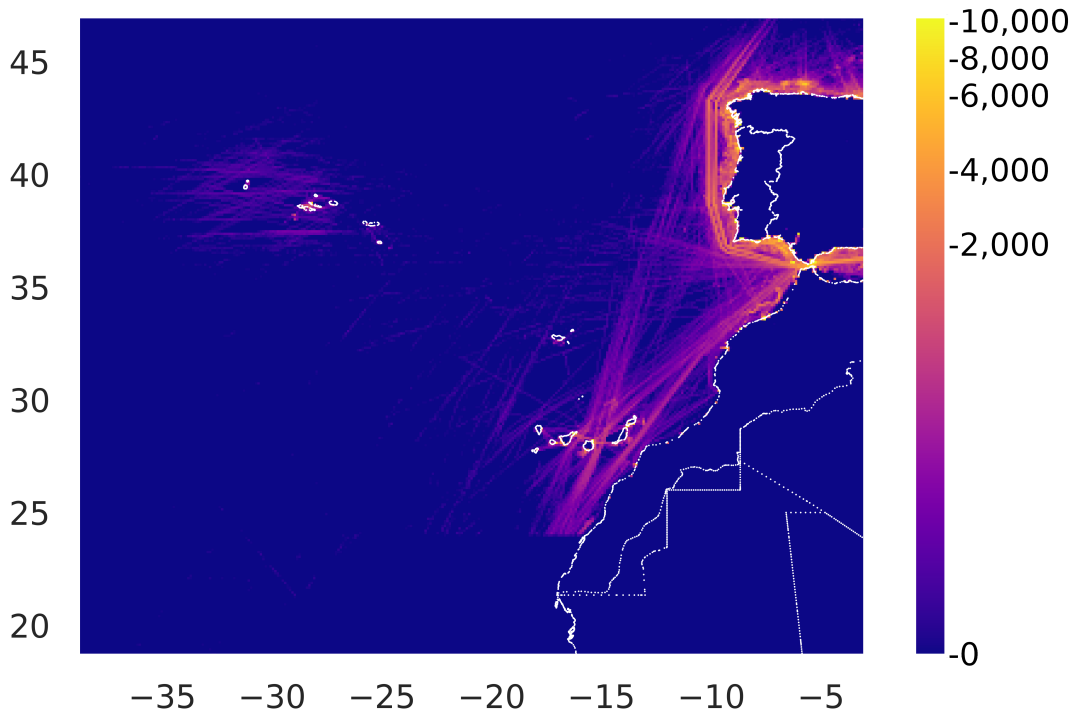

FIGURE 5.2: Density map, with all the approximately 2.2 Million points.

5.2 RB-ADS Experiment

This experiment was conducted, in order to validate the real time capacities of the RB-ADS Module. In order to validate such capacities, we focused this Experiment on the validation by analysis of the anomalies which were generated, and whether if this anomalies were generated in near-real time. This experiment was simultaneously conducted as the previous Experiment 3.2.1. This was possible as the incoming messages after being ingested and pre-processed they were stored in the Trajectory Extraction module as BPs but simultaneously the same BPs were inputted in the RB-ADS module. The RB-ADS module managed the incoming BPs, using the implemented service queue, which we explained in 4.7. For this Experiment we set the service queue N, size to 2. This made the *RB-ADS* to be executed for any vessel queue which had at least two BPs. The set of rules concidered for this Experiment, are represented under in Table 5.3.

TABLE 5.3: RB-ADS Experiment, Rule configurations, where the columns represent the features, and the rows the respective Rules.

Rule	SOG variation	GOG variation	SOG min.	Time Elapsed
R1	-	-	-	15min. (post.)
R2	>15 knot	-	-	-
R3	-	>25º	>0.5 knot	-

From the 5 days of executing the MAD-F, as mentioned in Section 5.1 a total of $2,259,615 BPs$ were ingested and pre-processed and finaly inputed into the RB-ADS module. With the presented set of rules a total of $191,481$ anomalies were generated. These number of anomalies is rather large. Representing approximately 8% from all received BPs to be considered anomalous for at least one of the presented rules. In Table 5.4 we detail more explicitly the total number of anomalies, by analysing the rules which generated such anomalies. As well as the matching the defined rules for this Experiment with the Anomaly Requirements, which were defined in 3.1.

TABLE 5.4: Anomalies found for each Rule with the respective Anomaly Requirement.

Rule	R1	R2	R3	Total
Count	82,866	2,144	106,471	191,481
Anomaly Requirement	AR3	AR2	AR1	-

From the results presented above what is possible to analyse, is that the most occurring anomaly was the *abnormal change of direction*, this despite the filtering of normal course variations on vessels which were stopped. We further analysed the time difference between same vessel transmissions. What we found out was that the *mean transmission rate*, was of approximately **10min.**, which is high for a real AIS feed. Although, if the BPs which were considered anomalous from the $R1$, were not considered for the calculation of this *mean*, the *mean transmission rate*

would be **5min.**. Thus, if a new rule were to be applied where from the $R3$ only messages that had been transmitted with a time difference inferior to 5 minutes, which could be represented as $COG.diff > 25\ and\ TimeDiff. < 5minutes$: only **7** anomaly occurrences would occur.

The additional rule presented above, was not validated with the live NMEA feed, as this was a live stream. Nevertheless, this rule was still validated with the same data. As the results from the Experiment 5.1 were stored in the trajectory database, these could be accessed multiple times. For the purpose of these current work we developed a BPs simulator which from the stored trajectories would simulate the real reception of AIS streams (from the perspective of the RB-ADS module).

