Master's Thesis
Computer Science
September 2012

Open Data for Anomaly Detection in Maritime Surveillance

Shahrooz Abghari
Samira Kazemi

School of Computing
Blekinge Institute of Technology
SE – 371 79 Karlskrona
Sweden

This thesis is submitted to the School of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:
Authors:
Shahrooz Abghari
E-mail: shahroozabghari@gmail.com

Samira Kazemi
E-mail: kazemi.samira@gmail.com

University advisors:
Dr. Henric Johnson,
School of Computing
Blekinge Institute of Technology

Dr. Niklas Lavesson,
School of Computing
Blekinge Institute of Technology

School of Computing
Blekinge Institute of Technology
SE – 371 79 Karlskrona Sweden

Internet : www.bth.se/com
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

Abstract

Context Maritime Surveillance (MS) has received increased attention from a civilian perspective in recent years. Anomaly detection (AD) is one of the many techniques available for improving the safety and security in the MS domain. Maritime authorities utilize various confidential data sources for monitoring the maritime activities; however, a paradigm shift on the Internet has created new sources of data for MS. These newly identified data sources, which provide publicly accessible data, are the open data sources. Taking advantage of the open data sources in addition to the traditional sources of data in the AD process will increase the accuracy of the MS systems.

Objectives The goal is to investigate the potential open data as a complementary resource for AD in the MS domain. To achieve this goal, the first step is to identify the applicable open data sources for AD. Then, a framework for AD based on the integration of open and closed data sources is proposed. Finally, according to the proposed framework, an AD system with the ability of using open data sources is developed and the accuracy of the system and the validity of its results are evaluated.

Methods In order to measure the system accuracy, an experiment is performed by means of a two stage random sampling on the vessel traffic data and the number of true/false positive and negative alarms in the system is verified. To evaluate the validity of the system results, the system is used for a period of time by the subject matter experts from the Swedish Coastguard. The experts check the detected anomalies against the available data at the Coastguard in order to obtain the number of true and false alarms.

Results The experimental outcomes indicate that the accuracy of the system is 99%. In addition, the Coastguard validation results show that among the evaluated anomalies, 64.47% are true alarms, 26.32% are false and 9.21% belong to the vessels that remain unchecked due to the lack of corresponding data in the Coastguard data sources.

Conclusions This thesis concludes that using open data as a complementary resource for detecting anomalous behavior in the MS domain is not only feasible but also will improve the efficiency of the surveillance systems by increasing the accuracy and covering some unseen aspects of maritime activities.

Keywords: open data, anomaly detection, maritime security, maritime domain awareness

Acknowledgment

First and foremost, we would like to thank our supervisors Dr. Niklas Lavesson and Dr. Henric Johnson for their patient guidance, encouragement and advice throughout this thesis.

We would also like to acknowledge the invaluable support of the Swedish Coastguard in this thesis. In particular, we are grateful to Peter Ryman, the law enforcement officer at the Coastguard, without whose knowledge and kind support this study would not have been successful.

Special thanks are due to Martin Boldt for all his help and support during the tough time of system installation.

Last but not least, we would like to express our heartfelt thanks to our beloved families, for their understanding and endless love through the duration of our studies.

TABLE OF CONTENTS

1. INTRODUCTION

MS is the effective understanding of all maritime activities that could impact the security, safety, economy or environment[1]. In recent years, the MS domain has received increased attention because of terrorism, smuggling activities and illegal immigration. An efficient MS system requires a complete Recognized Maritime Picture (RMP), which can be defined as a composite picture of maritime activities over an area of interest (Lefebvre & Helleur, 2001). For national maritime sovereignty, the RMP should include all activities within the 200 nautical miles Exclusive Economic Zone (EEZ). However, for some purposes such as the detection of illegal vessel transits, the RMP could extend beyond this region (Ponsford, D'Souza, & Kirubarajan, 2009). Using today's technology, continuous tracking of all maritime activities by a single sensor data is not sufficient since it cannot monitor everything that happens in the surveillance area. On the other hand, there are large amounts of data in the MS domain that are gathered from a variety of sensors, databases and information systems. Therefore, by taking advantage of all the available data sources it would be possible to obtain a complete RMP.

In addition to having a complete RMP, the way that the MS systems are used by human operators plays an important role in the efficiency of the surveillance operations. Monitoring vast sea areas and trying to establish Maritime Domain Awareness (MDA) for human operators is a difficult and time-consuming task (Riveiro, Falkman, & Ziemke, 2008a). This is due in part to the large amounts of heterogeneous data from multiple sources but also to the difficulties in detecting anomalous behavior from normal maritime activities. Therefore, having an automatic detector of unusual activities would help decision makers to efficiently monitor the ongoing activities in the surveillance area.

According to the Department of Homeland Security[2], AD is one of the enabling techniques for MDA. However, there are various AD techniques available and it is essential to choose the appropriate techniques that accomplish the MS goals. Data-driven AD approaches find the anomalous behavior by constructing a model from normal data and calculating the deviation from that model. However, relying only on data-driven approaches for surveillance systems is not sufficient due to the lack of user involvement in the detection process (Riveiro & Falkman, 2010). Furthermore, because of the diverse and complicated nature of the activities in the surveillance area, some of the suspicious behaviors are not directly observable. Therefore, finding all types of anomalies by using data-driven approaches seems to be impossible. On the other hand, maritime domain experts have the required knowledge and experience for finding maritime anomalies. Including the expert's knowledge about suspicious activities in the detection process can result in improved AD.

As well as the AD techniques, the use of different data sources will highly influence the detection of suspicious activities. Usually, only the data received from sensors are used for AD but there are a number of additional data sources regarding maritime activities that can be useful for this purpose. These sources of data consist of open and closed data about vessels, cargos, crew, etc. The closed data are only accessible to the maritime authorities, such as the Coastguard, but, by contrast, the open data are available online and freely accessible and reusable to the public. For instance, there are different organizations such as ports that publish their vessel traffic data or their facilities information online. In addition to the organizations, there are different online communities such as blogs, forums and social

[1] Integrating Maritime Surveillance, common information sharing environment (cise). Retrieved from *http://ec.europa.eu/maritimeaffairs/policy/integrated_maritime_surveillance/documents/integrating_maritime_surveillance_en.pdf*

[2] National plan to achieve maritime domain awareness for the national strategy for maritime security. Retrieved from *www.dhs.gov/xlibrary/assets/HSPD_MDAPlan.pdf*

networks which provide the possibility of sharing information about maritime events. Some of the advantages of open data in addition to the availability for the public and free accessibility are: first, open data can reveal some facts that are not reported to the maritime authorities or available in their databases and second, open data can be used in the global context and are not suffered from legitimate limitations of exchanging data between different countries. By applying the open data to the detection process, the AD can be done more wisely and the results can have more facts of interests for the maritime experts.

1.1 Problem Statement

In the maritime domain, there are different kinds of data sources that provide heterogeneous data regarding maritime activities. The majority of data, which are used by the MS systems, belong to the surveillance area of each country and are obtained from a variety of sensors and databases that are only accessible by the countries authorities. For detecting some of the anomalous activities such as smuggling, the maritime data beyond the surveillance area of each country are required. In order to assure security, maritime organizations in different countries need to exchange their privileged data and for this purpose they should deal with the diverse regulation of the data protection in each land. Exchanging data among countries is difficult, time-consuming and in some cases impossible because of the legislative issues. Moreover, there are activities that are neither reported to the maritime organizations, nor recorded in their data sources but they can be useful for the surveillance purpose. The publicly accessible and reusable data that are free from the legislative issues and revealing the unseen aspects of maritime activities are referred as open data. Consequently, employing the open data along with other confidential data sources would be beneficial for the MS systems to achieve their goals.

1.2 Research Questions

Given the context of available data sources and MS systems at the Swedish Coastguard and the types of anomalies that the subject matter experts at the Coastguard are interested in, the first research question is:

1. How accurate and valid are the results of an AD system that exploits open data as a complement to the available closed data?

In addition, by considering the accuracy as the degree to which the aforementioned AD system is able to distinguish between the normal and anomalous activities, and the validity as the degree to which the system results are true in real life, the next question is:

2. What is the performance difference between the system accuracy and the validity of results?

1.3 Aims and Objectives

The aim is to investigate the potential open data as a complementary resource for AD in the MS domain.

Objectives:

- Identify existing open data sources in the maritime domain.
- Identify those open data sources that are suitable for being used for AD in the MS domain.
- Propose a framework for AD in the MS domain based on using the open data.
- Develop an AD system based on the proposed framework and evaluate the accuracy of the system.
- Validate the implemented AD system in real life.

1.4 Contribution

This thesis contributes with a deeper understanding of open data as a complementary resource for effectively establishing the MS operations. It provides a framework for using the open data sources together with other sources of data for AD in the MS domain. According to the framework, an AD system is developed which employs a number of algorithms to implement the expert rules for detecting anomalies. The final contribution is the evaluation of the implemented AD system via the Coastguard validation in real life and also an experiment.

1.5 Outline

The remainder of this work is organized as follows: Section 2 reviews the background and related work regarding the open data, AD and data fusion in the MS domain. Section 3 presents the research methodology. Validity threats and verification are described in sections 4 and 5, respectively. Section 6 presents the experiment results and the validation results are shown in section 7. Section 8 features a detailed discussion about the obtained results. Finally, section 9 concludes the research with a discussion on the possible directions for future work.

2. BACKGROUND

The idea behind open data has been established for a long time. Open data can be used in a variety of domains and can be obtained from any resource. The two major sources of open data are the open data in science and the open data in government[1]. The longstanding concept of open data in science tries to overcome the difficulties in the current system of scientific publishing such as the inability to access data or usage limitation that is applied by the publishers or data providers (Molloy, 2011). Different groups, individuals and organizations are gathered to participate in a movement toward reforming the process of scientific publication (Molloy, 2011). One of the outcomes of the open data movement in science is the online availability of large number of scientific datasets for the public by different organizations. As well as the open data movement in science, governments for over a decade attempt to publish government data online and make it publicly accessible, readily available, understandable and usable (Alonso et al., 2009). Sharing the government data with the public provides openness and transparency with citizens. It can also improve the degree of participation in the society activities and the efficiency and effectiveness of the government services and the operations within and between the governments (Dietrich et al., 2009).

According to the estimation by Dedijer and Jéquier (1987), 90% of all information is open source, 9% is grey information (preprints of scientific articles, rumors in business circles, project proposals submitted to a research-funding agency, discussions with well-informed specialists, etc.), 0.9% is secret and 0.1% is non-existent information (i.e. the information you have, but you are not aware of it). Considering the large ratio of the open data sources, there should be a great value in using them in different domains. In the MS systems, the majority of the exploited data are obtained from the confidential sources. However, in recent years the new concept of the Web, which takes the network as a platform for information sharing, interoperability and collaboration, has created new sources of data for MS. There are organizations and communities that provide their maritime related data online and make them accessible for the public. Therefore, it would be beneficial for the MS

[1] Open definition. Retrieved from *opendefinition.org/okd/*

systems if they can take advantage of the open data to increase the safety and security in their surveillance area.

2.1 Terminology

According to the Department of Homeland Security[1], MDA would be achieved by monitoring the maritime activities, fusing and analyzing the data in a way that normal activities can be identified and anomalies differentiated. Therefore, AD and *data fusion* techniques are important technologies for MDA.

Data fusion involves the process of combining data from multiple sources or sensors and making inferences that may not be possible from a single source or sensor (Hall & McMullen, 2004).

AD is widely used in the areas such as video surveillance, network security and military surveillance. Chandola, Banerjee and Kumar (2009) define *AD* as:

The problem of finding patterns in data that do not conform to expected behavior.

Depending on the domain of study, the non-conforming patterns are called by different names such as anomalies, outliers, exceptions, etc. In the MS domain, these non-conforming patterns are referred as anomalies. Defense R&D Canada (Roy, 2008) provides the following definition for the term *anomaly* in the context of the MS domain:

Something peculiar (odd, curious, weird, bizarre, atypical) because it is inconsistent with or deviating from what is usual, normal, or expected, or because it is not conforming to rules, laws or customs.

The term *Open Data* refers to the idea of making data freely available to use, reuse or redistribute without any restriction. The open data movement follows the other open movements such as *Open Access* and *Open Source*. According to the Open Knowledge Foundation[2], a community based organization that promote open knowledge (whether it is content, data or information-based), an open work should be available as a whole, with a reasonable reproduction cost, preferably downloading via the Internet without any charge and in a convenient and modifiable form. Furthermore, it should be possible to modify and distribute the work without any discrimination against persons, groups, fields or endeavor. In the scope of this thesis, the open data term refers to the publicly available data that may or may not require free registration.

2.2 JDL Data Fusion Model

One of the most widely used data fusion models in the literature is the JDL data fusion model. It was developed by the US Joint Directors of Laboratories (JDL) data fusion sub-panel in 1985. A current version of the model consists of six different levels. Table 1 shows the description of these six levels (Hall & McMullen, 2004).

In this work the main focus is on AD which is done in the Level 2 (situation assessment) of the JDL model. The input of this level is the identified entities and their related information regarding each other or the environment and the output would be the assessment of a situation as normal or anomalous (Brax, Niklasson, & Smedberg, 2008).

[1] National plan to achieve maritime domain awareness for the national strategy for maritime security. Retrieved from *www.dhs.gov/xlibrary/assets/HSPD_MDAPlan.pdf*
[2] Open definition. Retrieved from *opendefinition.org/okd/*

Table 1

Summary of the JDL Data Fusion Model Components

JDL model component	Description
Level 0 processing (Source preprocessing)	At this level, preprocessing of data from sensors and databases would be done by means of image processing, signal processing and conditioning, unit conversions, bias corrections or feature extractions, etc.
Level 1 processing (Object refinement)	This level focuses on combining data from sensors and databases in order to obtain the most accurate and reliable estimates of an entity's position, movement, attributes, characteristics and identity.
Level 2 processing (Situation assessment)	According to the obtained result of previous level, a description of current relationships among entities and their relationship to the environment would be developed in order to determine the interpretation of the situation.
Level 3 processing (Impact assessment)	This level focuses on the estimation and prediction of alternative futures and hypotheses concerning the current situation to determine the potential impacts or threats.
Level 4 processing (Process refinement)	This level is a meta-process that monitors the whole data fusion process to optimize the utilization of data sources and algorithms and improve the performance of the ongoing data fusion.
Level 5 processing (Cognitive refinement)	Level 5 focuses on transforming the result of data fusion in to the displays and understandable information for the user and improvement of human/computer effectiveness.

2.3 Related Work

Data fusion techniques have been used for a long time in the MS domain. The majority of studies focused on target tracking, tactical situation awareness and threat assessment (Akselrod, Tharmarasa, Kirubarajan, Zhen Ding, & Ponsford, 2009; Bick & Barock, 2005; Danu, Sinha, Kirubarajan, Farooq, & Brookes, 2007; Di Lallo et al., 2006; Gad, 2009; Giompapa, Farina, Gini, Graziano, & Di Stefano, 2007; Giompapa et al., 2008; Hatch, Kaina, Mahler, & Myre, 1998; Henrich, Kausch, & Opitz, 2004; Jouan, Valin, Gagnon, & Bosse, 1999; Lefebvre & Helleur, 2001; Maresca et al., 2010; Vespe, Sciotti, & Battistello, 2008). Typically, in these works the combination of two or more sensors such as: Automatic Identification System (AIS), Infrared (IR), video camera, Synthetic Aperture Radar (SAR), Vessel Traffic Service (VTS) radar, Over The Horizon (OTH) radar, High Frequency Surface Wave Radar (HFSWR) and microwave radar is used and the surveillance area is limited to the coastal regions.

In recent years, the number of studies that address the use of AD in the MS domain is increasingly growing. AD techniques are divided into two groups, namely data-driven and knowledge-driven approaches. There are a couple of works that proposed knowledge-based systems with different representation techniques and reasoning paradigms such as rule-based, description logic and case-based reasoning (Guyard, Roy, & Defence R&D Canada-Valcartier, 2009; Nilsson, van Laere, Ziemke, & Edlund, 2008; Roy & Davenport, 2010). A prototype for a rule-based expert system based on the maritime domain ontologies was developed (Edlund, Gronkvist, Lingvall, & Sviestins, 2006) that could detect some of the anomalies regarding the spatial and kinematic relation between objects such as simple scenarios for hijacking, piloting and smuggling. Another rule-based prototype was developed

by Defense R&D Canada (Roy, 2008, 2010). The aforementioned prototype employed various maritime situational facts about both the kinematic and static data in the domain to make a rule-based automated reasoning engine for finding anomalies. One of the popular data-driven AD approaches is the Bayesian network (Fooladvandi, Brax, Gustavsson, & Fredin, 2009; Johansson & Falkman, 2007; Lane, Nevell, Hayward, & Beaney, 2010). Johansson and Falkman (2007) used the kinematic data for creating the network; however, in the work that was done by Fooladvandi et al. (2009) expert's knowledge as well as the kinematic data was utilized in the detection process. Moreover, Lane et al. (2010) presented five unusual vessel behaviors and the way of formulating them in an AD system that the estimation of the overall threat was performed by using a Bayesian network. Unsupervised learning techniques have been widely used for data-driven AD such as Trajectory Clustering (Dahlbom & Niklasson, 2007), Self Organizing Map (Riveiro, Johansson, Falkman, & Ziemke, 2008) and fuzzy ARTMAP neural network (Rhodes, Bomberger, Seibert, & Waxmamn, 2005). Some statistical approaches, such as Gaussian mixture model (Laxhammar, 2008), hidden Markov model (Andersson & Johansson, 2010), adaptive kernel density estimator (Ristic, La Scala, Morelande, & Gordon, 2008) and precise/imprecise state-based anomaly detection (Dahlbom & Niklasson, 2007) have been used in this context. The majority of the works that have been done in the context of AD only used the AIS data.

There are a number of studies that employed data fusion techniques to fuse data from different sensors in AD systems (Carthel et al., 2007; Guerriero, Willett, Coraluppi, & Carthel, 2008; Rhodes, Bomberger, Seibert, & Waxman, 2006; Vespe, Sciotti, Burro, Battistello, & Sorge, 2008). In these studies, the surveillance area was restricted to the coastal regions and the combination of data from AIS, SAR, IR, video and radar was used in the fusion process to obtain the vessel tracks. Furthermore, there are some other works that focused on the fusion of both sensor and non-sensor data (Andler et al., 2009; Ding, Kannappan, Benameur, Kirubarajan, & Farooq, 2003; Fooladvandi et al., 2009; Lefebvre & Helleur, 2004; Mano, Georgé, & Gleizes, 2010; Riveiro & Falkman, 2009). For example, Lefebvre and Helleur (2004) and Riveiro and Falkman (2009) treated the expert's knowledge as the non-sensor data. Riveiro and Falkman (2009) introduced a normal model of vessel behavior based on AIS data by using self organizing map and a Gaussian mixture model. According to the model, the expert's knowledge about the common characteristic of the maritime traffic was captured as IF-THEN rules and the AD procedure was supposed to find the deviation from the expected value in the data. Lefebvre and Helleur (2004) fused radar data with user's knowledge about the vessels of interests. The sensor data were modeled as track and the non-sensor data were modeled as templates. The track-template association was done by defining mathematical models for tracks and using fuzzy membership functions for association possibilities. Mano et al. (2010) proposed a prototype for the MS system that could collect data from different types of sensors and databases and regroup them for each vessel. Sensors like AIS, HFSWR and classical radars and databases such as environmental database, Lloyd's Insurance and TF2000 Vessel DB were included in this prototype. By using multi-agent technology an agent was assigned to each vessel and anomalies could be detected by employing a rule-based inference engine. When the combination of anomalies exceeded a threshold, vessel status was informed to the user as an anomaly. The work presented by Ding, Kannappan, Benameur, Kirubarajan and Farooq (2003), proposed the architecture of a centralized integrated maritime surveillance system for the Canadian coasts. Sensors and databases included in this architecture were: HFSWR, ADS (Automatic Dependant Surveillance) reports, visual reports, information sources, microwave radar and radar sat. A common data structure was defined for storing data that were collected from different sensors. Andler et al. (2009) also described a conceptual MS system that integrated all available information such as databases and sensor systems (AIS, LRIT, intelligence reports, registers/databases of vessels, harbors, and crews) to help user to detect and visualize anomalies in the vessel traffic data in a worldwide scale. Furthermore, the authors suggested using open data in addition to other resources in the fusion process.

In conclusion, the main focus of the projects that have been done in the context of AD in the MS domain was related to using sensors data and mainly the AIS data to find the anomalies in the coastal region. Detection of some suspicious activities such as smuggling requires vessel traffic data beyond the coastal region. Maritime authorities in each country have overall information of maritime activities in their surveillance area. But exchanging information among different countries is a complicated procedure because of the diverse regulation of data protection in each land. Therefore, using data sources that are free from legislative procedures can be a good solution for providing information that belongs to the regions outside the land territory. Furthermore, all the information about maritime activities is not recorded in the authorities' databases or reported to them. On the other hand, there are numerous open data sources consists of different websites, blogs and social networks that can be useful for observing the hidden aspects of maritime activities. Hence, this thesis will investigate the potential open data sources for maritime activities and exploit them to build an AD system.

3. RESEARCH METHODOLOGY

This thesis starts by investigating the applicable open data sources for AD in the MS domain. Then, potential maritime anomalies that can be detected by the obtained open data sources and the way that the open data can be applied to the AD process are identified. The next step is to design and implement an anomaly detector system that can use open data in the detection process. For this purpose, a general framework for AD based on using both open and closed data sources in the MS domain is proposed. Then, an anomaly detector system is developed according to the proposed framework. In order to understand to what extent the system results are valid, the system is evaluated in real life by the subject matter experts from the Swedish Coastguard. However, the system accuracy should be evaluated before using the system in real life. Therefore, an experiment is conducted to measure the system accuracy. The following sections describe these steps in detail.

3.1 Open Data in the MS Domain

Applicable data sources for AD in the MS domain can be divided into three categories. The first and main category consists of sensors. Sensors provide kinematic data for each object in their coverage area and can be categorized as passive and active. Active sensors do not require cooperation from objects and collect the data by active probing the environment such as radar and sonar (A. M. Ponsford et al., 2009). However, passive sensors rely on the data that are broadcasted by objects intentionally such as AIS or unintentionally such as Electronic Intelligence (ELINT) and SIGnal INTelligence (SIGINT) systems (A. M. Ponsford et al., 2009). More information about the main maritime sensors can be found in (İnce, Topuz, & Panayirci, 1999; Vespe, Sciotti, & Battistello, 2008). The second category of data sources includes the authorized databases which contain information about vessels, cargos, crew, etc[1]. The first and second categories are only accessible to the maritime authorities such as the Coastguard and can be referred as closed data sources. The third category belongs to the open data sources which are publicly available via the Internet and are free to access or reuse. These data sources consists of vessels traffic data and reports or news that are related to the maritime domain and can be found in different blogs, websites or social networks.

[1] Integrated maritime policy for the EU : Working document III on maritime surveillance systems. Retrieved from *epub.sub.uni-hamburg.de/epub/volltexte/2009/1750/pdf/maritime_surveillance_en.pdf*

To obtain the applicable open data for AD, the first step is started by looking through the information resources document[1] provided by the International Maritime Organization (IMO). IMO is the United Nations specialized agency with responsibility for the safety and security of shipping and the prevention of marine pollution by vessels. The document introduces 29 governmental and intergovernmental organizations that work in different fields related to the MS domain such as maritime safety, prevention of pollution from vessels, liability and insurance issues, shipping information, etc. All these 29 organizations' websites and the links provided by each of them are investigated and a list of online data sources is prepared. The prepared list contains data sources that have information that can be qualified for the MS purpose such as information related to AIS, vessel characteristics, ports, maritime companies, suppliers' information, weather, etc. Table A.1 of Appendix A, presents the obtained data sources. These sources of data are available online but having access to some of them needs non-free registration. Moreover, in the process of finding open data sources, it is attempted to obtain sources of data that are related to the Baltic region and mostly Sweden by use of the previously observed data sources and also Google search engine.

3.2 Identification of Maritime Anomalies

An important aspect of the literature review is to find out more about the potential maritime anomalies. The main sources of information about maritime anomalies are reports of the two workshops that were held in Canada (Roy, 2008) and Sweden (Andler et al., 2009; van Laere & Nilsson, 2009). In these two workshops attendees were experts in the maritime domain and a variety of maritime anomalies were identified. The outcome of the Swedish workshop was the identification of the 31 most desired anomalous behaviors for the Swedish stakeholders. The identified anomalies belong to the following categories: tampering, owner/crew, history, rendezvous (object, location), movement and cargo. The outcome of the Canadian workshop was a taxonomy of maritime anomalies that categorized anomalies as static and dynamic. Static anomalies were related to the vessel characteristics such as name, IMO number, etc. Dynamic anomalies were divided into two groups, kinematic and non-kinematic anomalies. The kinematic anomalies included anomalies related to location, course, speed, reporting and maneuver of vessels. The non-kinematic anomalies were related to passengers, crew list, cargo list, last port of call and next port of call.

According to the identified anomalies by the two workshops, a list of some potential maritime anomalies that can be detected by use of the available open data sources is prepared. Then, in a meeting with representatives of the Swedish Coastguard the types of anomalies that are interesting for them and the possibility of using open data for AD are discussed. During the meeting the prepared list of anomalies is presented and they are asked about the possibility of occurrence and their degree of interest for each anomaly. As an outcome of the meeting a number of scenarios are created and based on them 11 rules are defined. The first scenario refers to the anomalies related to the vessel static information such as name, owner, IMO number, dimensions, type and the status (in service or laid up). For example, sailing a vessel with a draught of 22 meters over an area with a 9 meters depth or observing a vessel that should be laid up or changing the name or the owner of a vessel during its voyage indicate the existence of suspicious activities. The second scenario is related to the prior arrival notification for vessels. Vessels should inform their arrival time to the ports at least 24 hours in advance. Each port also provides an online timetable for the incoming vessels. Therefore, any mismatch between the reported AIS data regarding the destination or the arrival time of a vessel and the destination port timetable needs to be checked by coastguards. The third scenario is related to ordering pilots. Usually, large vessels because of their size and weight need to be guided by pilots through dangerous and congested waters. Therefore, vessels need to submit their request for a pilot and also inform

[1] Information resources on maritime security and ISPS code. Retrieved from *www.imo.org/knowledgecentre/informationresourcesoncurrenttopics/maritimesecurityandispscode/documents/information%20resources%20on%20maritime%20security%20and%20isps%20code.pdf*

the destination port. However, in some cases vessels order a pilot without informing the port. Such situations should be investigated.

In the next meeting, the scenarios and the rules are presented to the representatives of the Swedish Coastguard and they are asked to comment or suggest new scenarios or rules. By getting the final approval from the Coastguard experts, one new rule (rule number 5) is added to the list. Table 2 shows the admitted rules by the experts. These identified maritime anomalies can be detected by use of AIS data, vessel traffic timetables in ports and pilots websites and the vessel characteristic data that are available in data sources such as Lloyd's. A name is given to each anomaly that can be detected by the rules, and for the rest of this thesis anomalies will be referred by their names.

Table 2

The identified anomalies that can be detected by open data (confirmed by the Swedish Coastguard)

No.	Expert rules	Anomaly
1	If a vessel destination does not exist in the port schedule then anomaly.	VESSEL_NOT_INFORMED_PORT (A1)
2	If a vessel ETA does not match with the port ETA for the vessel then anomaly.	ARRIVAL_TIME_MISMATCHED (A2)
3	If a vessel entered a port without informing the port then anomaly.	VESSEL_ENTERED_PORT_WITHOUT_ NOTICE (A3)
4	If a vessel has requested a pilot but has not used the service then anomaly.	VESSEL_NOT_USED_PILOT (A4)
5	If vessel A which normally travels between ports X and Y, suddenly goes to port Z then anomaly.	UNUSUAL_TRIP_PATTERN (A5)
6	If a vessel has not left a port according to the port schedule then anomaly.	VESSEL_NOT_LEFT_PORT (A6)
7	If a vessel exists in a port schedule but it has not entered the port then anomaly.	VESSEL_NOT_ENTERED_PORT (A7)
8	If a vessel does not exist in the port schedule and the vessel has requested a pilot then anomaly.	VESSEL_ORDERED_PILOT_AND_ NOT_INFORMED_PORT (A8)
9	If a vessel has moored in a port and has been observed somewhere else then anomaly.	VESSEL_MOORED_IN_PORT (A9)
10	If vessel A has not entered a port according to the port schedule instead vessel B enters the port at the same time slot then anomaly.	WRONG_VESSEL_ENTERED (A10)
11	If a vessel with the laid up status has been observed somewhere else then anomaly.	VESSEL_LAID_UP (A11)

Note. ETA = estimated time of arrival

3.3 Framework Design

The Open Data Anomaly Detection System (ODADS) is designed for traffic monitoring and detecting anomalies in the MS domain by using open and closed data sources. Figure 1 depicts the ODADS architecture. ODADS consists of three modules: 1) *Data Collector*, 2) *Anomaly Detector* and 3) *Display Client*. The Data Collector module is responsible for collecting open data from the Internet, preprocessing and storing the data in a database. The data can be related to vessel traffic (such as AIS reports, ports and pilots timetables), vessel characteristics, ports equipments and facilities, companies that are involved in maritime activities, news or reports about maritime events and activities available in different social media platforms (such as blogs and social networks), etc. The Data Store comprises a set of databases that contain data belong to different types of sensors, authorized databases and open data sources. The data in the Data Store can be fused or integrated before being used in the detection process. When the Data Collector completes its task, Anomaly Detector starts to work. The Anomaly Detector module analyzes the available data (open and closed data) and detects possible anomalies by utilizing both knowledge-driven and data-driven techniques. Different AD techniques are employed due to the distinct nature of anomalies and the complexity of the environment in the MS domain. Previously known anomalies can be detected by knowledge based techniques such as rule-based, but in real life an AD system must be able to detect the unseen anomalies, too. This is one of the benefits of using data-driven methods such as machine learning techniques. Therefore, detecting different types of anomalies seems to be possible by exploiting different techniques. The Display Client module is the user interface of the system. This module represents the cognitive refinement level (Level 5) of the JDL model. M. J. Hall, Hall and Tate (2000), argued that the effectiveness of a system can be affected by the way that the system produced information is comprehended by the human user. The cognitive refinement process involves traditional Human-Computer Interaction (HCI) utilities such as geographical display or advanced methods that support functionalities such as cognitive aids, negative reasoning enhancement, focus/defocus of attention and representing uncertainty. Section 3.4 describes the implementation details for each module.

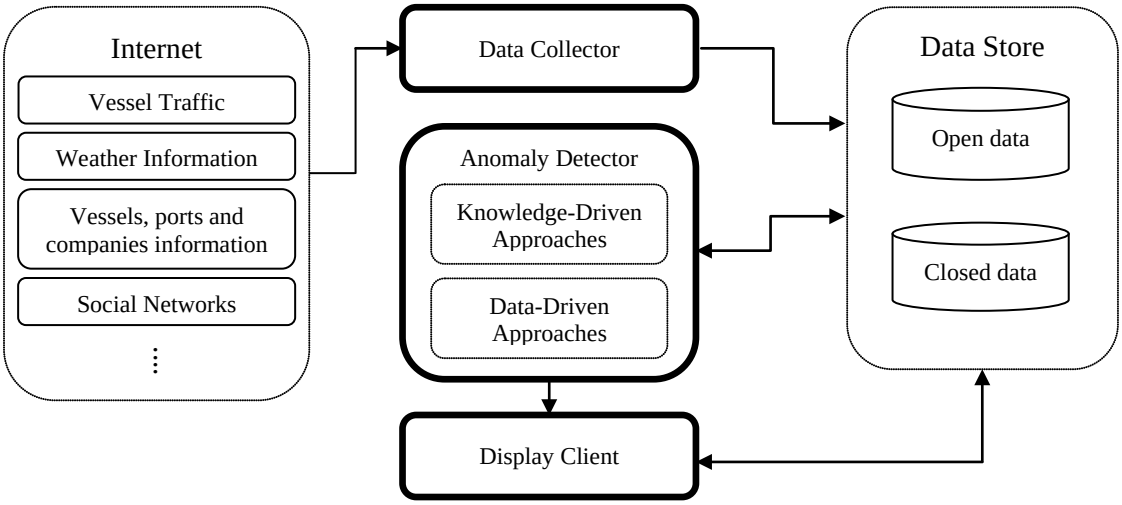

Figure 1. The Open Data Anomaly Detection System (ODADS) architecture. The Data Collector module collects data from the Internet and stores them in the database. The Anomaly Detector module detects anomalies by taking advantage of different techniques. The Display Client module displays the detected anomalies to the user and enables system-user interaction.

3.4 Implementation

ODADS is implemented by taking advantage of the identified maritime anomalies and the obtained open data sources that were discussed in the previous sections. To limit the scope, first of all only four types of vessels: passenger, ferry, cargo and tanker are considered and other types of vessels such as fishing and sailing vessels, which ports usually do not provide any information about them, are omitted. Secondly, the rule related to the vessel static information is ignored. Furthermore, the WRONG_VESSEL_ENTERED anomaly is excluded due to its complexity. As well as the A1-A9 anomalies, in further collaboration with the Coastguard representatives during the implementation phase, a new type of anomaly is proposed. This anomaly is called UNDER_SURVEILLANCE_VESSEL and occurs when a vessel of interest has any of the A1-A9 anomalies and the vessel exists in the vessels blacklist.