The simulator gathers BPs from the trajectories (or a group of) stored in the trajectory database, and send this BPs to the RB-ADS, as presented in 5.3.

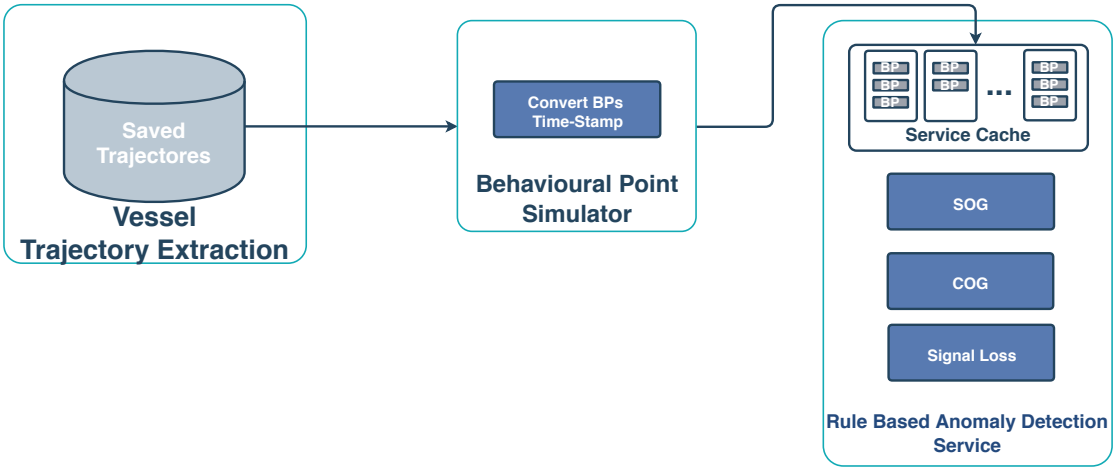

FIGURE 5.3: BPs Simulator.

The simulator worked by setting the a **Initial-Simulated-Time** as if it was the current Time. This time was the first timestamp of the stored BPs. Based on this **Initial-Simulated-Time**, the following BPs would be sent based on the time difference from this **Initial-Simulated-Time**. As it would be impractical to wait 5 days for the simulation of the reception of this data, the developed simulator was implemented with a *speed up factor*. Therefore, for every time considered by

the simulator would be divided by the *speed up factor*. This some what allowed us to replicate the reception of the same data, at extremely higher throughput, while simultaneously allowing the analysis from the results of different rules.

5.3 Anomaly Detection Service Experiment

ADS Experiment for this current work represents our validation process of the developed offline AD functionalities. The steps towards this experiment were similar to the Experiment 5.2. As this module was developed to access more complex anomaly detection methods by the use of batches of historical data. We did not conduct this experiment same data as presented in 5.1 or had a similar approach as for the Experiment 5.2. As we already had preconceived knowledge of the dataset which we used in our initial data-analysis (Section 4.1). We conducted this Experiment with the same dataset, as it in fact represented an huge batch of historical data. presented an huge batch of data.

Before performing of the *ADS Experiment* itself, the raw dataset was injected in the MAD-F as a single batch of data. This transformed a the historical AIS dataset into a normalised set of BPs, which was kept stored in *Trajectory Database*. What is to note is that if this group of BPs were to be stored as files, these files would be nearly 5 GB(if stored as .csv type files). As the transformation of the dataset in BPs made the dataset pass through the "pre-processement" pipeline. This cleaned the whole dataset which was of initially 18.84 Million rows (AIS messages), from 4,555 different Vessel, into approximately 17.10 Million BPs. After the BPs were store as trajectories, and additional "manual" filtering was done. We filtered the trajectories with a size inferior of 100 BPs. This filtering only slightly reduced the number of total BPs considered for this experiment to 17.06 Million, although the number of considered Vessels was dramatically reduced to 1,588 Vessels.

The ADS experiment was divided into two Sections. The first section presents the results obtained for the Vessel Rendezvous detection, and the second section

presents the results for the Incoherent Navigational Status and Time Space incompatibility. The results are presented as the generated anomalies from the ADS module. From each subsection we present an explanatory analysis of the generated anomalies.

5.3.1 ADS - Rendezvous Experiment

This subsection, shows the results and our analysis of the results obtained from the Rendezvous sub-experiment. This sub-experiment was conducted on the historical batch of data which was described above. The Rendezvous detection, as any other module of the proposed MAD-F was developed to be configured with the set of parameters most adequate for the situation which would be deployed. This choice of parameters in any real scenario would be done by Maritime Experts. Although for the sole purpose of this Experiment, the choice of parameters was done by us. This Experiment was conducted with four different sets of configurations(Table 5.5).

Table 5.5: Anomaly Detection Service - Rendezvous input parameters.

Rendezvous Parameter	BPs	Time-Window	Distance Threshold
C1	17.1M	10 min.	50 yards
C2	17.1M	2 min.	50 yards
C3	17.1M	10 min.	10 yards
C4	17.1M	2 min.	50 yards

The four different set of configurations were chosen in order to demonstrate the rendezvous anomaly detection capabilities. By varying the configurations, the rendezvous detection can either be done in a more precise way, or in a more efficient way. The variation of *Time-Windows* directly impacts the granularity of the detection and the variation of the *Distance Threshold* impacts the proximity the vessels were to each other. In the Table 5.6 we present the results obtained with the configurations presented above.

Table 5.6: Rendezvous experiment results, with the variation of the configuration parameters.