3.4.1 Data Description

The required vessel traffic data can be obtained from AIS reports and ports and pilots timetables. The surveillance area is restricted to the north of the Baltic Sea and a part of the Gulf of Finland, the regional area between three European countries Sweden, Finland and Estonia. Figure 2 shows the surveillance area where the geographic coordinates lie between latitudes 58.49º - 60.24º N and longitudes16.19º – 25.00º E. This region is one of the high-traffic regions in the Baltic Sea and is surrounded by the four highly used ports. More information regarding each port can be found in Appendix B.

Figure 2. The area of interest is restricted to the north of the Baltic Sea and a part of the Gulf of Finland. Ports from left to right are Norrköping, Stockholm group (Nynäshamn, Stockholm and Kapellskär), Helsinki and Tallinn (The image is adapted from Google Earth).

3.4.1.1 AIS data

Due to inaccessibility to the raw AIS data in this thesis, the AIS reports that are provided by the *MarineTraffic.com* website[1] are exploited. These reports consist of both static and dynamic types of data for each vessel during its voyage. Vessel static data include name, type, built year, size, draught, flag, call sign, Maritime Mobile Service Identity (MMSI), IMO identification number, origin (last known port), destination, Estimated Time of Arrival (ETA) and dynamic data are related to speed (max and average), position (longitude and latitude), Course Over Ground (COG), heading.

[1] *www.marinetraffic2.aegean.gr/ais/getkml.aspx*

3.4.1.2 Ports and Pilots data

Ports timetables are obtained from the websites of the high-traffic ports in the area of interest: Stockholm group (Stockholm, Kapellskär and Nynäshamn) and Norrkoping ports in Sweden, Helsinki port in Finland and Tallinn port in Estonia. The pilots timetables belong to the vessel traffic in Stockholm pilotage area in Sweden. Table A.2 of Appendix A, presents each data source in detail.

3.4.2 **Data Collector Module**

The Data Collector module is responsible for extracting data from different open data sources via the Internet, data preprocessing and data storage. This module consists of two parts. The first part is the *AIS Logger*, a shell script program, which downloads and extracts the processed vessel AIS reports and tracks[1] data from *MarineTraffic.com* website. Since the *MarineTraffic.com* website updates the data at 10-minute intervals, data extraction is done every 10 minutes and the collected data are stored in a file. When the file is prepared, the second part of the module, the *Web Scraper*, starts to work. It is a Java program that harvests data from different websites and stores them into the database. These data include ports and pilots timetables from their websites and the vessels details information from the M*arinetraffic.com.* Figure 3 illustrates the different parts of the Data Collector module.

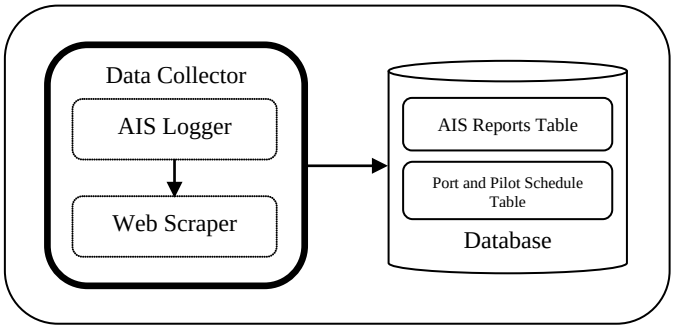

Figure 3. The Data collector module consists of two parts, *AIS Logger* that provides a list of vessels in the area of interest and the *Web Scraper*, a program which gathers vessels details information and ports and pilots timetables from the Internet.

3.4.2.1 Preprocessing

Data preprocessing can be performed by a number of techniques such as data cleaning to remove noise, data integration to merge multiple data sources into an understandable data, data transformation such as normalization to improve the accuracy and data reduction to eliminate the redundant features (Han et al., 2011).

In the first step of preprocessing, text values such as vessel name, origin, destination and company name are transformed to the same format (lowercase characters) and all special characters and whitespaces are removed. Since the data are collected from different sources in different countries, time values are converted to Central European Time (CET). Every data source has its own format of data representation. For instance, vessel arrival time in ports can be stored as two separated parts of date and time, a combination of date and time values or just as date. Moreover, some of the ports provide additional vessel information such as weight and length as well as the timetables. This diversity makes some parts of the provided data remain unused. Therefore, a common data representation format for ports and pilots data is defined that contains vessel name, vessel type, origin, destination, company name, vessel status and arrival/departure time.

[1] *www.marinetraffic.com/ais/gettrackxml.aspx?mmsi=xxxxxxxx*

3.4.3 **Database**

The Database contains four tables to store the extracted open data: 1) *AIS Reports table*, 2) *Ports and Pilots Schedule table*, 3) *Vessel Trip History table* and 4) *Detected Anomaly table*. There is also a table for storing vessels that are kept under surveillance because of their previous involvement in the criminal activities. The under surveillance vessels data are provided by the Coastguard officer via the Display Client module and stored in the *Blacklist Table*. The *AIS Reports table* contains the processed AIS data such as vessel name, type, flag, origin, destination, etc. The *Ports and Pilots Schedule table* is used for storing the vessel data including arrival and departure time gathered from ports and pilots data sources. In the *Vessel Trip History table*, data related to the frequency of vessels trips between different ports are stored and the *Detected Anomaly* table contains the history of all types of detected anomalies.

3.4.4 **Anomaly Detector Module**

The Anomaly Detector module is the main part of the system which employs different techniques to detect anomalies. After investigating the nature of the anomalies and the potential techniques, it is determined that except for the UNUSUAL_TRIP_PATTERN anomaly, the other types can be detected by exploiting search techniques. Detection of the UNUSUAL_TRIP_PATTERN anomaly requires data-driven approaches such as machine learning or statistical techniques and also a history of the vessel traffic data for training the system. In case of using machine learning techniques, the system will detect the anomaly by learning the normal vessel trip pattern and highlighting the unusual trip and by using statistical approaches the system will make decision based on the frequency or the possibility of a vessel trip between two different ports. The next section describes the detection algorithm for each anomaly in detail. This module is depicted in Figure 4.

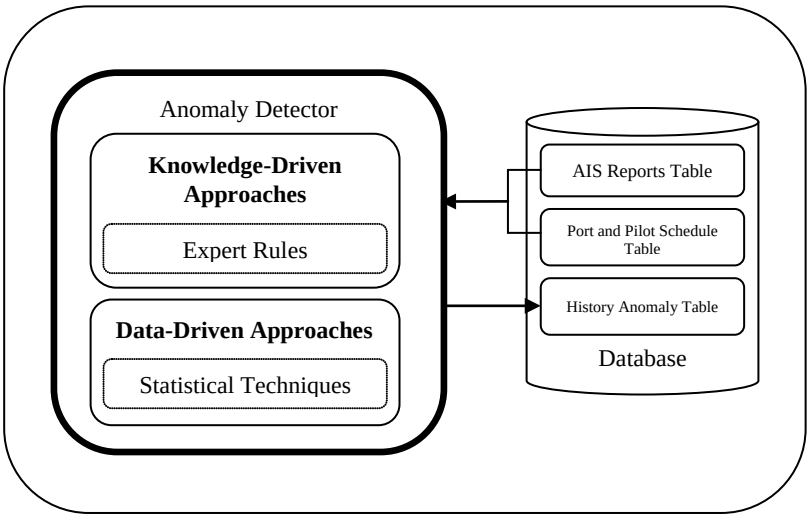

Figure 4. The Anomaly Detector module

3.4.4.1 Algorithms

To define the appropriate detection algorithms, the first step is to specify what types of data are required for detecting each anomaly (Table A.3 of Appendix A). Detection of each individual anomaly (except the UNUSUAL_TRIP_PATTERN anomaly) can be done by performing a search in the specified data for finding the desired match. If the match is not found then the vessel would be marked as an anomaly. Appendix C gives a detailed description about each algorithm. Algorithm 1 shows the main procedure of the Anomaly Detector module. Algorithms 2-4 and 6-8 present the described search process in detail. For the UNUSUAL_TRIP_PATTERN anomaly, data related to six months (September 15, 2011-March 15, 2012) of vessel traffic in the surveillance area are gathered. By monitoring the

activities during this period, the system will be able to find the normal pattern of vessels trips in the area of interest. For detecting this anomaly a simple statistical approach is used. A look up table is created and for each vessel the number of times that the vessel travels between two different ports is stored. For each vessel, if the frequency of travelling between its origin and destination is less than a predefined threshold then the vessel will be reported as an anomaly. Algorithms 5 and 16 describe the process in detail.

After executing all the algorithms, the final type of anomaly for each vessel should be determined. There are some situations that multiple anomalies can occur in the same time for a specific vessel. These situations are presented in Table 5. For the combination of anomalies that do not have any feature in common new types of anomalies are defined.

Table 5

The combined anomalies

Anomaly	Condition	Anomaly Type
A1,A5	Vessel did not inform its arrival to the destination port and its trip to the port is not common.	UNUSUAL_TRIP_PATTERN_AND_ VESSEL_NOT_INFORMED_PORT
A1,A8	Vessel did not inform its arrival to the destination port and it ordered a pilot.	VESSEL_ORDERED_PILOT_AND_ NOT_INFORMED_PORT
A1,A5,A8	Vessel did not inform its arrival to the destination port, it ordered a pilot and its trip to the port is not common.	UNUSUAL_TRIP_PATTERN_AND_ VESSEL_ORDERED_PILOT_AND_ NOT_INFORMED_PORT_ INFORMED_PORT
A2,A4	Vessel has delay and it ordered a pilot but did not use it.	VESSEL_ARRIVAL_TIME_ MISMATCHED_AND_VESSEL_NOT_ USED_PILOT
A2,A5	Vessel has delay and its trip to the port is not common.	UNUSUAL_TRIP_PATTERN_AND_ VESSEL_ARRIVAL_TIME_ MISMATCHED
A2,A4,A5	Vessel has delay and it ordered a pilot but did not use it and its trip to the port is not common.	UNUSUAL_TRIP_PATTERN_AND_ VESSEL_NOT_USED_PILOT_AND_ VESSEL_ARRIVAL_TIME_ MISMATCHED
A3,A6	Vessel entered a port without prior notification and has not left the port according to the plan.	VESSEL_ENTERED_PORT_ WITHOUT_NOTICE_AND_NOT_ LEFT_PORT_ON_TIME
A4,A5	Vessel ordered a pilot and informed to the port however it did not used the pilot and its trip to the port is not common.	UNUSUAL_TRIP_PATTERN_AND_ VESSEL_NOT_USED_PILOT
A4,A7	Vessel ordered a pilot but did not use it and it has not entered the port.	VESSEL_NOT_ENTERED_PORT_ AND_NOT_USED_PILOT
A4,A8	Vessel ordered a pilot but did not inform the port and it did not use the pilot.	VESSEL_ORDERED_PILOT_AND_ NOT_INFORMED_PORT_AND_ VESSEL_NOT_USED_PILOT

(Continued)

Anomaly	Condition	Anomaly Type
A5,A7	Vessel trip to the destination port is not common and it has not entered the port.	UNUSUAL_TRIP_PATTERN_AND_ VESSEL_NOT_ENTERED_PORT
A7,A4,A5	Vessel ordered a pilot but did not use it and has an unusual trip and it has not entered the port.	UNUSUAL_TRIP_PATTERN_AND_ VESSEL_NOT_ENTERED_PORT_ AND_VESSEL_NOT_USED_PILOT
A8,A5	Vessel ordered a pilot but did not inform the port and the trip to the destination port is not common.	UNUSUAL_TRIP_PATTERN_AND_ VESSEL_ORDERED_PILOT_AND_ NOT_INFORMED_PORT
A9,[A1-A8]	Vessel uses the information that belongs to a moored vessel in a port; since the vessel provides unreal information, it is normal that other anomalies are set for this vessel.	VESSEL_MOORED_IN_PORT
A8,A4,A5	Vessel ordered a pilot but did not inform the port and the trip to the destination port is not common and the vessel also did not use the pilot.	UNUSUAL_TRIP_PATTERN_AND_ VESSEL_ORDERED_PILOT_AND_ NOT_INFORMED_PORT_AND_ VESSEL_NOT_USED_PILOT

3.4.4.2 String matching techniques

Using exact string matching techniques for comparing vessel information from different data sources is inapplicable due to the potential errors that might occur because of different notations or even human operator mistakes during data entry. Therefore, a metric should be used for measuring the degree of similarity between two vessels from different sources. For this purpose, some string comparison methods are evaluated such as *Damerau-Levenshtein* (*D-L*) also known as *edit distance* (Damerau, 1964; Levenshtein, 1966), the *n-gram* technique (P. A. V. Hall & Dowling, 1980), *Jaro* (Jaro, 1972, 1989) and *JaroWinkler* (Winkler, 1990). *D-L* or *edit distance* is equal to the minimum number of edits (substitution, delete, insert and transposition) required to change one string to the other. The use of *n-gram* for string matching is performed by first constructing the n-grams (contiguous sequences of n items) for each string and then comparing the n-grams of both strings in order to find the number of common n-grams. *Jaro* and *JaroWinkler* (a variant of *Jaro*) measure the number and order of common characters in two strings and also the number of transposition that is needed to change one of the strings to the other. *Jaro* metrics are more similar to the human decision making compared to the *D-L* distance (Denk & Hackl, 2003).

To evaluate these metrics, the *D-L* distance is implemented and for the other metrics *SimMetrics*[1], an open source library that contains a set of similarity metrics, is used. Regarding the *n-gram* technique, choosing the length of n-grams is an important factor since it can affect the processing time and accuracy of the method. According to Salton and McGill (1986) and Zamora, Pollock, and Zamora (1981), trigram is one of the *n-gram* methods (when n=3) that can achieve the best results in retrieving similar words.

The collected data related to one day of vessel traffic in the area of interest are taken with distinct vessels names, origins and destinations and stored in separate groups. Then, the string matching methods are applied to each group by considering all pair wise combination of the elements in each group. To evaluate the results, the obtained similarities are sorted in ascending order and it is verified whether or not the values in each pair can be considered similar by a human operator. Two values are considered equal if they are exactly the same or a bit different because of misspellings or shortening. Finally, based on the results it is

[1] *www.aktors.org/technologies/simmetrics/index.html*

concluded that for the available data, *JaroWinkler* has the best performance among all the other metrics and its results are closer to reality. *Jaro* and *D-L* are the second and the third best respectively and *trigram* is the worst.

According to the degrees of similarity that are obtained from the result of *JaroWinkler* and the data characteristics, two similarity thresholds, one for the vessel name and the other for the origin/destination are defined. The vessel name similarity threshold is considered as 0.03 and value for the trip threshold is 0.06 (Given p as the similarity value, $0 \leq p \leq 1$, $p= 0$ shows the exact match and $p=1$ indicates that the two strings are dissimilar). These thresholds are referred as VESSEL_NAME_SIMILARITY_THRESHOLD and TRIP_ SIMILARITY_THRESHOLD in Appendix C.

3.4.5 **Display Client**

The Display Client module is the graphical user interface of ODADS. It is a web-based application and supports the functionalities of the JDL Level 5 processing such as HCI utilities, cognitive aids and focus/defocus attention. While designing the user interface, the six principles of the user interface design that are based on the usage-centered design approach are considered. According to Constantine and Lockwood (1999), these principles are: structure, simplicity, visibility, feedback, tolerance, and reuse.

Figure D.1 of Appendix D presents a snapshot of the user interface. When a user logs in to the system, the main view of the ODADS user interface is displayed. It consists of a geographical display and four tabs, namely *Anomaly*, *Dataset*, *Reports* and *Setting*. The geographical display utilizes the Google maps API features to enable the vessel traffic monitoring in the area of interest. Vessels are represented on the map by different colors according to their types. *Ferries* and *passenger* vessels are green, *cargo* vessels are cyan and *tankers* are yellow. In order to give a better perspective of the surveillance area, *tug* and *pilot* vessels are also displayed on the map. The *tug* and *pilot* vessels are presented in gray and dark blue, respectively. By clicking on each vessel, an information window appears which contains the vessel information such as name, MMSI number, IMO number, flag, heading, origin, destination, EAT and anomaly type. The AD process is performed every 10 minutes and consequently, the map is updated and if any new anomaly is detected, the *Anomaly* tab becomes red in order to inform the user about the incoming suspicious activities. The newly detected anomalous vessels are displayed in red and the previously detected anomalies are shown in light red. The *Anomaly* tab presents the overall information about the number of detected anomalies, total number of vessels and the last data collection time. Moreover, anomalous vessels are shown in a table which has the search, pagination and zoom to map utilities. The *Dataset* tab shows the ports and pilots timetables within a time interval and in a common representation format. This common data representation format contains vessel information such as name, type, company name, origin, destination, time, status and the name of the data source. The *Reports* tab consists of two sections. The first section provides reports regarding the total number of each anomaly and detailed information about anomalous vessels in a time interval in the HTML or EXCEL formats. The second section presents graphical reports about daily statistics of the detected anomalies in a stacked chart and overall statistics of all detected anomalies in a pie chart. The *Setting* tab is designed for getting any kind of information or system settings from the user. In the current implementation this tab is responsible for presenting the information about vessels that should be kept under surveillance and providing the features to add, remove, edit or disable/enable a vessel. Any changes in this tab would be considered in the next period of the AD process.