Parameters	Rendezvous Detected	Time Groups	Time Elapsed (aprox.)
C1	35,667	131,760	50s
C2	120,773	26,352	4min.
C3	5,704	131,760	2min.
C4	18,993	26,352	40s

From the results presented above, the first thing we noticed was that the number of occurrences was larger than expected. Regarding the variation of configurations what was found out, was that the variation of distance threshold impacts the number of possible rendezvous detentions, which was expected. What was not expected was the number of occurrences increasing with the decrease of the time-groups sizes. Although, after analysing the results, this results did exactly what the method was developed for. As for this work, we considered an anomaly to be a single instance in time, and not the time group of which the anomaly had occurred. When considering lower time-groups sizes if two vessels had report twice in same position, two anomalies would be created. For the purpose of this analysis, and in order to mitigate this duplication of technically the same anomaly, thus gaining insight of how many rendezvous had occurred. We grouped the anomalies, therefore, if the anomalies were generated by the same group of vessels, in consequent time-groups they would be considered the anomaly with same with a larger duration. What was discovered from this grouping of consequent anomalies was that, with the configuration parameters **C4**, only **75** combinations of two vessels generated rendezvous anomalies. Although each combination of vessels generated multiple times rendezvous occurrences, with some combinations generating up to **7,436** times.

After analysing the frequency of occurrence, we analysed the location where the possible rendezvous had occurred. This led us to conclude that most of the detected rendezvous occurrences occurred nearby port. As the only truly way to

validate the results was by providing this results to a Maritime Officer, which could not be achieved for the sole purpose of this present work. We decided to represent the Rendezvous occurrences which we detected on a distance of above 2 Km of the closest Port, thus creating a footprint of the possible Rendezvous occurrences for the whole dataset. A similar analysis is done by the authors in [35].

In Figure 5.4, we present the a visual representation of the locations where the Rendezvous anomaly had occurred, centered on the area nearby the port of Brest(France). By displaying only the rendezvous events that occurred 2Km away from the closest port, the number of rendezvous occurrences is significantly reduced, as it is possible to visualise in Figure 5.4(Right). Nevertheless, as the geographical representation of *a port*, for this work was considered as a single point and a port is larger than just a single point. Filtering by the distance to port, by just considering as a point can cause a occurrence to not be filtered, nut still be in port area. This problem is solved by knowing the geographic area of a port, and will be addressed for Future Work.

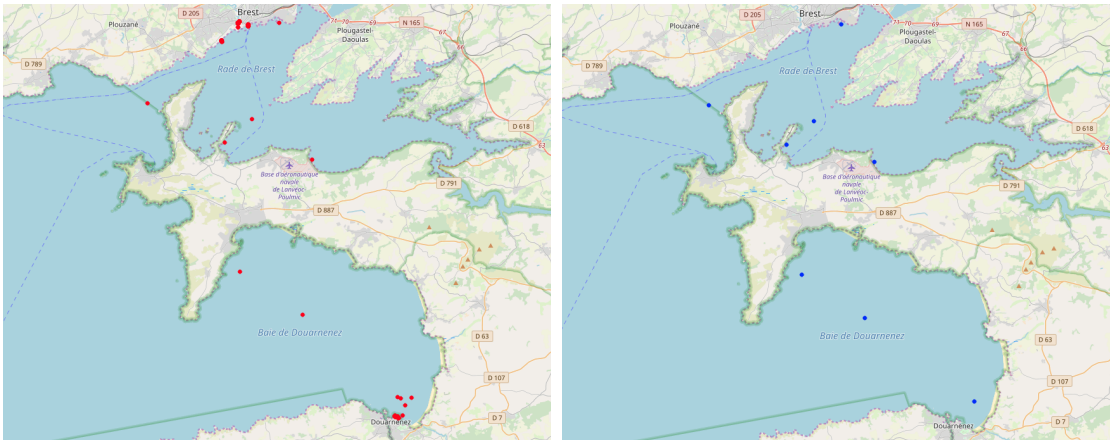

FIGURE 5.4: Rendezvous results, on the left no distance filter is applied and in right 2Km distance to port is applied.

5.3.2 ADS - Time Space Incompatibility Experiment

The time space incompatibility Experiment, serves for this present work as a Experiment in which we analyse the results from the ADS time-space incompatibility

anomaly detection module. In order to achieve this, we took two different analysis for the this Experiment. First we analyse the overall anomalies generated, by varying the input parameters of this service. Secondly, as this service represents a some what first approach towards a vessel positional estimation, we applied the implemented linear estimation to the vessel trajectory, which was presented in Section 4.6.1, to a single trajectory which was presented in Section 4.5.1.

By following a similar approach as the one presented for the rendezvous Experiment, we used the historical batch of data previously described also for the detection of the space time incompatibility. As the *Distance Factor Threshold*, is the configuration parameter for this specific anomaly detection, and it should be configured by a Maritime Expert depending on the scenario. For this Experiment we varied the dft and analysed the results. In Table 5.7 we present the number of occurrences of the what was interpreted for this present work as the **AR5**(presented in Section 3.1).

TABLE 5.7: Time space incompatibility occurrences, by varying the dft, and comparing with the BP time shift.

Distance Factor Threshold	Delta Time	<2min.	<5min.	<15min.	>15min.
500m	6,581	391	866	1,978	4,559
1km	4,289	187	229	685	3,569
2.5km	2,373	53	53	95	2,257
5km	1,353	38	38	40	1,295

What was expected from this anomaly detection, was that the number of detected anomalies would increase with the decreasing of dft, this was confirmed by our Experiment. From this conclusion we further analysed this results by comparing the number of detected anomalies with the time shift from the previous BP^{T-1}, as we described in Section 5.3.2. What is possible to analyse is the correlation from the time elapsed with the number of detected time space incompatibility occurrences. Thus, an Signal Loss Anomaly would also trigger a time space incompatibility anomaly.