The technologies that are used for implementing this module are Java, Spring Framework, JSP, JavaScript, jQuery and Google maps API.

3.5 Experimental Design

Before using ODADS in real life, it is important to figure out to what extent the results of the system are accurate. Accuracy is the degree to which the estimates or measurements of a quantity correctly describe the exact value of that quantity. In other words, accuracy is the proportion of true results in the population. To evaluate the system accuracy, the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) are needed. Accuracy is calculated by the following formula (Han, Kamber, & Pei, 2011):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The first step in designing the experiment is to identify the population. The population consists of the vessel traffic data in the surveillance area. Since the population is too large and it is impossible to look into all members manually to count the number of TP, FP, TN and FN, a sample should be taken from the population. The second step is to identify the sampling procedure. The important factors in sampling are the sampling frame, method and the sample size. The sampling frame consists of all members of the population that have the chance to be included in a sample and must be representative of the population. In this thesis the sampling frame is the vessel traffic data related to AIS, ports and pilots in the surveillance area, which are provided by ODADS. Due to the high volume of traffic through the surveillance area, it is expected that the majority of anomalies can be observed in one week execution of the system. Therefore, one week of vessel traffic data in April 2012 is used as the sample frame. For choosing a sampling method, it is important to decide whether or not to use probability sampling. In probability sampling, all members of the population have the chance of being selected in the sample. If the members have an equal probability of selection, the sample can be unbiased. In this thesis it is possible to perform a random sampling with equal probabilities. However, because of the ODADS periodic data collection (every 10 minutes), the data regarding to a specific vessel can be selected multiple times. To omit the effects of repeated data in a sample, a multistage sampling is performed. In the first stage, a simple random sampling without replacement is done for selecting the time slots that ODADS attempts to collect and analyze the data. After selecting the time slots, the corresponding data for each time slot will be selected by a stratified sampling. Three strata are defined according to the type of vessels: ferry and passenger, cargo and tanker vessels. Selection of vessels is also limited to the vessels that are originated from or targeted to the four particular ports.

The total number of time slots in the sample frame is 835. This means that on average, ODADS collects data 139 times a day. In the first stage of sampling a random timeslot is selected for each day, which results in 7 time slots for one week. Then, by considering the described limitation in the selection process, the average number of entire data in a selected time slot is about 100 records. Among these records, 30 records are selected by stratification. Almost 73 % of the vessels in each time slot are moored. Since the majority of the anomalies are related to the vessels trips, a limitation on the number of moored vessels in the samples is defined. In this way, it can be possible to check more anomalies in the evaluation process. The second stage of sampling is repeated by taking into consideration that the number of moored vessel in the sample cannot exceed from the half of the sample size (in this case 15).

3.6 Validation Design

After completing the implementation of ODADS, a meeting is held with the Coastguard representatives. During the meeting the system is presented and it is determined that ODADS will be used by the Coastguard officers for a period of time to verify the validity of the detected anomalies and evaluate the usefulness of such system for the Coastguard. Therefore, ODADS is installed on an Internet server and another meeting is arranged with

the Coastguard to define the validation process. As the outcome of the meeting, it is decided the validation will take place at the Coastguard headquarters office in Karlskrona, Sweden for four weeks (April 23, 2012 to May 18, 2012), at any time during working hours (08:00 to 17:00). The officers are supposed to evaluate the detected anomalies by checking them against the available data in the systems and data sources that are used during the normal operational activities at the Coastguard. They are asked to provide weekly report about their evaluation results in order to decrease the possible malfunctioning of the system and the validation process. An Excel form is provided to the officers in order to receive the reports in a common format. While investigating the validity of an alarm for a specific vessel, the following information should be reported: detection time, vessel information (such as name, MMSI and IMO), trip information, type of the alarm (true/false), systems or data sources that are used to investigate the detected anomaly. The Coastguard officers are also asked for providing any information about vessels that are marked as anomaly according to the available systems at the Coastguard, but ODADS is unable to recognize them. Although having such information is useful for analyzing the validation results, it is impossible for the Coastguard to provide this information. A detected anomaly for a vessel is true if it can be confirmed by the available data sources at the Coastguard and consequently it is false if the authorized data sources provide any information that declines the detected anomaly. No further assessment is done regarding the classification of the detected anomalies to true and false alarms.

4. VALIDITY THREATS

In this thesis there are some issues that can threat the validity of the results and they should be considered before developing the system and performing the evaluation and validation. Construct validity refers to the extent to which the results of a study reflect the theory or the concept behind (Shadish, Cook, & Campbell, 2002). The main issue that may threat the construct validity is the design and reliability of implementation. Results can be affected by the potential faults that may happen in the implementation either because of programming faults or lack of tuning. In addition, the inaccurate nature of the open data that are used may have some effect on the results. The open data can have errors due to the human operator mistakes. They do not follow a similar format and can be unavailable for a while or are not updated immediately. To diminish the undesirable effect of the data, creating a common data representation format, filtering of the data and using approximate string matching techniques would be helpful. For decreasing the effect of programming faults, the validity of implemented application should be tested different times with both real and manipulated data that contain anomalies. There are also some parameters in the application that are needed to be selected correctly. For instance, in order to match the vessel information, string matching techniques are required and to determine that two strings are matched with each other, a similarity threshold should be used. Choosing an inappropriate value for such parameters will lead to incorrect results. The other threat of validity, which can occur while performing the evaluation and validation, targets both the internal and external validities. Internal validity ensures that the observed relationship between the treatment and outcome is due to a casual relationship and it is not because of an uncontrolled factor. External validity refers to the ability of generalizing the result of the study to other domains, times or places (Shadish et al., 2002). The threat to internal and consequently the external validity may occur if the data that are used in the evaluation process are biased and not representative of the population. In such situation generalizing the results of the treatments to the whole system is unrealistic. To prevent this issue, the samples, which are taken from the population in the experiment, should be real representative of the population. As well as the experiment, while performing the validation by the subject matter experts the same procedure for selecting and checking the detected anomalies should be followed.

5. VERIFICATION

To ensure that ODADS works properly, the system is tested manually with both real and manipulated data. The tests are performed during the implementation and also after completing the system. At first, a number of vessels with different types of anomalies are inserted to the real collected data to check whether all types of anomalies can be detected by ODADS. Then, the system is run for a period of time and the detected anomalies are checked manually against the available data to make sure about their correctness. During the test phase the Anomaly Detector module is updated and some of the detection conditions are narrowed down. The process is repeated until the system can detect all the anomalies correctly.

Before starting the validation process, the system is used by the subject matter experts from the Coastguard and according to their comments some of the algorithms in the Anomaly Detector module are updated. For example, ferries usually provide their schedule monthly or once in a couple of months; therefore, their arrival is not always available in daily schedule of the ports and the VESSEL_NOT_INFORMED_PORT and UNUSUAL_ TRIP_PATTERN anomalies are often set for them. For this reason, these two anomalies will not be checked for ferries.

6. EXPERIMENTAL RESULTS

The results of the system execution during the specified week are provided. To provide an overview of the detected anomalies by ODADS, Table 6 illustrates the total number of vessels in the surveillance area. Table 7 shows the number of detected types of anomaly for each day and finally Table 8 provides the total percentage of each anomaly type during that week.

Table 6

Total number of vessels in the surveillance area

	Days							
	1	2	3	4	5	6	7	Avg
Total number of vessels	614	623	663	665	669	695	784	673.29
Cargo, Tanker, Passenger and Ferry vessels	345	349	365	369	370	372	394	366.29
Cargo, Tanker, Passenger and Ferry vessels that are originated from or targeted to the specified ports	136	142	141	137	145	149	142	141.71

Table 7

Total and the average number of detected anomalies during one week of execution

	Days							
Anomaly	1	2	3	4	5	6	7	Avg
ARRIVAL_TIME_MISMATCHED	23	24	27	23	23	26	21	23.86
VESSEL_NOT_INFORMED_PORT	8	22	13	12	16	14	18	14.71

(Continued)

Anomaly	Days							Avg
	1	2	3	4	5	6	7	
VESSEL_NOT_LEFT_PORT	5	11	10	11	8	9	4	8.29
UNUSUAL_TRIP_PATTERN	1	6	3	5	4	4	3	3.71
VESSEL_NOT_USED_PILOT	1	3	2	2	2	2	2	2.00
UNUSUAL_TRIP_PATTERN_AND_VESSEL_ ARRIVAL_TIME_MISMATCHED	3	1	1	1	4	1	1	1.71
VESSEL_MOORED_IN_PORT	2	1	1	1	1	1	2	1.29
UNUSUAL_TRIP_PATTERN_AND_ VESSEL_NOT_INFORMED_PORT	0	0	0	1	1	1	1	0.57
VESSEL_ENTERED_PORT_WITHOUT_NOTIE	0	0	1	0	0	1	0	0.29
VESSEL_NOT_ENTERED_PORT	0	2	0	0	0	0	0	0.29
UNDER_SURVEILLANCE_VESSEL	1	1	0	0	0	0	0	0.29
VESSEL_ARRIVAL_TIME_MISMATCHED_ AND_VESSEL_NOT_USED_PILOT	0	0	1	0	0	1	0	0.29
VESSEL_ORDERED_PILOT_AND_NOT_ INFORMED_PORT	0	1	0	0	0	0	0	0.14
UNUSUAL_TRIP_PATTERN_AND_VESSEL_ NOT_USED_PILOT	0	0	0	0	0	0	1	0.14
VESSEL_ORDERED_PILOT_AND_NOT_ INFORMED_PORT_AND_VESSEL_NOT_ USED_PILOT	0	1	0	0	0	0	0	0.14
UNUSUAL_TRIP_PATTERN_AND_ VESSEL_ORDERED_PILOT_AND_NOT_ INFORMED_PORT	0	0	0	0	0	0	0	0.00
UNUSUAL_TRIP_PATTERN_AND_VESSEL_ NOT_USED_PILOT_AND_VESSEL_ARRIVAL_ TIME_MISMATCHED	0	0	0	0	0	0	0	0.00
VESSEL_ENTERED_PORT_WITHOUT_ NOTICE_AND_NOT_LEFT_PORT_ON_TIME	0	0	0	0	0	0	0	0.00
VESSEL_NOT_ENTERED_PORT_AND_NOT_ USED_PILOT	0	0	0	0	0	0	0	0.00
UNUSUAL_TRIP_PATTERN_AND_VESSEL_ NOT_ENTERED_PORT	0	0	0	0	0	0	0	0.00
UNUSUAL_TRIP_PATTERN_AND_ VESSEL_NOT_ENTERED_PORT_AND_ VESSEL_NOT_USED_PILOT	0	0	0	0	0	0	0	0.00
UNUSUAL_TRIP_PATTERN_AND_VESSEL_ ORDERED_PILOT_AND_NOT_INFORMED_ PORT_AND_VESSEL_NOT_USED_PILOT	0	0	0	0	0	0	0	0.00

Table 8

The percentage of each anomaly during one week of execution

Anomaly	Count (%)
ARRIVAL_TIME_MISMATCHED	41.34
VESSEL_NOT_INFORMED_PORT	25.49
VESSEL_NOT_LEFT_PORT	14.36
UNUSUAL_TRIP_PATTERN	6.43
VESSEL_NOT_USED_PILOT	3.47
UNUSUAL_TRIP_PATTERN_AND_VESSEL_ARRIVAL_TIME_MISMATCHED	2.96
VESSEL_MOORED_PORT	2.23
UNUSUAL_TRIP_PATTERN_AND_VESSEL_NOT_INFORMED_PORT	0.99
VESSEL_ENTERED_PORT_WITHOUT_NOTICE	0.50
VESSEL_NOT_ENTERED_PORT	0.50
UNDER_SURVEILLANCE_VESSEL	0.50
VESSEL_ARRIVAL_TIME_MISMATCHED_AND_VESSEL_NOT_USED_PILOT	0.50
VESSEL_ORDERED_PILOT_AND_NOT_INFORMED_PORT	0.24
UNUSUAL_TRIP_PATTERN_AND_VESSEL_NOT_USED_PILOT	0.24
VESSEL_ORDERED_PILOT_AND_NOT_INFORMED_PORT_AND_ VESSEL_NOT_USED_PILOT	0.24
UNUSUAL_TRIP_PATTERN_AND_VESSEL_NOT_ENTERED_PORT	0.00
UNUSUAL_TRIP_PATTERN_AND_VESSEL_ORDERED_PILOT_AND_NOT_ INFORMED_PORT	0.00
VESSEL_ENTERED_PORT_WITHOUT_NOTICE_AND_NOT_LEFT_PORT_ON_ TIME	0.00
VESSEL_NOT_ENTERED_PORT_AND_NOT_USED_PILOT	0.00
UNUSUAL_TRIP_PATTERN_AND_VESSEL_NOT_USED_PILOT_AND_ VESSEL_ARRIVAL_TIME_MISMATCHED	0.00
UNUSUAL_TRIP_PATTERN_AND_VESSEL_NOT_ENTERED_ PORT_AND_VESSEL_NOT_USED_PILOT	0.00
UNUSUAL_TRIP_PATTERN_AND_VESSEL_ORDERED_PILOT_AND_NOT_ INFORMED_PORT_AND_VESSEL_NOT_USED_PILOT	0.00

After carrying out the sampling, all the samples are checked against the primary identified anomalies (A1-A9). To compute the number of TP, FP, TN and FN, a confusion

matrix is created based on the nine classes of anomalies and the normal class (Table 9). The accuracy of the system is calculated as follow:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{17 + 192}{17 + 192 + 0 + 1} = 0.99$$

The existing FN for the ARRIVAL_TIME_MISMATCHED anomaly is due to the wrong provided AIS data by the vessel or possibly the *MarineTraffic.com* website and also the limitation of the system for considering all conditions. In this case, the vessel arrival time belongs to a couple of days before the current date and for this reason it is ignored by the system. However, it is quite possible to handle such situations if additional sources of AIS data are available.

Table 9

Confusion matrix for the nine classes of anomalies and the normal class

		Predicted class										
		A1	A2	A3	A4	A5	A6	A7	A8	A9	Normal	Total
Actual class	A1	6	-	-	-	-	-	-	-	-	-	6
	A2	-	7	-	-	-	-	-	-	-	1	8
	A3	-	-	-	-	-	-	-	-	-	-	0
	A4	-	-	-	-	-	-	-	-	-	-	0
	A5	-	-	-	-	3	-	-	-	-	-	3
	A6	-	-	-	-	-	1	-	-	-	-	1
	A7	-	-	-	-	-	-	-	-	-	-	0
	A8	-	-	-	-	-	-	-	-	-	-	0
	A9	-	-	-	-	-	-	-	-	-	-	0
	Normal	-	-	-	-	-	-	-	-	-	192	192
	Total	6	7	0	0	3	1	0	0	0	193	17

7. VALIDATION RESULTS

The sea monitoring system that is used by the Coastguard officer is called *SJÖBASIS*[1]. *SJÖBASIS* aggregates the maritime data from different systems and agencies with the aim of improving the efficiency of MS. In *SJÖBASIS*, the required data that contain vessel position, speed, heading, arrival/departure time and trip, are obtained from the following sources:

- *SafeSeaNet*[2]: A vessel traffic monitoring and information system, which is the centralized European platform for maritime data exchange and linking maritime authorities across Europe. It enables the member countries to provide and receive information about vessels, vessels movements and hazardous cargoes. Main sources of information include AIS-based position reports and notification messages that are sent by designated authorities in the participating countries. This system is available via the Swedish Maritime Administration.

[1] *www.kustbevakningen.se/sv/granslos-samverkan/sjoovervakningsuppdraget/samverkan-sjoinformation/*
[2] *www.emsa.europa.eu/operations/maritime-surveillance/safeseanet.html*

- *SjöC*: The Swedish naval forces surveillance center, which uses military sensors for correlating positions with AIS.
- AIS: The international system for presenting the vessels position and identity.
- *HELCOM AIS*[1]: The national AIS data that is consolidated in cooperation of countries around the Baltic Sea.