Additionally for this Experiment, we applied the enriched Linear Estimation Equation which was presented in Section 4.6.1 to a single trajectory of a Vessel, as we present under in Figure5.5.

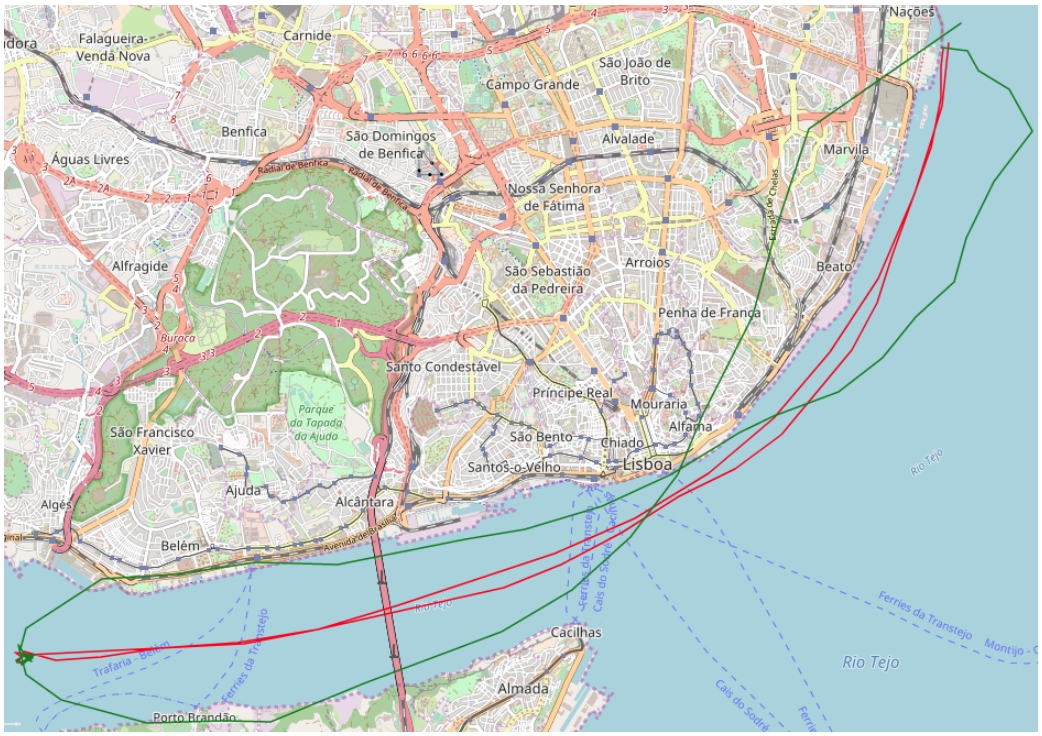

FIGURE 5.5: Linear Trajectory estimation (Green), applied to the Vessel Trajectory (Red) presented in Section 4.5.1.

For any historical trajectory it is possible to know where the was each for every transmission, by calculating the haversine distance between the estimated position at T^{-1} and the actual vessel position at time T. This distance represents the distance estimation error. For this trajectory which length was of 138 BPs, the mean distance estimation error was of **361 meters**. This result for this trajectory were suboptimal, as firstly a Vessel position should never be estimated to be in land, and secondly this specific trajectory had 98 BPs with a reported SOG under 1knot. Despite all, the presented sub-experiment, serves as a baseline, for the implementation of more advanced trajectory estimation methods as we will discuss in Future Work.

5.3.3 ADS - Navigational Status Validation Experiment

The Navigational Status Validation Experiment, was conducted with similar approach as the Experiment 5.3.1. This experiment, started with the analysis of the usage frequency of each Navigational Status, as it is presented under in Table 5.8.

TABLE 5.8: Navigational Status Counts, where the % is rounded to two decimal places.

Navigational Status	Count	Count(%)
0	8,895,694	52%
15	5,334,804	31%
5	1,030,712	6%
7	1,012,271	6%
3	391,141	2%
1	177,925	1%
8	71,664	0.0%
2	23,306	0.0%
6	14,955	0.0%
4	62	0.0%

In Table 5.8 what is possible to notice is that the distribution of the reported navigational status is extremely skewed. With approximately 83% of all the analysed BPs were reported as either Status 0(Under Way Using Engine) or 15(Default State). Despite this skewed distribution, the experiment was still conduced for the statuses that were quantifiable in a stopped or moving expert label, which was explained in Section 4.6.2. This ultimately reduced the BPs which were evaluated by this experiment. Nevertheless the experiment was conducted with 10.1 Million BPs, where the results under in Table 5.9.

Table 5.9: Results for Navigational Status Validation Experiment, with the Stopped or Moving approach.

Navigational Status	Count	Incoherent Count	Incoherent %
0 (using engine)	8,895,694	5,225,362	58.74%
1 (at anchor)	177,925	54,085	30.40%
5 (moored)	1,030,712	246,430	23.91%
6 (aground)	14,955	2,091	13.98%
8 (sailing)	71,664	24,607	34.34%
Total	10,190,950	5,552,575	54.49%

From the results presented above, it is clear that the major part of the used Navigational Statuses were reported wrongly. Similar results were found in [42], using a different dataset. A possible reason for such high number of miss used navigational status, might be justified by the fact that the Navigational Status is set by the crew on the AIS device. Although to try to better understand this results we started by analysing the areas where the miss-use of navigational status would occur. This analysis is presented in the form of a density plot, which was done using the same packages, as in Experiment 5.1.

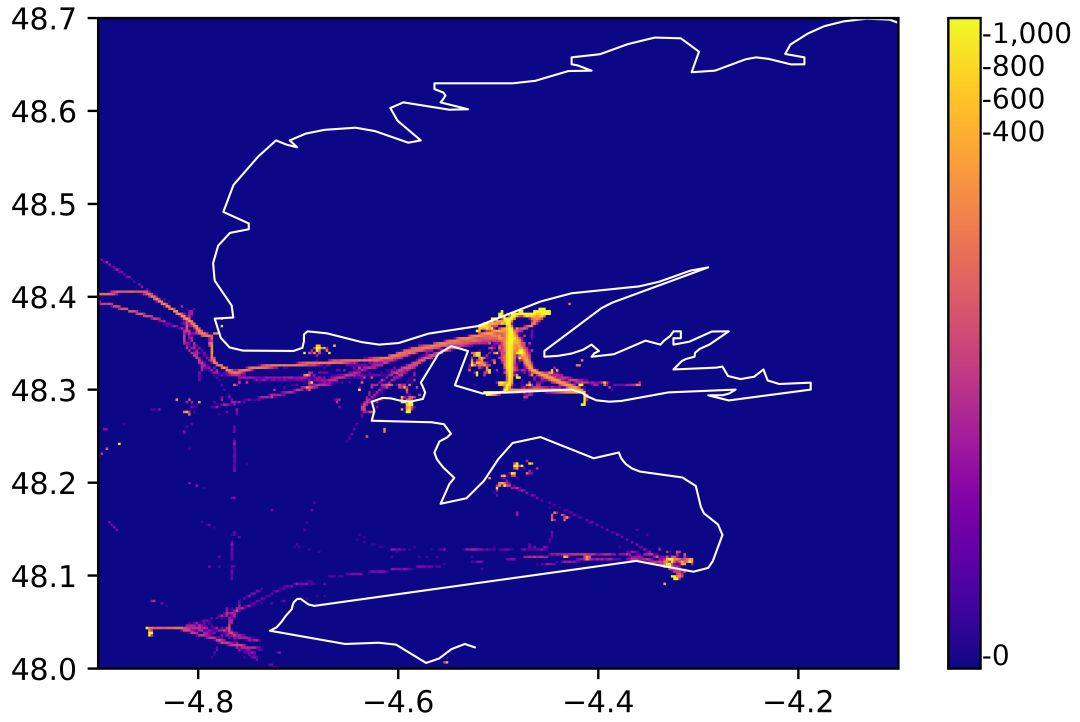

FIGURE 5.6: Density map, with all the 5.5M occurrences of wrong Navigational Status.

What we were able to analyse in Figure 5.6, is that the high density areas where the Navigational Status is reported wrongly, are areas really close to port. As the AIS navigational status needs to be changed each time the vessel arrives at port. This led us to believe that the crew members "forgets" to change the navigational status, while in port. Which is then represented on the data, as the Vessel being stopped on port for long periods of time with the navigational status 0(under way using engine).

5.3.3.1 ADS - Fishing Status Validation Experiment

Fishing status validation is a sub-Experiment related to the Experiment presented above. As to the best of our knowledge there are no current *public* classified fishing trajectories datasets, the approach taken to validate the usage of this specific status was merely by our analysis. What we done for this sup-Experiment, was the usage of the already pre-defined Gaussian Mixture Model presented in Section 4.6.3. The

model was applied on the sub-set of BPs which had the AIS navigational status reported as **7 - engaged in fishing**. Despite the acknowledged generalisation and the limited validation of the presented model the presented results serve as our first steps towards the implementation of a MAD-F *fishing detection module*. This will be discusses for Future Work in Chapter 6.

For this sub-experiment we first provide an exploratory analysis of the reported BPs and what would be expected to be reported from fishing vessels. From approximately 3 million BPs, transmitted by the fishing vessels (vessel type 30) only about 30% was in fact transmitted with the navigational status 7(engaged at fishing), as the rest half of them were reported with the default AIS status and as we present in Table 5.10.

TABLE 5.10: Fishing Vessels, Navigational Status Counts, where the % is rounded to two decimal places.

Navigational Status	Count	Count(%)
15	1,290,264	0.42
7	919,515	0.30
0	749,419	0.25
3	37,092	0.01
5	25,505	0.01
8	6,394	0.00
2	2,164	0.00
6	2,055	0.00
1	19	0.00

We further analysed the transmitted navigational status by analysing, if weather any other vessels had transmitted the engaged at fishing navigational status. As presented under in Table 5.11, we noticed that approximately 10% of the reported BPs were in fact transmitted by other types of Vessel.

TABLE 5.11: Vessel types which had reported the engaged at fishing navigational status.

Vessel Type	Vessel Type Nº	Count
Fishing Vessel	30	919,515
Other	90	92,576

After this initial analysis, we applied a fishing navigational status validation, by classifying the reported engaged at fishing *BPs* into a steaming (high speed) or fishing (low speed). From this results we further analyse this classification by comparing this results with the mean speed average, from each group, as we present under in Table 5.12.

TABLE 5.12: Fishing, not Fishing classification using the Gaussian Mix Module.