The officer checks the validity of anomalies according to the priority that each anomaly has for him. Table 10 provides the priority of the anomalies for an officer with an interest in law enforcement.

Table 10

Prioritization of the identified anomalies from a law enforcement officer's point of view

Anomaly	Priority
ARRIVAL_TIME_MISMATCHED	5
VESSEL_ENTERED_PORT_WITHOUT_NOTICE	5
VESSEL_NOT_ENTERED_PORT	5
UNDER_SURVEILLANCE_VESSEL	5
VESSEL_NOT_LEFT_PORT	4
VESSEL_MOORED_IN_PORT	4
UNUSUAL_TRIP_PATTERN	3
VESSEL_NOT_INFORMED_PORT	2
VESSEL_NOT_USED_PILOT	1
VESSEL_ORDERED_PILOT_AND_NOT_INFORMED_PORT	1

Note. 5 shows the highest priority and 1shows the lowest.

During the four-week validation period, ODADS is used at the Coastguard for 12 working days and in total 76 of the detected anomalies are evaluated. Table 11 presents the validation results. Among the evaluated anomalies, there are a number of anomalous vessels that remain unchecked due to a lack of corresponding data in the Coastguard systems. A large number of detected anomalies are related to the ARRIVAL_TIME_MISMATCHED anomaly that in many cases can be due to the inconsistent time formats (UTC, CET and EET either with or without daylight saving) in different data sources and various settings in the AIS transmitters. In these cases, the detected anomalies by ODADS are logically true, however, in actual fact they are not. For example, a vessel that is going from Helsinki to Stockholm is reporting its arrival time according to the Finland local time (EET format) instead of using the UTC format. Therefore, this time difference causes that the arrival time reported by the vessel does not match with the expected arrival time of the vessel at the destination port, which leads to set the ARRIVAL_TIME_MISMATCHED anomaly for that vessel.

[1] *www.helcom.fi/BSAP/ActionPlan/en_GB/SegmentSummary/*

Table 11

Validation result of the Coastguard

Anomaly	Alarms		
	True	False	Not Checked
VESSEL_NOT_INFORMED_PORT	7	3	4
ARRIVAL_TIME_MISMATCHED	19	5	3
VESSEL_NOT_USED_PILOT	2	-	-
UNUSUAL_TRIP_PATTERN	7	6	-
VESSEL_NOT_LEFT_PORT	1	1	-
VESSEL_ORDERED_PILOT_AND_NOT_INFORMED_PORT	2	-	-
VESSEL_MOORED_IN_PORT	-	3	-
UNDER_SURVEILLANCE_VESSEL	1	-	-
UNUSUAL_TRIP_PATTERN_AND_VESSEL_NOT_ INFORMED_PORT	5	1	-
UNUSUAL_TRIP_PATTERN_AND_VESSEL_ ARRIVAL_ TIME_ MISMATCHED	4	-	-
UNUSUAL_TRIP_PATTERN_AND_VESSEL_ORDERED_ PILOT_AND_NOT_INFORMED_PORT	-	1	-
VESSEL_ORDERED_PILOT_AND_NOT_INFORMED_ PORT_AND_VESSEL_NOT_USED_PILOT	1	-	-
Total count	49	20	7
Total count (%)	64.47	26.32	9.21

In addition to the comparative analysis in the validation process, modus operandi of using an anomaly detector that takes advantage of open data is investigated. The Coastguard officer uses an analysis tool[1] to analyze the ODADS excel reports and draw conclusions regarding the modus operandi of the system in emergency situations that can have impact on the MS operations. One of the possible analyses of the system reports can be the investigation of vessels with multiple anomalies. Here are some examples of the vessels with multiple anomalies. A cargo vessel has recurring anomalies related to the arrival/departure time, trip and notification to the port. From the types of anomalies that are detected for this vessel, it can be concluded that the vessel behavior points to a higher threat concerning customs and border. For a passenger vessel the following anomalies are often detected: VESSEL_NOT_INFORMED_PORT, ARRIVAL_TIME_MISMATCHED, VESSEL_NOT_ ENTERED_PORT and VESSEL_NOT_LEFT_PORT. These anomalies may happen because of inaccurate or wrong provided data by the vessel. The conclusion and modus operandi for this vessel is that the duty officer will contact the vessel to highlight the importance of submitting accurate information. In real life, occurring anomalies such as VESSEL_NOT_ENTERED_PORT or VESSEL_NOT_LEFT_PORT for a passenger vessel can be related to serious issues such as an accident and it will result in increased scrutiny for that vessel.

It is also possible to look into the relation between the types of vessels and the detected anomalies. Such assessment can be used for strategic and risk analysis. For tanker vessels the

[1]IBM i2 Analyst's Notebook available at *www.i2group.com/us/products/analysis-product-line/ibm-i2-analysts-notebook*

most popular anomaly is VESSEL_NOT_INFORMED_PORT. This anomaly has a high priority for emergency preparedness for accidents involving tankers. Tankers with UNUSUAL_TRIP_PATTERN anomaly are a potential risk to other vessels and can cause accidents. From a risk assessment point of view the combination of this anomaly with the VESSEL_NOT_ENTERED_PORT or VESSEL_NOT_USED_PILOT anomalies can lead to high-risk situations. The most occurring anomaly for ferries and passenger vessels is ARRIVAL_TIME_MISMATCHED. In several cases this anomaly is detected incorrectly because of the wrong reported arrival time; however, this anomaly has great importance for the authorities to plan their operations regarding ferries and passenger vessels arrivals effectively. The most serious anomaly for ferries and passenger vessels is VESSEL_NOT_ LEFT_PORT and the authorities should suspect that some form of accident or difficulty is arising regarding to the departure of the vessels. For cargo vessels, the most recurring anomalies are VESSEL_NOT_INFORMED_PORT and ARRIVAL_TIME_MISMATCHED. However, some of the ARRIVAL_TIME_MISMATCHED anomalies are false alarms because of the incorrect data. The VESSEL_NOT_INFORMED_PORT anomaly is important for ports security and safety. The prior notification to the ports is obligatory for vessels, but the fine for breaking this rule is negligible which lets the vessels that are involved in illegal activities such as smuggling disobey this rule. The most serious anomalies for the cargo vessels are the VESSEL_NOT_ENTERED_PORT and UNDER_SURVEILLANCE_ VESSEL anomalies.

Furthermore, looking into the most frequent anomalies for different ports will assists the maritime authorities to make their decisions more efficiently. According to the report analysis, the most frequent anomaly in Stockholm and Nynäshamn ports is ARRIVAL_ TIME_MISMATCHED, in port of Kapellskär is VESSEL_NOT_INFORMED_PORT and in Norrköping port is UNUSUAL_TRIP_PATTERN. One possible conclusion from the most popular anomalies for the ports is that the ports authorities should be informed of the divergence in the traffic flow and the operational management functions in order to plan and allocate resources efficiently. On the other hand, in some cases the anomalies are too common for a port because of the inaccurate provided data and they can be disregarded by the officer.

The received feedback from the Coastguard representatives during the validation process shows that ODADS is implemented in a way that it can assist the operator to have a better understanding of the ongoing maritime activities. They believe that the ODADS results are reliable and the quality of the open data that are used is good and can be used in real life. The functionality, usability and visualization of ODADS are satisfying. In addition to illustrating the vessel traffic data in a simple, clear and informative way, ODADS can provide clear statement about the anomalies and its statistical reports are beneficial when the authorities and freight companies conduct strategic analysis of maritime traffic and risk assessment. Finally, the capability of automated detection of anomalies by utilizing open data and the form of data representation could prove to be a valuable asset to the Coastguard.

8. DISCUSSION

Taking advantage of AD systems will assist authorities to tighten security in the MS domain. There are a number of studies which focused on developing AD systems by using knowledge-driven and data-driven approaches. For instance, Defense R&D Canada (Roy, 2008, 2010) developed a rule-based prototype for AD by exploiting maritime situational facts about both kinematic and static data of the domain. Edlund et al (2006) developed another prototype for a rule-based expert system to detect the anomalies regarding spatial and kinematic relation between objects. Riveiro and Falkman (2009) proposed using a combination of data-driven and knowledge-driven approaches to detect anomalies by use of a normal model of vessel behavior based on AIS data and experts rules. In the majority of

studies that addressed AD, the exploited data for the AD process were obtained from closed data sources and there is a lack of investigation on using open data sources for AD in the MS domain. Therefore, in this thesis ODADS is implemented by employing expert rules to investigate the potential open data as a complement to the closed data for AD in the MS domain. The accuracy of ODADS is evaluated via an experiment and the experimental results indicate an accuracy of 99% for the system. In addition, validity of the system results is evaluated in real life by the experts from the Swedish Coastguard. Despite the inaccurate nature of open data and by considering the fact that only open data sources are used in the system, the high number of true alarms (64.47%) in the validation process admits the validity of the system outcomes. Furthermore, there are no corresponding data in the authorized databases for 9.21% of the evaluated anomalies by the Coastguard. This fact refers to a potential information gap in the closed data sources. However, the considerable number of false alarms (26.32%) for a surveillance system is unsatisfying. The number of false alarms indicates the difference between the accuracy of the system and the validity of the results. Even though the data that are used in this thesis are obtained from relatively trusted data sources such as ports, the false alarms occur mostly because of the inaccurate data. The open data that are exploited by ODADS suffer from the errors due to human operator mistakes, irregular data update, data update latency and incompatible data format. In ODADS, there are situations that a detected anomaly is disappeared in the next periods of the system execution because of new arrival of the correct data. Frequent occurrence of false alarms distracts operator's attention from the real anomalies in the surveillance area. To decrease the false alarms in ODADS, the main solution is to integrate the open data with the closed data which can cover the lack of information or inaccuracy in the open data. In addition, defining a probability for detected anomalies can decrease the number of false alarms. This would be possible by analyzing the history of vessels behavior as well as the current situation and defining a probability threshold to omit the anomalies that have a lower probability than the threshold. Furthermore, having extra information regarding vessels such as crew and cargo information, can affect the probability of being a real anomaly for a specific vessel. For example, if a vessel has the ARRIVAL_TIME_MISMATCHED anomaly and it has a crew member with a criminal record or a special cargo, then there is a possibility that the vessel is stopped somewhere to exchange something. Therefore, in such situation, the probability of being a true anomaly is high.

According to the validation results, the UNUSUAL_TRIP_PATTERN anomaly creates the majority of false alarms. This is due in part to the statistical approach that is used for detecting this anomaly and also the wrong origin and destination information that the vessels provide. The lookup table that is created for storing the frequency of the trips between different places is not updated periodically. While populating the table, the ports timetables are used which can be incomplete. An alternative detection approach can be using of machine learning techniques which attempt to detect the anomaly according to the pattern of movements for individual vessels instead of the reported trip data by the vessels or ports.

In local areas such as the surveillance area in this thesis, mainly because of large amount of quality assured data and the limited size of the surveillance area, it is easier for the maritime authorities to track and control the vessels activities. Therefore, use of open AIS data in this region is not required and it should be prohibited to decrease the negative impacts of open data on the system results. On the other hand, when the vessel information beyond the EEZ is required, the value of open data becomes more obvious.

9. CONCLUSION AND FUTURE WORK

This thesis investigated the potential open data as a complementary resource for Anomaly Detection (AD) in the Maritime Surveillance (MS) domain. A framework for AD was proposed based on the usage of open data sources along with other traditional sources of

data. According to the proposed AD framework and the algorithms for implementing the expert rules, the Open Data Anomaly Detection System (ODADS) was developed. To evaluate the accuracy of the system, an experiment on the vessel traffic data was conducted and an accuracy of 99% was obtained for the system. There was a false negative case in the system results that decreased the accuracy. It was due to incorrect AIS data in a special situation that was not possible to be handled by the detection rules in the scope of this thesis. The validity of the results was investigated by the subject matter experts from the Swedish Coastguard. The validation results showed that the majority of the ODADS evaluated anomalies were true alarms. Moreover, a potential information gap in the closed data sources was observed during the validation process. Despite the high number of true alarms, the number of false alarms was also considerable that was mainly because of the inaccurate open data. This thesis provided insights into the open data as a complement to the common data sources in the MS domain and is concluded that using open data will improve the efficiency of the surveillance systems by increasing the accuracy and covering some unseen aspects of maritime activities.

In the future, it is important to investigate how the open data sources in the maritime domain can be used in a global perspective. In this thesis, the surveillance area was limited to a local area which is fully covered by the authorities' data sources. When the data beyond the exclusive economic zone are needed, it is more valuable to use open data sources. By taking advantage of the subject matter experts' knowledge about MS, it would be possible to figure out how the global open data should be exploited for the surveillance purpose. Integration of the open data with maritime confidential data can improve the efficiency of MS and should be considered as a further improvement of the system. Another improvement can be considering a probability for each detected anomaly according to the history of the vessels behavior and the current situation. Moreover, further investigation on the other sources of open data such as social data, which is created and shared through social media platforms, and online videos from the ports activities in the high risk regions, will be useful. The data that are used in ODADS are relatively trusted, but in case of using other open data sources in the MS domain for AD, the quality assurance of the data should be investigated. As well as using knowledge based systems, taking advantage of data-driven approaches such as machine learning techniques can increase the efficiency of the MS systems. Finally, the next step for improving the MS systems after being equipped with the AD functionality is to predict the future threats or incoming anomalies based on the analysis of the current situation.

APPENDIX A: OPEN AND CLOSED DATA SOURCES

Table A.1

Maritime open and closed data sources that are available via the Internet

Organization name	Categories	NAR	AR		Provider
			FR	NFR	
Baltic and International Maritime Council www.bimco.org/	Shipping information Consulting services			●	Denmark
Association of Ship Brokers and Agents, Inc. www.asba.org/	Ship brokers information			●	USA
Bureau International des Containers www.bic-code.org/	Freight containers-coding Identification Marking			●	France
European Maritime Safety Agency www.emsa.europa.eu/oil-recovery-vessels/vessel-technical-specifications.html	Maritime safety Prevention of pollution from ships	●			Portugal
ICC Commercial Crime Services www.icc-ccs.org/	Fraud in international trade	●		●	UK
International Association of Classification Societies www.iacs.org.uk/shipdata/default.aspx	Maritime safety Regulation	●		●	UK
International Group of P&I Clubs www.igpandi.org/Home	liability and insurance issues	●		●	UK
International Association of Independent Tanker Owners www.intertanko.com/	Transportation safety Prevention of pollution from ships Free competition			●	UK

(Continued)

Organization name	Categories	AR			Provider
		NAR	FR	NFR	
Oil Companies International Marine Forum www.ocimf.com/Home	Transportation safety Prevention of pollution from ships			●	UK
Internet Ships Register www.ships-register.com/	Ships information			●	UK
World Shipping Register www.world-register.org/	Ships information Ports information Companies information	●			—
Lloyd's Register Ships In Class www.lrshipsinclass.lrfairplay.com/default.aspx	Ships information		●		UK
International Telecommunication Union www.itu.int/ITU-R/index.asp?category=terrestrial&rlink=mars&lang=en	Ships information Coasts information Addresses of accounting authorities, administrations which notify information MMSI assigned to search and rescue aircraft MMSI assigned to AIS Aids to Navigation	●		●	Switzerland
International Maritime Consultancy Specialists in VTS www.maritime-vts.co.uk/	Vessel traffic services and maritime organization	●			UK
Port Directory www.port-directory.com/	Ports information	●			UK
Equasis www.equasis.org/EquasisWeb/public/HomePage?fs=HomePage	Ships information Companies information		●		Portugal
Q88.COM www.q88.com/Home.aspx?c=1	Questionnaire generator Ships information	●		●	USA
InforMare www.informare.it/indexuk.htm	Shipping information	●			Italy

(Continued)

Organization name	Categories	NAR	AR		Provider
			FR	NFR	
Paris Mou www.parismou.org/	Port State control	●			Netherlands
Tokyo Mou www.tokyo-mou.org/	Port State control	●			Japan
Australian Government Bureau of Meteorology www.bom.gov.au/	Weather, climate and water	●		●	Australia
Finnish Meteorological Institute en.ilmatieteenlaitos.fi/home	Weather, climate and water	●			Finland
Meteo France france.meteofrance.com/	Weather, climate and water	●		●	France
Earth Science Office NASA weather.msfc.nasa.gov/GOES/	Weather, climate and water	●			—
The Weather Channel www.weather.com/	Weather, climate and water	●			—
Ocean Color WEB NASA oceancolor.gsfc.nasa.gov/	Weather, climate and water	●		●	—
Earth European Space Agency earth.esa.int/ers/eeo4.10075/atsr_med.html	Weather, climate and water		●	●	Italy
Weather BBC www.bbc.co.uk/weather/	Weather, climate and water	●			UK
Sailwx.info www.sailwx.info/	Live marin information	●			USA
Ship.gr www.ship.gr/	Ship brokers information Ship suppliers Companies information	●			—