Label	Count	Mean SOG
Fishing	1,552,816	6.89
Not Fishing	1,487,490	2.60
Total	3,040,306	8.72

5.4 Marisa Validation Trials

This section presents how the validation of the developed MAD-F will be processed. The evaluation presented above serves as our own analysis and examination of the results produced by the developed MAD-F. Even though the present work was developed in a highly collaborative spirit, the results used must be evaluated under the project context in order for any method to be truly validated. In MARISA, this is achieved by including the different partners into what was defined in the project as **Trials**. A Trial represents a defined operational scenario where the project end-users will test the developed MARISA services. Despite the intermediate validation procedures applied throughout this work, the ultimate validation of the developed MAD-F is solely conditional on the performance of the project

trials. By aggregating the users by region of activity, five different trials were defined for the MARISA Project and they will take place until the end of 2018. INOV will be present in three of these Trials. For the purpose of the present work, we shall describe the Trial for which INOV contributed the most for the preparation of the Iberian Trial. The Iberian Trial will be conducted by the Portuguese Navy and the Spanish Guarda Civil, and will involve INOV and numerous other project partners. This will occur during the first fortnight of November, around the region of the Algarve (Portugal), as shown in Figure 5.7.

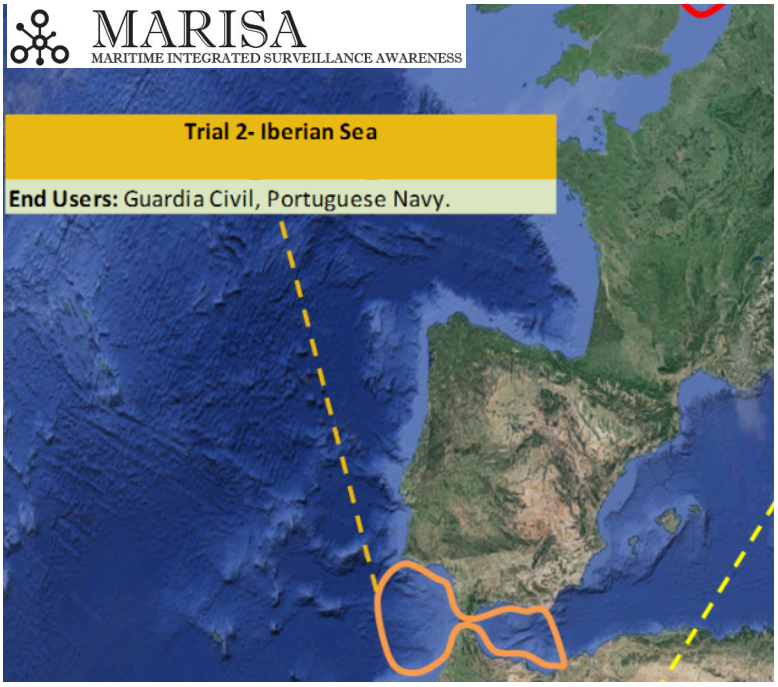

FIGURE 5.7: MARISA Iberian Trial operational area.

INOV was present in previous meetings with the Portuguese Navy, where the overall execution of the Iberian Trial was discussed. Within the scope of this work's efforts, we are currently ready and able to correlate the Trial activities with the effective validation of the Services and thus the MAD-F. The Iberian Trial will be conducted with real assets, including military vessels from both the Portuguese Navy and the Spanish Guarda Civil, the end-users of this project. Maritime Agents will perform a somewhat choreographed set of vessel manoeuvres able to trigger anomalies. In Table 5.13, we present the choreography exercise, which will be performed by the Maritime Agents, and how they are correlated

with the developed MAD-F (more specifically by correlating with the Anomaly Requirements which were presented in Section 3.1).

TABLE 5.13: Part of the Choreography conducted for the Iberian Trial. Where V1 and V2 represent a vessel from each end-user, and OC the Operation Control.

Provided Description	Choreography	Remarks
COMMCHECKS	at 1000: - OC - communications checks.	
TRANSIT TO HIGH SEA	at 1100: - V1 - turns to portside 40º, heading to 140º	
TRANSIT TO HIGH SEA	at 1130: - V1 - turns to starboard side 40º, heading to 180º - V1 - increases speed to 25KTS and maintains for 10min.	Aim is to detect Change in Course Over Ground (COG) and Speed Over Ground (COG).
TRANSIT TO HIGH SEA	at 1200: - V1 - Turns the AIS System from 1200 to 1220. - OC - Checks if it is detected a change in the AIS System.	Aim is to detect non-broadcasting (AIS)
RENDEZVOUS	at 1230: - V2 - approaches towards P1(TBD). - V2 - approaches towards P1(TBD).	
RENDEZVOUS	at 1300: - V1 - stops at P1 for more than 5min. - V2 - stops at P1 for more than 5min.	Aims to detect the repeated rendezvous of 2 Vessels.
TRANSIT ALONG SIDE COASTLINE	at 1400: - OC - manipulates the V1 and V2 AIS signal. - V1 - heads to port. - V2 - heads to port.	Aims to detect Incoherent Position and Navigational Status.

The choreography exercise just detailed, within the greater scope of the Iberian Trial, will test conclusively the capabilities of our methodologies. It is worth mentioning that a large portion of Data Science projects suffer from either absent or weak validation stages. This project, in turn, relies on actual, verifiable validation schemes which were devised and coordinated appropriately, given the excellent collaboration links provided by the MARISA project.

Chapter 6

Conclusion and Future Work

In summary, this work is concerned with the detection and identification of anomalies at seas. This task in itself, as was discussed throughout this work, is rather complex and can only be meaningful when a suitable definition of anomaly is applied to the problem at hand. Without such particular constraints, both the aim and eventual conclusions become of general scope. Greatly justified by this reason, we were fortunate to have discussed these technicalities with Maritime experts via the MARISA project. Their insights and feedback lead to a workable interaction level which ultimately allowed us for the development of a number of data-driven methods to be applied, resulting in our MAD-F.