(Continued)

Organization name	Categories	AR			Provider
		NAR	FR	NFR	
American Bureau of Shipping www.eagle.org/eagleExternalPortalWEB/appmanager/absEagle/absEagleDesktop?_nfpb=true&_pageLabel=abs_eagle_portal_home_page	Classification Societies	●			USA
Det Norske Veritas www.dnv.com/	Classification Societies	●		●	Norway
Bureau Veritas Groups www.bureauveritas.com/wps/wcm/connect/bv_com/Group/Footer/Home/	Classification Societies	●		●	—
China Classifiaction Societies www.ccs.org.cn/en/index.htm	Classification Societies			●	China
HELLENIC REGISTER OF SHIPPING www.hrs.gr/index.htm	Classification Societies	●			Greece
Nippon Kaiji Kyokai www.classnk.or.jp/hp/en/index.aspx	Classification Societies	●			Japan
Vesseltracker.com www.vesseltracker.com/en/VesselArchive.html	AIS Data Ships information	●	●		Germany
Digital Seas www.digital-seas.com/start.html	AIS Data Ships information	●	●		Germany
MarineTraffic.com www.marinetraffic.com/ais/	AIS Data Ships information	●			Greece
Shipspotting.com www.shipspotting.com/	AIS Data Ships information	●	●		—
International Maritime Organization www5.imo.org/SharePoint/mainframe.asp?topic_id=334&offset	Piracy reports	●			UK
Copenhagen Malmö Port www.cmport.com/	Port Authorities	●			Denmark

(Continued)

Organization name	Categories	NAR	AR		Provider
			FR	NFR	
PORT OF GOTHENBURG www.portgot.se/prod/hamnen/ghab/dalis2b.nsf	Port Authorities	●			Sweden
Genoa Port Authority www.porto.genova.it/index.php/en	Port Authorities	●			Italy
Port of Klaipeda www.portofklaipeda.lt/en.php	Port Authorities	●			Lithuania
Philippine Ports Authority www.ppa.com.ph/	Port Authorities	●			Philippine
Panama Canal Authority www.pancanal.com/eng/index.html	Port Authorities	●			USA
UK P&I CLUB www.ukpandi.com/	liability and insurance issues	●	●		UK
The American Club www.american-club.com/	liability and insurance issues	●	●		USA
Steamship Mutual www.simsl.com/	liability and insurance issues	●	●		Bermuda
SKULD www.skuld.com/	liability and insurance issues	●			Norway
North of England P&I Association www.nepia.com/home/	liability and insurance issues	●			UK
The standard club www.standard-club.com/	liability and insurance issues	●		●	UK
Baltic Ports Organization www.bpoports.com/	Port coordinator			●	Denmark

Note. NAR = no authorization required; AR= authorization required; FR = free registration; NFR = non-free registration; Dashes indicate undisclosed information.

Table A.2

Data sources that are used in the implementation

Website name	Data type	Website URL
Marinetraffic.com	Real time information based on AIS systems	www.marinetraffic.com/ais/
Swedish Maritime Administration (Sjofartsverket)	Stockholm pilotage area	www.sjofartsverket.se/sv/Infrastruktur-amp-Sjotrafik/Lotsning/Lotsinfo/
Ports of Stockholm, Kapellskär and Nynäshamn	Vessels in port and expected arrival	www.stockholmshamnar.se/en/Karta/Vessel-calls/
Port of Norrköping	Vessels in port	www.norrkoping-port.se/anlop.php?page=snabb_fih&link=110|111
	Expected vessels arrival	www.norrkoping-port.se/anlop.php?page=snabb_fih_ank&link=110|111
Port of Helsinki	Cargo vessels in port	www.portofhelsinki.fi/cargo_traffic/vessels_in_ports
	Expected cargo vessels arrival	www.portofhelsinki.fi/cargo_traffic/arrival_ships
	Expected passenger vessels departure	www.portofhelsinki.fi/passengers/departure_times_and_terminals
	Expected passenger vessels arrival	www.portofhelsinki.fi/passengers/arrival_times_and_terminals
	Passenger vessels have visited the port before	www.portofhelsinki.fi/passengers/cruise_ships_that_have_visited_the_port
Port of Tallinn	Vessels in port	www.ts.ee/?op=ships_in_port&lang=eng
	Expected cargo vessels arrival	www.ts.ee/?op=cargo_ships_arrivals&lang=eng
	Expected passenger vessels arrival	www.ts.ee/?op=passenger_ship_arrivals&lang=eng
	Expected passenger vessels departure	www.ts.ee/?op=passenger_ship_departures&lang=eng

Table A.3

Required data for detection of each anomaly

Anomaly	Data
VESSEL_NOT_INFORMED_PORT	AIS - Port vessel arrival data
ARRIVAL_TIME_MISMATCHED	AIS - Port vessel arrival data
VESSEL_ENTERED_PORT_ WITHOUT_NOTICE	AIS - Port vessel arrival data - Port vessel departure/ in-port data
VESSEL_NOT_USED_PILOT	AIS - Pilot vessels service data
UNUSUAL_TRIP_PATTERN	AIS - Port vessel arrival data – Port vessel departure/ in-port data
VESSEL_NOT_LEFT_PORT	AIS - Port vessel departure data
VESSEL_NOT_ENTERED_PORT	AIS - Port vessel arrival data – Port vessel departure/ in-port data
VESSEL_ORDERED_PILOT_AND_ NOT_INFORMED_PORT	AIS – Port vessel arrival data – Pilot vessels service data
VESSEL_MOORED_IN_PORT	AIS – Port vessel departure/ in-port data

APPENDIX B: PORTS REGIONS

This section provides the geographical information about each port and its harbors.

Table B.1

Ports areas and their geographical positions

Port	Number of area[a]	Minimum longitude	Maximum longitude	Minimum latitude	Maximum latitude
Helsinki	1	24.899414	24.985057	60.140472	60.18381
Kapellskär	1	19.064606	19.079539	59.715444	59.728009
Norrköping	3	16.180591	16.228976	58.59246	58.614909
		16.231232	16.254241	58.608807	58.625638
		16.260810	16.269291	58.635730	58.640399
Nynäshamn	1	17.949054	17.982726	58.898194	58.932706
Stockholm	6	18.060464	18.108335	59.29992	59.332768
		18.101335	18.147693	59.33865	59.3577276
		18.023163	18.061236	59.318440	59.328193
		18.017862	18.034127	59.312907	59.317289
		17.821146	17.825557	59.360349	59.362994
		18.163756	18.174964	59.318795	59.322375
Tallinn	6	24.624438	25.010307	59.425881	59.524365
		24.021952	24.098137	59.327022	59.362901
		22.221453	22.249035	58.526984	58.541255
		24.793068	24.877004	59.553744	59.602800
		24.452785	24.585510	59.464992	59.497719
		23.845721	23.957913	59.345430	59.389721

Note. [a] In order to have a more precise AD, harbors and areas around each port that are used by vessels are identified.

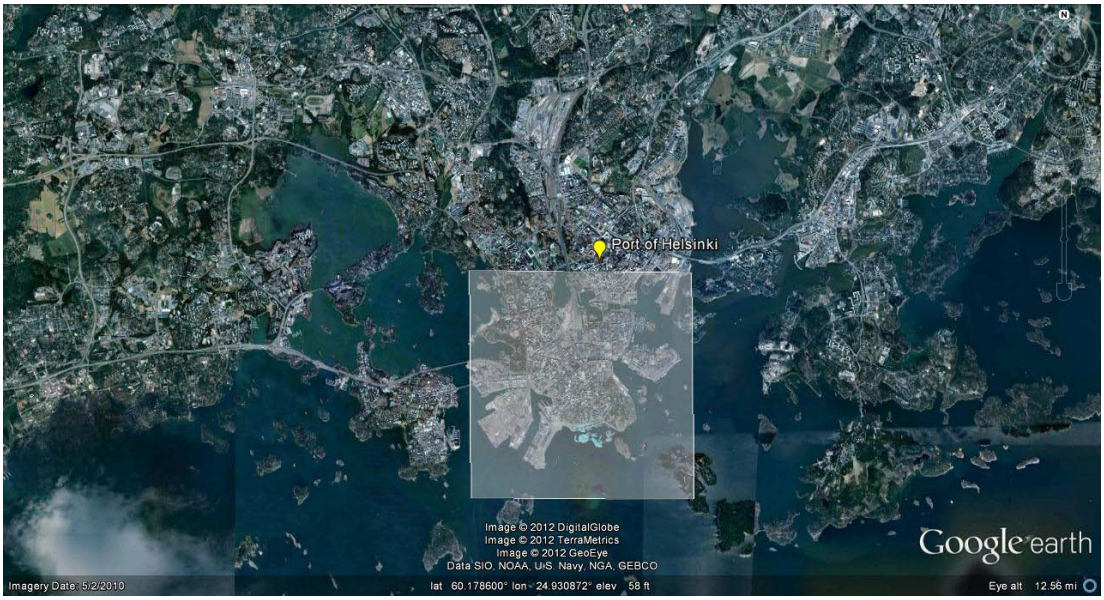

Figure B.1. Helsinki port area (The image is adapted from Google Earth).

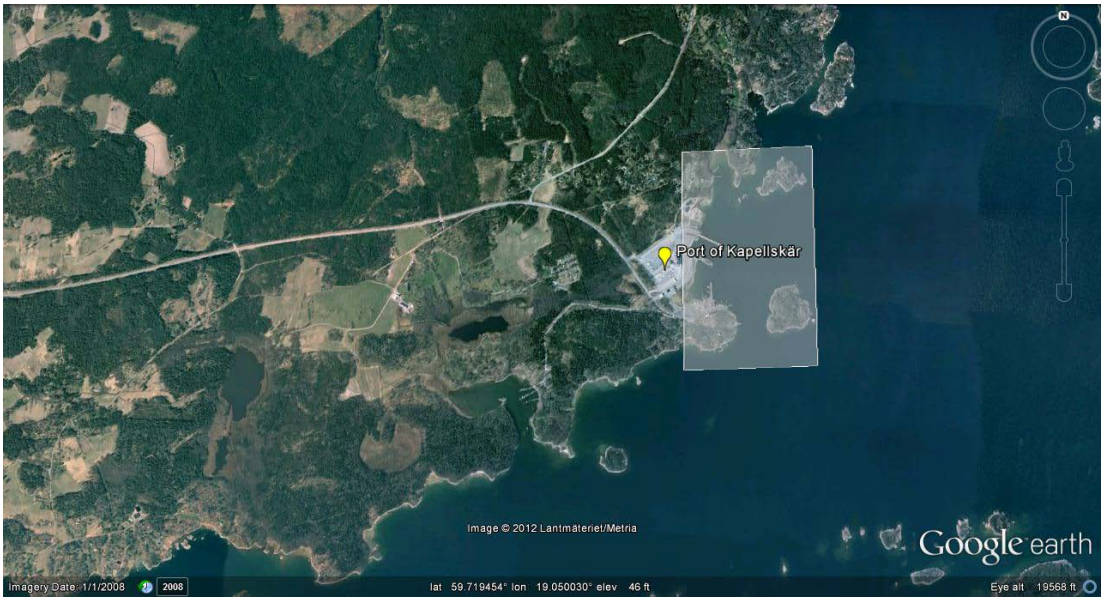

Figure B.2. Kapellskär port area, Stockholm group (The image is adapted from Google Earth).

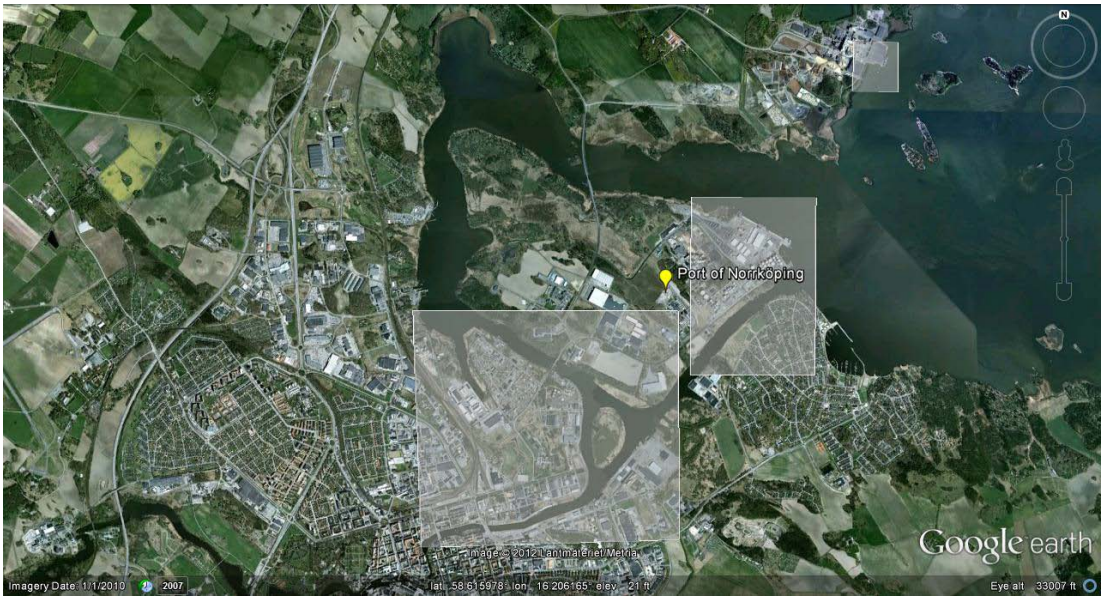

Figure B.3. Norrköping port area (The image is adapted from Google Earth).

Figure B.4. Nynäshamn port area, Stockholm group (The image is adapted from Google Earth).

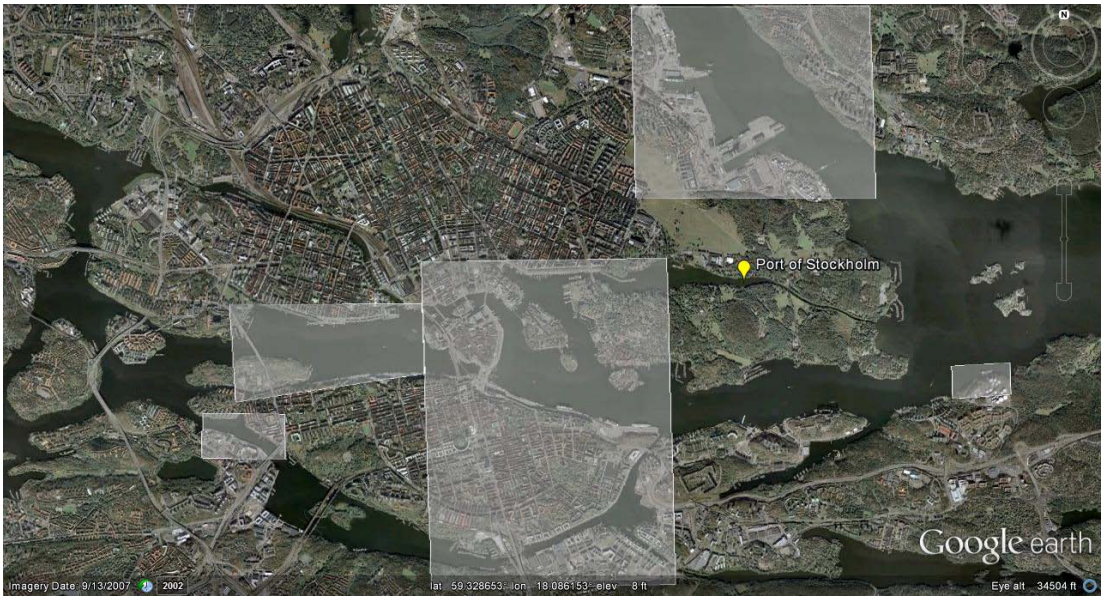

*Figure B.5.*Stockholm port areas, Stockholm group (The image is adapted from Google Earth).

Figure B.6. Tallinn port areas (The image is adapted from Google Earth).

APPENDIX C: ALGORITHMS

This section provides a list of all algorithms that are used for implementing Anomaly Detection module.