The implemented MAD-F which was developed in accordance with the defined objectives presented in Section 1.1 is able to ingest and process high throughputs of real maritime data. This was achieved with our selection of technologies, namely Python's Pandas package and Apache's Kafka stream processing platform. This provided, generally speaking, a very effective and flexible framework to handle and structure the data. Despite the fact that these tools render the Framework modular and hence scalable, we believe improvements could be achieved by using Apache's Spark[1] cluster-computing framework.

[1] https://spark.apache.org/

Regarding the transformation of the data into sequential form, we introduced the concept of Behavioural Point BP, with which the maritime data features were transformed as to become more suitable for further manipulation. The output data at this stage serves as the adequate input for applying the Anomaly Detection. The AD modules implemented for the current MAD-F are rule-based, as opposed to more sophisticated methods. Despite the ever-growing demand for far more complex and involved methodologies, this procedure still yields very satisfactory results.

There are many reasons for choosing this procedure. Firstly, no publicly-available classified datasets exist for this source of data, to the best of our knowledge. This invariably forbids us from labelling anomalies without the explicit guidance of experts in the field. Secondly, applying more sophisticated methods requires a longer project execution time, which was a delicate issue from the outset. In spite of all these reasons presented above, we are confident the Framework is easily scalable for future development and offers an easy integration of new modules in the future.

MARISA is admittedly an ambitious endeavour which takes in contributions from multiple partners from all around Europe, as well as institutional agencies of different countries. Progress in highly technical matters is therefore inevitably slower than what would be expected from smaller projects.

It is our intention to look into new modules to complement and further expand the scope of this Framework. The linear estimation of the vessel trajectories used in our framework would benefit greatly by employing more intricate estimation methods such as Kalman filter [43, 44]. An explicit module for detection of fishing activity, as discussed at length by the Global Fish Watch project [2], would bring added value to the Framework. Lastly, the study of the vessel trajectories, which was based on a point-based analysis up to now, could be upgraded to a time series analysis. Related procedures such as multivariate time series clustering, already mentioned in the state of art in Subsection, 2.4.3 and 2.4.2.

[2]http://globalfishingwatch.org/publications/

Additionally, some modules from the developed *Modular Anomaly Detection Framework*, originated the work [42], which provided a particular emphasis on the detection of the Rendezvous anomaly detection.

Bibliography

[1] C. Bueger, "What is maritime security?," *Marine Policy*, vol. 53, pp. 159–164, 3 2015.

[2] S. Mao, E. Tu, G. Zhang, L. Rachmawati, E. Rajabally, and G.-B. Huang, "An Automatic Identification System (AIS) Database for Maritime Trajectory Prediction and Data Mining," in *Proceedings of ELM-2016*, pp. 241–257, Springer, 2018.

[3] J.-G. Lee, J. Han, and X. Li, "Trajectory Outlier Detection: A Partition-and-Detect Framework," in *2008 IEEE 24th International Conference on Data Engineering*, pp. 140–149, IEEE, 4 2008.

[4] International Maritime Organisation, "IMO - Regulations for carriage of AIS," tech. rep., International Maritime Organisation, 2015.

[5] A. Harati-Mokhtari, A. Wall, P. Brooks, and J. Wang, "Automatic Identification System (AIS): Data Reliability and Human Error Implications," *Journal of Navigation*, vol. 60, p. 373, 9 2007.

[6] P. R. Lei, "A framework for anomaly detection in maritime trajectory behavior," *Knowledge and Information Systems*, vol. 47, no. 1, pp. 189–214, 2016.

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

[8] H. Y. Shahir, U. Glasser, A. Y. Shahir, and H. Wehn, "Maritime situation analysis framework: Vessel interaction classification and anomaly detection," *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pp. 1279–1289, 2015.

[9] R. Laxhammar, "Anomaly detection for sea surveillance," *2008 11th International Conference on Information Fusion*, pp. 55–62, 2008.

[10] M. Hassel, I. B. Utne, and J. E. Vinnem, "Allision risk analysis of offshore petroleum installations on the Norwegian Continental Shelf—an empirical study of vessel traffic patterns," *WMU Journal of Maritime Affairs*, vol. 16, pp. 175–195, 5 2017.

[11] Z. Feng and Y. Zhu, "A Survey on Trajectory Data Mining: Techniques and Applications," *IEEE Access*, vol. 4, pp. 2056–2067, 2016.

[12] N. Le Guillarme and X. Lerouvreur, "Unsupervised extraction of knowledge from S-AIS data for maritime situational awareness," *Information Fusion (FUSION), 2013 16th International Conference on*, pp. 2025–2032, 2013.

[13] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1114–1127, 2008.

[14] M. Ivanovic and V. Kurbalija, "Time series analysis and possible applications," *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2016 - Proceedings*, pp. 473–479, 2016.

[15] T. Górecki and M. Łuczak, "Multivariate time series classification with parametric derivative dynamic time warping," *Expert Systems with Applications*, vol. 42, pp. 2305–2312, 4 2015.

[16] T. Warren Liao, "Clustering of time series data—a survey," *Pattern Recognition*, vol. 38, pp. 1857–1874, 11 2005.

[17] E. A. Maharaj, "Comparison and classification of stationary multivariate time series," *Pattern Recognition*, vol. 32, pp. 1129–1138, 1999.

[18] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, p. 40, 2010.

[19] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The VLDB Journal The International Journal on Very Large Data Bases*, vol. 8, pp. 237–253, 2 2000.

[20] Y. Cai and R. Ng, "Indexing spatio-temporal trajectories with Chebyshev polynomials," in *Proceedings of the 2004 ACM SIGMOD international conference on Management of data - SIGMOD '04*, (New York, New York, USA), p. 599, ACM Press, 2004.