Algorithm 1 Main procedure for the Anomaly Detector module

while true

1. Extract AIS, Ports and Pilot data from database regarding to the latest time of data collection by the Data Collector module
2. Call methods regarding to the following algorithms:
    a. Algorithm 2
    b. Algorithm 3
    c. Algorithm 4
    d. Algorithm 5
    e. Algorithm 6
    f. Algorithm 7
    g. Algorithm 8
3. Determine the final type of anomaly for vessels (Algorithm 19)
4. Store detected anomalies in database
5. Wait until the new data are collected by the Data Collector module

end while

Algorithm 2 Detect VESSEL_NOT_INFORMED_PORT and ARRIVAL_TIME_MIS-MATCHED anomalies

```
ARRIVAL_TIME_DELAY_THRESHOLD ← 30 minutes
for each vessel ∈ AIS data do
  destination ← vessel destination
  if destination ≠ null and isRelatedToPorts(destination) and
       not isInPortsArea(vessel) then
    vessel_found ← false
    port_schedule ← vessels that are expected to arrive at destination port

    // For vessels that have not updated their trip information at the beginning of their trip
    if destination contains vessel origin and
         vesselLeftPortRecently(destination port, vessel) and not
          isTargetedPort(destination port, vessel) then
      continue
    end if
    for each port_vessel ∈ port_schedule do
      if (vessel arrival date = port_vessel arrival date or
           current date = port_vessel arrival date or

           // To consider an interval for changing dates
           |vessel arrival time - port_vessel arrival time| < FIVE_HOURS ) and
           isVesselSimilar(vessel, port_vessel) then
        vessel_found ← true
        if vessel arrival time ≠ port_vessel arrival time and
             |vessel arrival time - port_vessel arrival time| >
             ARRIVAL_TIME_DELAY_THRESHOLD then
          add(Vessel anomaly, ARRIVAL_TIME_MISMATCHED)
```

Algorithm 2 Detect VESSEL_NOT_INFORMED_PORT and ARRIVAL_TIME_MISMATCHED anomalies

```
            end if
            break
          end if
        end for
        if vessel_found ≠ true then
          add(Vessel anomaly, VESSEL_NOT_INFORMED_PORT)
        end if
      end if
    end for
```

Algorithm 3 Detect VESSEL_ENTERED_PORT_WITHOUT_NOTICE anomaly

```
for each port ∈ available ports
  port_vessels ← vessels that are currently available in port and arrived one hour ago
  for each port_vessel ∈ port_vessels do
    expected_vessels ← vessels that were expected to enter the port at vessel arrival time
    vessel_found ← false
    for each exp_vessel ∈ expected_vessels do
      if isVesselSimilar(port_vessel, exp_vessel) then
        vessel_found ← true
        break
      end if
    end for
    if vessel_found ≠ true then
      add(port_vessel anomaly, VESSEL_ENTERED_PORT_WITHOUT_NOTICE)
       (port_vessel)
    end if
  end for
end for
```

Algorithm 4 Detect VESSEL_NOT_USED_PILOT and VESSEL_ORDERED_PILOT_AND_NOT_INFORMED_PORT anomaly

```
PILOT_START_TIME_DELAY_THRESHOLD ← 30 minutes
for each vessel ∈ AIS data do
  for each pilot_vessel ∈ pilot schedule do
    if current date = pilot_vessel date and isVesselSimilar(vessel, pilot_vessel) then
      if vessel anomaly contains VESSEL_NOT_INFORMED_PORT then
        add(vessel anomaly,VESSEL_ORDERED_PILOT_AND_NOT_INFORMED_PORT)
      end if
      if pilot_vessel has not received service and pilot_vessel start_operation_time +
          PILOT_START_TIME_DELAY_THRESHOLD < current time then
        add(vessel anomaly , VESSEL_NOT_USED_PILOT)
      end if

      // To solve the issue for duplicate records with different status in pilot data
      if pilot_vessel has received service and vessel anomaly contains
          VESSEL_NOT_USED_PILOT then
        remove(vessel anomaly, VESSEL_NOT_USED_PILOT)
        break
```

Algorithm 4 Detect VESSEL_NOT_USED_PILOT and VESSEL_ORDERED_PILOT_AND_NOT_INFORMED_PORT anomaly

```
        end if
      end if
    end for
end for
```

Algorithm 5 Detect UNUSUAL_TRIP_PATTERN anomaly

```
TRIP_COUNT_THRESHOLD ← 2
for each vessel ∈ AIS data do
  if vessel origin ≠ null and vessel destination ≠ null and vessel type ≠ FERRY and
      (isRelatedToPorts(vessel origin) or isRelatedToPorts(vessel destination)) and not
       isInPortsArea(vessel) then
    count ← getTripHistoryForVessel(vessel)
    if count ≠ -1 and count < TRIP_COUNT_THRESHOLD then
      add(vessel anomaly , UNUSUAL_TRIP_PATTERN)
    end if
  end if
end for
```

Algorithm 6 Detect VESSEL_NOT_ENTERED_PORT anomaly

```
for each port ∈ available ports
  expected_vessels ← vessels that were expected to enter the port since one hour ago
  port_vessels ← vessels that are currently available in port
  for each exp_vessel ∈ expected_vessels do
    vessel_found ← false
    for each port_vessel ∈ port_vessels do
      if port_vessel time ≥ exp_vessel time and
          isVesselSimilar(exp_vessel, port_vessel) then
        vessel_found ← true
        break
      end if
    end for
    if vessel_found ≠ true then
      add(port_vessel anomaly, VESSEL_NOT_ENTERED_PORT)
      findVesselMatchInAISData(port_vessel)
    end if
  end for
end for
```

Algorithm 7 Detect VESSEL_NOT_LEFT_PORT anomaly

```
for each port ∈ available ports
  port_vessels ← vessels that were expected to leave the port since one hour ago
  for each port_vessel ∈ port_vessels do
    for each vessel ∈ AIS data do
      if isVesselSimilar(vessel, port_vessel) and  isMooredInPortsArea (vessel) then
        add(vessel anomaly , VESSEL_NOT_LEFT_PORT)
```

Algorithm 7 Detect VESSEL_NOT_LEFT_PORT anomaly

```
            break
          end if
        end for
      end for
    end for
```

Algorithm 8 Detect VESSEL_MOORED_IN_PORT anomaly

```
for each vessel ∈ AIS data do
  if not isInPortsArea(vessel) then
    vessel_found ← false
    for each port ∈ available ports do
      if vessel_found ≠ true then
        port_vessels ← vessels that are available in port for at least 30 minutes later
        for each port_vessel ∈ port_vessels do
          if isVesselSimilar(vessel, port_vessel) and
              vessel type = port_vessel type and
              isOtherSimilarVesselInPort(vessel) then
            add(vessel anomaly , VESSEL_MOORED_IN_PORT)
            vessel_found ← true
            break
          end if
        end for
      end if
    end for
  end if
end for
```

Algorithm 9 isInPortsArea(vessel)

```
in_area ← false
for each port ∈ available ports do
  rectangle ← get the pre-defined port area
  if vessel position is within the rectangle then
    in_area ← true
    break;
  end if
end for
return in_area
```

Algorithm 10 isMooredInPortsArea(vessel)

```
in_area ← false
for each port ∈ available ports do
  create a rectangle from pre-defined coordinates for the port area
  if vessel is not moving and vessel position is within the rectangle then
    in_area ← true
    break;
```

Algorithm 10 isMooredInPortsArea(vessel)

```
    end if
end for
return in_area
```

Algorithm 11 isVesselSimilar(vessel_1, vessel_2)

```
VESSEL_NAME_SIMILARITY_THRESHOLD ← 0.03
similarity ← calculate the similarity between the vessel_1 name and vessel_2 name using
             a string matching technique
is_similar ← false
if similarity < VESSEL_NAME_SIMILARITY_THRESHOLD then
   is_similar ← true
end if
return is_similar
```

Algorithm 12 isRelatedToPorts(name)

```
is_related ← false
for each port ∈ available ports do
   if name = port name or name contains port name or
        isTripSimilar(name, port name) then
      is_ related ← true
      break
   end if
end for
return is_related
```

Algorithm 13 isTripSimilar (trip1, trip2)

```
TRIP_SIMILARITY_THRESHOLD ← 0.06
similarity ← calculate the similarity between the two trips using a string matching technique
is_similar ← false
if similarity < TRIP_SIMILARITY_THRESHOLD then
   is_similar ← true
end if
return is_similar
```

Algorithm 14 vesselLeftPortRecently(port, vessel)

```
left_port ← false
port_vessels ← vessels that were expected to leave the port since twelve hours ago
for each port ∈ available ports do
   if isVesselSimilar(vessel, port_vessel) then
      left_port ← true
      break
   end if
end for
```

Algorithm 14 vesselLeftPortRecently(port, vessel)

return left_port

Algorithm 15 isTargetedPort(port, vessel)

```
is_targeted ← false
horizontal_line ← get the pre-defined horizontal line for the port
vertical_line ← get the pre-defined vertical line for the port
if vessel position is on the left side of vertical_line and
      vessel position is below the horizontal_line and
      vessel heading is to the north east then
   is_targeted ← true
else if vessel position is on the right side of vertical_line and
      vessel position is below the horizontal_line and
      vessel heading is to the north west then
   is_targeted ← true
else if vessel position is on the right side of vertical_line and
      vessel position is above the horizontal_line and
      vessel heading is to the south west then
   is_targeted ← true
else if vessel position is on the left side of vertical_line and
      vessel position is above the horizontal_line and
      vessel heading is to the south east then
   is_targeted ← true
end if
 return is_targeted
```

Algorithm 16 getTripHistoryForVessel(vessel)

```
vessel_found ← false
count ← 0
for each history_vessel ∈ vessel trip history do
  if isVesselSimilar(vessel, history_vessel) then
     vessel_found ← true
      trip ← vessel origin and destination
     if isTripSimilar(trip, history_vessel trip) then
        count ← history_vessel trip_count
       break
     end if
  end if
end for
if vessel_found ≠ true then
   count ← -1
return count
```

Algorithm 17 isOtherSimilarVesselInPort(ais_vessel)

```
is_similar ← false
for each vessel ∈ AIS data do
   if (vessel name = ais_vessel name and vessel type = ais_vessel type and
```

Algorithm 17 isOtherSimilarVesselInPort(ais_vessel)

```
        vessel mmsi ≠ ais_vessel mmsi)) and isInPortArea(vessel) then
      is_similar ← true
   end if
end for
return is_similar
```

Algorithm 18 findVesselMatchInAISData(port_vessel)

```
for each vessel ∈ AIS data do
   if isVesselSimilar(vessel, port_vessel) and vessel type = port_vessel type and
        vessel type ≠ FERRY then
      if port_vessel anomaly = VESSEL_ENTERED_PORT_WITHOUT_NOTICE and
           not isInPortsArea(vessel) then
        continue
      else if port_vessel anomaly = VESSEL_NOT_ENTERED_PORT and
          (vessel has ARRIVAL_TIME_MISMATCHED or isInPortsArea(vessel)) then
        continue
      else
        add(vessel anomaly , port_vessel anomaly)
      end if
   end if
end for
```

Algorithm 19 Determine the final type of anomaly for vessels

```
for each vessel ∈ AIS data do

   //A10
   if vessel ∈ smuggling list and vessel anomaly count ≠ 0 then
      vessel anomaly ← UNDER_SURVEILLANCE_VESSEL
      continue
  end if
   if vessel anomaly count > 1 then

      //A9
      if vessel anomaly contains VESSEL_MOORED_IN_PORT then
        anomaly ← VESSEL_MOORED_IN_PORT

      //A1,A5
      else if vessel anomaly contains VESSEL_NOT_INFORMED_PORT and
             vessel anomaly contains UNUSUAL_TRIP_PATTERN then
      anomaly ← UNUSUAL_TRIP_PATTERN_AND_VESSEL_NOT_INFORMED
                  _PORT

     //A2,A5
      else if vessel anomaly contains ARRIVAL_TIME_MISMATCHED and
           vessel anomaly contains UNUSUAL_TRIP_PATTERN then
        anomaly ← UNUSUAL_TRIP_VESSEL_ARRIVAL_TIME_MISMATCHED

      //A7,A5
      else if vessel anomaly contains VESSEL_NOT_ENTERED_PORT and
```

Algorithm 19 Determine the final type of anomaly for vessels

```
        vessel anomaly contains  UNUSUAL_TRIP_PATTERN then
       anomaly ← UNUSUAL_TRIP_PATTERN_AND_VESSEL_NOT_ENTERED
                    _PORT

  //A8,A5
  else if vessel anomaly contains  VESSEL_ORDERED_PILOT_AND_NOT_
        INFORMED_PORT and vessel anomaly contains UNUSUAL_TRIP
        _PATTERN then
    anomaly ←UNUSUAL_TRIP_PATTERN_AND_VESSEL_ORDERED_PILOT
                _AND_NOT_INFORMED_PORT

 //A3,A6
  else if vessel anomaly contains  VESSEL_ENTERED_PORT_WITHOUT_NOTICE
         and vessel anomaly contains  VESSEL_NOT_LEFT_PORT then
    anomaly ← VESSEL_ENTERED_WITHOUT_INFORMATION_NOT_LEFT
                _PORT_ ON_TIME

   //A2,A4

 else if  vessel anomaly contains  ARRIVAL_TIME_MISMATCHED and
         vessel anomaly contains  VESSEL_NOT_USED_PILOT then
    anomaly ←  VESSEL_ARRIVAL_TIME_MISMATCHED_AND
                _VESSEL_NOT_USED_PILOT

  //A2,A4,A5
  else if vessel anomaly contains ARRIVAL_TIME_MISMATCHED and
        vessel anomaly contains VESSEL_NOT_USED_PILOT and  vessel
        anomaly contains UNUSUAL_TRIP_PATTERN then
    anomaly ← UNUSUAL_TRIP_PATTERN_AND_VESSEL_NOT_USED_PILOT
                _AND_VESSEL_ARRIVAL_TIME_MISMATCHED

  //A4,A5
  else if vessel anomaly contains VESSEL_NOT_USED_PILOT and
       vessel anomaly contains UNUSUAL_TRIP_PATTERN then
    anomaly ← UNUSUAL_TRIP_PATTERN_AND_VESSEL_NOT_USED_PILOT

  //A7,A4
  else if  vessel anomaly contains VESSEL_NOT_ENTERED_PORT and
         vessel anomaly contains VESSEL_NOT_USED_PILOT then
    anomaly ← VESSEL_NOT_ENTERED_PORT_AND_NOT_USED_PILOT

  //A7,A4,A5
  else if vessel anomaly contains VESSEL_NOT_ENTERED_PORT and
        vessel anomaly contains  VESSEL_NOT_USED_PILOT and
        vessel anomaly contains  UNUSUAL_TRIP_PATTERN then
    anomaly ← UNUSUAL_TRIP_PATTERN_AND_VESSEL_NOT_ENTERED
                _ PORT_ AND_VESSEL_NOT_USED_PILOT

  //A8,A4
  else if vessel anomaly contains VESSEL_ORDERED_PILOT_AND_NOT_
        INFORMED_PORT and vessel anomaly contains VESSEL_NOT_USED
        _PILOT then
    anomaly ← VESSEL_ORDERED_PILOT_AND_NOT_INFORMED_PORT
                _AND_VESSEL_NOT_USED_PILOT
```

Algorithm 19 Determine the final type of anomaly for vessels

```
    //A8,A4,A5
    else if vessel anomaly contains VESSEL_ORDERED_PILOT_AND_NOT_
          INFORMED_PORT and  vessel anomaly contains VESSEL_NOT_USED
          _PILOT and vessel anomaly contains UNUSUAL_TRIP_PATTERN then
       anomaly ← UNUSUAL_TRIP_PATTERN_AND_VESSEL_ORDERED_PILOT
                    _AND_NOT_INFORMED_PORT_AND_VESSEL_NOT_USED
                    _PILOT
    else
       // this part is a default state and executes when the combination of anomalies is not
       // logical and it is due to the wrong and inconsistent data
      anomaly ←  select one of the anomalies randomly
    end if
    vessel anomaly ← anomaly
  end if
end for
```