[21] C. Cassisi, P. Montalto, M. Aliotta, A. Cannata, and A. Pulvirenti, "Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining," in *Advances in Data Mining Knowledge Discovery and Applications*, no. May 2014, InTech, 9 2012.

[22] S. Salvador and P. Chan, "FastDTW : Toward Accurate Dynamic Time Warping in Linear Time and Space," *Intelligent Data Analysis*, vol. 11, pp. 561–580, 2007.

[23] S. Seto, W. Zhang, and Y. Zhou, "Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition," in *2015 IEEE Symposium Series on Computational Intelligence*, pp. 1399–1406, IEEE, 12 2015.

[24] M. D. Robards, G. K. Silber, J. D. Adams, J. Arroyo, D. Lorenzini, K. Schwehr, and J. Amos, "Conservation science and policy applications of the marine vessel Automatic Identification System (AIS)-A review," *Bulletin of Marine Science*, vol. 92, no. 1, pp. 75–103, 2016.

[25] N. Marz and J. Warren, *Big Data, Principles and best practices of scalable real-time data systems*, vol. 37. Manning, 2015.

[26] C. Ray, R. Dréo, E. Camossi, and A.-L. Jousselme, "Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance," 2 2018.

[27] E. Tu, G. Zhang, L. Rachmawati, E. Rajabally, and G.-B. Huang, "Exploiting AIS Data for Intelligent Maritime Navigation: A Comprehensive Survey From Data to Methodology," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, pp. 1559–1582, 5 2018.

[28] R. Moussa, "Scalable Maritime Traffic Map Inference and Real-time Prediction of Vessels' Future Locations on Apache Spark," in *Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems - DEBS '18*, (New York, New York, USA), pp. 213–216, ACM Press, 2018.

[29] V. Roşca, E. Onica, P. Diac, and C. Amariei, "Predicting Destinations by Nearest Neighbor Search on Training Vessel Routes," in *Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems - DEBS '18*, (New York, New York, USA), pp. 224–225, ACM Press, 2018.

[30] J. Sadowski and A. Czapiewska, "Algorithms for Ship Movement Prediction for Location Data Compression," *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, vol. 9, no. 1, pp. 75–81, 2015.

[31] D. A. Kroodsma, J. Mayorga, T. Hochberg, N. A. Miller, K. Boerder, F. Ferretti, A. Wilson, B. Bergman, T. D. White, B. A. Block, P. Woods, B. Sullivan, C. Costello, and B. Worm, "Tracking the global footprint of fisheries," *Science*, vol. 359, pp. 904–908, 2 2018.

[32] E. N. De Souza, K. Boerder, S. Matwin, and B. Worm, "Improving fishing pattern detection from satellite AIS using data mining and machine learning," *PLoS ONE*, vol. 11, no. 7, pp. 1–20, 2016.

[33] F. Natale, M. Gibin, A. Alessandrini, M. Vespe, and A. Paulrud, "Mapping fishing effort through AIS data," *PLoS ONE*, vol. 10, no. 6, pp. 1–16, 2015.

[34] F. Mazzarella, M. Vespe, D. Damalas, and G. Osio, "Discovering Vessel Activities at Sea using AIS Data : Mapping of Fishing Footprints," *17th International Conference on Information Fusion (FUSION)*, pp. 1–7, 2014.

[35] N. A. Miller, A. Roan, T. Hochberg, J. Amos, and D. A. Kroodsma, "Identifying Global Patterns of Transshipment Behavior," *Frontiers in Marine Science*, vol. 5, no. July, pp. 1–9, 2018.

[36] J. Edlund, M. Grönkvist, A. Lingvall, and E. Sviestins, "Rule-based situation assessment for sea surveillance," in *Proceedings of SPIE - The International Society for Optical Engineering* (B. V. Dasarathy, ed.), vol. 6242, p. 624203, 4 2006.

[37] S. Boinepalli and A. J. Knight, "A Rule-based Track Anomaly Detection Algorithm," 2014.

[38] J. Will, L. Peel, and C. Claxton, "Fast Maritime Anomaly Detection using Kd-Tree Gaussian Processes," in *IMA Maths in Defence Conference*, 2011.

[39] C. Ray, R. Gallen, C. Iphar, A. Napoli, and A. Bouju, "DeAIS project: Detection of AIS spoofing and resulting risks," *MTS/IEEE OCEANS 2015 - Genova: Discovering Sustainable Ocean Energy for a New World*, pp. 0–5, 2015.

[40] W. Yan, R. Wen, A. N. Zhang, and D. Yang, "Vessel movement analysis and pattern discovery using density-based clustering approach," in *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3798–3806, IEEE, 12 2016.

[41] P. A. Silveira, A. P. Teixeira, and C. G. Soares, "Use of AIS data to characterise marine traffic patterns and ship collision risk off the coast of Portugal," *Journal of Navigation*, vol. 66, no. 6, pp. 879–898, 2013.

[42] T. Machado, R. Maia, P. Santos, and J. Ferreira, "Vessel Trajectories Outliers," in *Ambient Intelligence – Software and Applications –, 9th International Symposium on Ambient Intelligence*, ch. 29, p. 1–9, Springer, 2019.

[43] P. Borkowski, "The Ship Movement Trajectory Prediction Algorithm Using Navigational Data Fusion," *Sensors*, vol. 17, p. 1432, 6 2017.

[44] L. P. Perera, P. Oliveira, and C. Guedes Soares, "Maritime Traffic Monitoring Based on Vessel Detection, Tracking, State Estimation, and Trajectory Prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, pp. 1188–1200, 9 2012.