APPENDIX D: USER INTERFACE

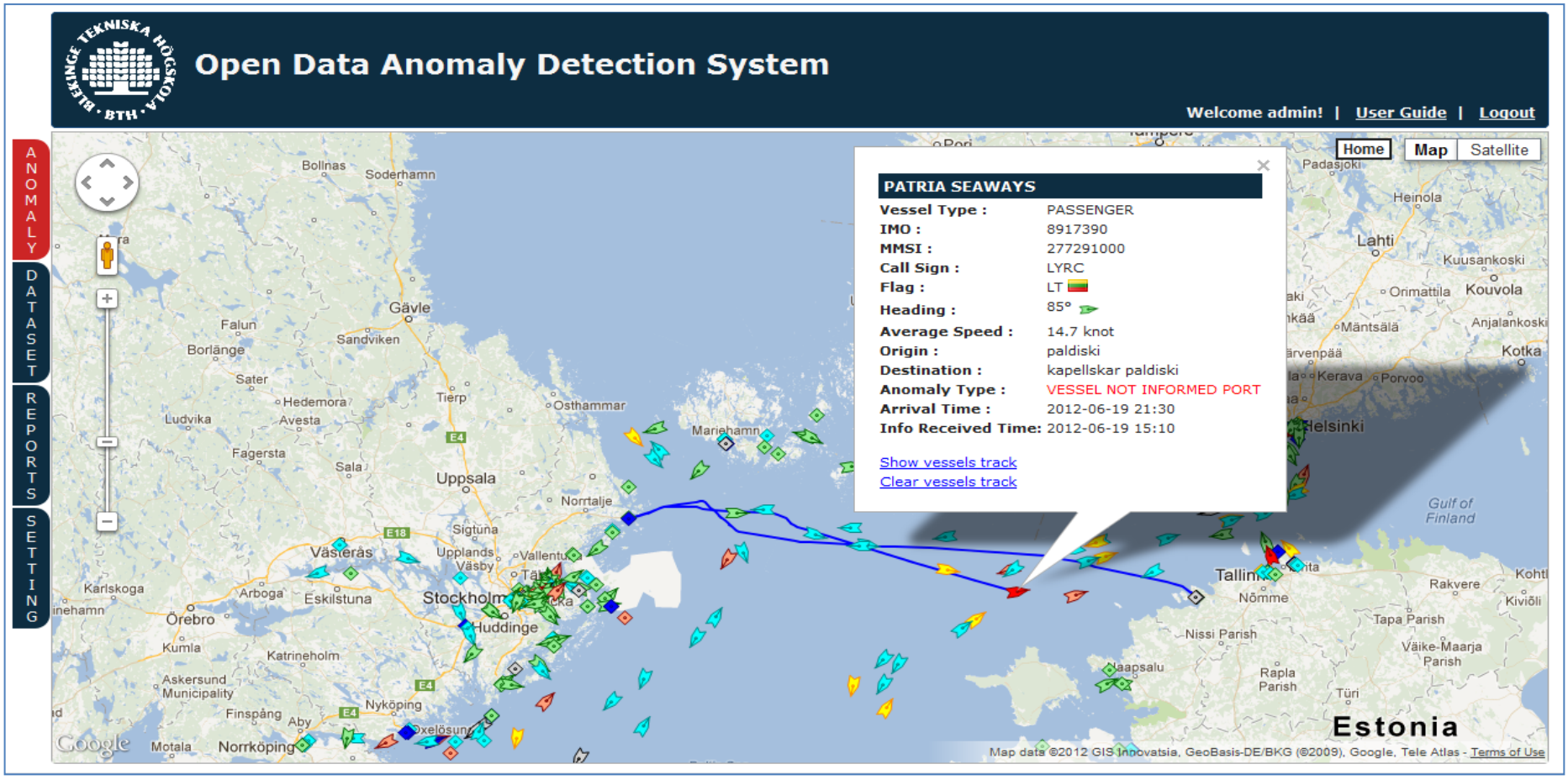

Figure D.1. The main view of the ODADS user interface, the map displays the surveillance area, green vessels are Ferry and Passenger, Cargo vessels are cyan and Tankers are yellow. Grey and dark blue colors belong to Tug and pilot vessels, respectively. The four tabs on the left side contain the data related to detected anomalies, ports and pilot timetables, system reports and settings. In order to inform operators about newly detected anomalies, the *Anomaly* tab on the left side and the anomalous vessels become red. Moreover, previously detected vessels are shown in light red. By clicking on each vessel it is possible to view more information about the vessel and its track.

APPENDIX E: ACRONYMS

AD	Anomaly Detection
ADS	Automatic Dependant Surveillance
AIS	Automatic Identification System
CET	Central European Time
COG	Course Over Ground
D-L	Damerau-Levenshtein
EET	Eastern European Time
EEZ	Exclusive Economic Zone
ELINT	Electronic Intelligence
ETA	Estimated Time of Arrival
FN	False Negatives
FP	False Positives
HCI	Human-Computer Interaction
HFSWR	High Frequency Surface Wave Radar
IMO	International Maritime Organization
IR	Infrared
JDL	Joint Directors of Laboratories
MDA	Maritime Domain Awareness
MMSI	Maritime Mobile Service Identity
MS	Maritime Surveillance
ODADS	Open Data Anomaly Detection System
OTH	Over The Horizon
RADAR	RAdio Detection And Ranging
RMP	Recognized Maritime Picture
SAR	Synthetic Aperture Radar
SIGINT	SIGnal INTelligence
TN	True Negatives
TP	True Positives
UTC	Coordinated Universal Time
VTS	Vessel Traffic Service

REFERENCES

Akselrod, D., Tharmarasa, R., Kirubarajan, T., Zhen Ding, & Ponsford, T. (2009). Multisensor-multitarget tracking testbed. *Proceedings of the IEEE Symposium: Computational Intelligence for Security and Defence Applications, Canada*, 1-6. doi:10.1109/CISDA.2009 .5356526

Alonso, J. M., Ambur, O., Amutio, M. A., Azañón, O., Bennett, D., Flagg, R., McAllister, D., et al. (2009, May 12). Improving access to government through better use of the web. Retrieved from World Wide Web Consortium website: www.w3.org/TR/ egov-improving/

Andersson, M., & Johansson, R. (2010, November). *Multiple sensor fusion for effective abnormal behaviour detection in counter-piracy operations.* Paper presented at the International Waterside Security Conference, Carrara, Italy. doi:10.1109/WSSC.2010 .5730221

Andler, S. F., Fredin, M., Gustavsson, P. M., van Laere, J., Nilsson, M., & Svenson, P. (2009). SMARTracIn: A concept for spoof resistant tracking of vessels and detection of adverse intentions. *Proceedings of the SPIE - The International Society for Optical Engineering, USA, 7305,* 73050G-1-73050G-9. doi:10.1117/12.818567

Avdic, A., Hedström, K., Rose, J., & Grönlund, Å. (Eds.). (2007). *Understanding eparticipation : Contemporary PhD eParticipation Research in Europe.* Örebro: Örebro universitet.

Bick, E. T., & Barock, R. T. (2005). CENTURION harbor surveillance test bed. *Proceedings of MTS/IEEE OCEANS, USA,* 2, 1358-1363. doi:10.1109/OCEANS.2005. 1639943

Brax, C., Niklasson, L., & Smedberg, M. (2008). Finding behavioural anomalies in public areas using video surveillance data. *Proceedings of the Eleventh International Conference on Information Fusion, Germany*. doi:10.1109/ICIF.2008.4632410

Carthel, C., Coraluppi, S., & Grignan, P. (2007). Multisensor tracking and fusion for maritime surveillance. *Proceedings of the Tenth International Conference on Information Fusion, Canada.* doi:10.1109/ICIF.2007.4408025

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 15:1–15:58. doi:10.1145/1541880.1541882

Constantine, L. L., & Lockwood, L. A. D. (1999). *Software for use: a practical guide to the models and methods of usage-centered design*. Addison Wesley.

Dahlbom, A., & Niklasson, L. (2007). Trajectory clustering for coastal surveillance. *Proceedings of the Tenth International Conference on Information Fusion, Canada.* doi:10.1109/ICIF .2007.4408114

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM, 7*(3), 171–176. doi:10.1145/363958.363994

Danu, D., Sinha, A., Kirubarajan, T., Farooq, M., & Brookes, D. (2007). Fusion of over-the-horizon radar and automatic identification systems for overall maritime picture. *Proceedings of the Tenth International Conference on Information Fusion, Canada.* doi:10.1109/ICIF.2007. 4408147

Dedijer, S., & Jéquier, N. (1987). *Intelligence for economic development : an inquiry into the role of the knowledge industry*. Oxford: Berg.

Di Lallo, A., Farina, A., Fulconi, R., Stile, A., Timmoneri, L., & Vigilante, D. (2006, October). *A real time test bed for 2D and 3D multi-radar tracking and data fusion with application to border control.* Paper presented at the CIE International Conference on Radar, Shanghai, China. doi:10.1109/ICR.2006.343162

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, P., Tait, J., & Zijlstra, T. (2009). *Open data handbook*. Retrieved from opendatahandbook.org/en/

Ding, Z., Kannappan, G., Benameur, K., Kirubarajan, T., & Farooq, M. (2003). Wide area integrated maritime surveillance: An updated architecture with data fusion. *Proceedings of the Sixth International Conference of Information Fusion, Australia, 2*, 1324-1333. doi:10.1109/ICIF.2003.177391

Edlund, J., Gronkvist, M., Lingvall, A., & Sviestins, E. (2006). Rule-based situation assessment for sea surveillance. *Proceedings of the SPIE - The International Society for Optical Engineering, USA, 6242*, 624203-1-624203-11. doi:10.1117/12.664410

Fooladvandi, F., Brax, C., Gustavsson, P., & Fredin, M. (2009). Signature-based activity detection based on Bayesian networks acquired from expert knowledge. *Proceedings of the Twelfth International Conference on Information Fusion, USA,* 436-443.

Gad, A. S. (2009). A fuzzy logic-based multisensor data fusion for maritime surveillance - real data testing. *Proceedings of the Twenty-Sixth National Radio Science Conference, Egypt.*

Giompapa, S., Farina, A., Gini, F., Graziano, A., & Di Stefano, R. (2007, April). *Computer simulation of an integrated multi-sensor system for maritime border control.* Paper presented at the IEEE Radar Conference, Boston, MA, USA, 308-313. doi:10.1109/ RADAR.2007.374233

Giompapa, S., Farina, A., Gini, F., Graziano, A., Croci, R., & Di Stefano, R. (2008, May). *Study of the classification task into an integrated multisensor system for maritime border control*. Paper presented at the IEEE Radar Conference, Rome, Italy. doi:10.1109/RADAR.2008 .4720807

Guerriero, M., Willett, P., Coraluppi, S., & Carthel, C. (2008). Radar/AIS data fusion and SAR tasking for maritime surveillance. *Proceedings of the Eleventh International Conference on Information Fusion, Germany*. doi.10.1109/ICIF.2008.4632409

Guyard, A. ., Roy, J., & Defence R&D Canada-Valcartier, V. Q. (2009). Towards case-based reasoning for maritime anomaly detection: A positioning paper. *Proceedings of the Twelfth IASTED International Conference on Intelligent Systems and Control*. USA.

Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE, 85(1)*, 6-23. doi:10.1109/5.554205

Hall, D. L., & McMullen, S. A. H. (2004). *Mathematical techniques in multisensor data fusion*. Boston, London: Artech House.

Hall, M. J., Hall, S. A., & Tate, T. (2000). Removing the HCI bottleneck: how the human computer interface (HCI) affect the performance of data fusion systems. *MSS National Symposium on Sensor and Data Fusion* (pp. 89-104). USA.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*. USA: Elsevier.

Hatch, M. D., Kaina, J. L., Mahler, R. P., & Myre, R. S. (1998, November). *Data fusion methodologies to support theater level and deployable surveillance systems*. Paper presented at the Conference Record of Thirty-Second Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 1, 563-567. doi:10.1109/ACSSC.1998 .750926

Henrich, W., Kausch, T., & Opitz, F. (2004). Data fusion for the finnish fast attack craft squadron 2000: Concept and architecture. *Proceedings of the Seventh International Conference on Information Fusion, Sweden, 2,* 842-847.

İnce, A. N., Topuz, E., & Panayirci, E. (1999). *Principles of integrated maritime surveillance systems*. Boston, Dordrecht, London: Springer.

Jaro, M. A. (1972). UNIMATCH: A computer system for generalized record linkage under conditions of uncertainty. *Proceedings of the spring joint computer conference, USA,* 523–530. doi:10.1145/1478873.1478943

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association, 84*(406), 414-420. doi:10.2307/2289924

Johansson, F., & Falkman, G. (2007, December). *Detection of vessel anomalies - a Bayesian network approach.* Paper presented at the Third International Conference on Intelligent Sensors, Sensor Networks and Information, Melbourne, Qld., Australia, 395-400. doi:10.1109/ISSNIP.2007.4496876

Jouan, A., Valin, P., Gagnon, L., & Bosse, E. (1999). Airborne fusion of imaging and non-imaging sensor information for maritime surveillance. *Proceedings of the Conference on Quality Control by Artificial Vision, Canada,* 281-286.

Lane, R. O., Nevell, D. A., Hayward, S. D., & Beaney, T. W. (2010). Maritime anomaly detection and threat assessment. *Proceedings of the Thirteenth International Conference on Information Fusion, UK*.

Laxhammar, R. (2008). Anomaly detection for sea surveillance. *Proceedings of the Eleventh International Conference on Information Fusion, Germany.*

Lefebvre, E., & Helleur, C. (2001). Multisource information adaptive fuzzy logic correlator for recognized maritime picture. *Proceedings of the Fourth International Conference on Information Fusion, Canada, 2,* ThB2-19-24. Retrieved from ftp.isif.org/fusion /proceedings/fusion01CD/fusion/searchengine/pdf/ThB23.pdf

Lefebvre, E., & Helleur, C. (2004). Automated association of track information from sensor sources with non-sensor information in the context of maritime surveillance. *Proceedings of the Seventh International Conference on Information Fusion, Sweden, 2,* 1251-1256.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics-Doklady, 10(8),* 707–710. Retrieved from www.freearchive.org/o/964e741877a3ffa8f2195ee253597fbaeed69a207e77df150439802fd370b2f8

Mano, J. P., Georgé, J. P., & Gleizes, M. P. (2010). Adaptive multi-agent system for multi-sensor maritime surveillance. In Y. Demazeau, F. Dignum, J. M. Corchado, & J. B. Pérez (Eds.), *Advances in Intelligent and Soft Computing Series: Vol. 70. Advances in Practical Applications of Agents and Multiagent Systems* (pp. 285-290). Berlin, Heidelberg, Germany: Springer. doi: 10.1007/978-3-642-12384-9_34

Maresca, S., Greco, M., Gini, F., Grasso, R., Coraluppi, S., & Horstmann, J. (2010, June). *Vessel detection and classification: An integrated maritime surveillance system in the Tyrrhenian Sea.* Paper presented at the Second International Workshop on Cognitive Information Processing, Elba, Italy, 40-45. doi:10.1109/CIP.2010.5604209

Molloy, J. C. (2011). The Open Knowledge Foundation: Open Data Means Better Science. *Public Library of Science Biology, 9*(12). doi:10.1371/journal.pbio.1001195

Nilsson, M., van Laere, J., Ziemke, T., & Edlund, J. (2008). Extracting rules from expert operators to support situation awareness in maritime surveillance. *Proceedings of the Eleventh International Conference on Information Fusion, Germany.* doi:10.1109 /ICIF.2008.4632308

Ponsford, A. M., D'Souza, I. A., & Kirubarajan, T. (2009, May). *Surveillance of the 200 nautical mile EEZ using HFSWR in association with a spaced-based AIS interceptor*.

Paper presented at the IEEE Conference on Technologies for Homeland Security, Boston, MA, USA, 87-92. doi:10.1109/THS.2009.5168019

Rhodes, B. J., Bomberger, N. A., Seibert, M., & Waxman, A. M. (2005, October). *Maritime situation monitoring and awareness using learning mechanisms.* Paper presented at the IEEE Military Communications Conference, Atlantic City, NJ, USA, 1, 646-652. doi:10.1109/MILCOM.2005.1605756

Rhodes, B. J., Bomberger, N. A., Seibert, M., & Waxman, A. M. (2006, October). *SeeCoast: Automated port scene understanding facilitated by normalcy learning.* Presented at the IEEE Military Communications Conference, Washington, DC, USA. doi:10.1109 /MILCOM.2006.302306

Ristic, B., La Scala, B., Morelande, M., & Gordon, N. (2008). Statistical analysis of motion patterns in AIS Data: Anomaly detection and motion prediction. *Proceedings of the Eleventh International Conference on Information Fusion, Germany.* doi:10.1109/ICIF .2008.4632190

Riveiro, M., & Falkman, G. (2009). Interactive visualization of normal behavioral models and expert rules for maritime anomaly detection. *Proceedings of the Sixth International Conference on Computer Graphics, Imaging and Visualization, China.* 459-466. doi:10.1109/CGIV.2009.54

Riveiro, M., & Falkman, G. (2010). Supporting the analytical reasoning process in maritime anomaly detection: Evaluation and experimental design. *Proceedings of the Fourteenth International Conference Information Visualisation, UK,* 170-178. doi:10.1109/IV.2010.34

Riveiro, M., Falkman, G., & Ziemke, T. (2008a). Visual analytics for the detection of anomalous maritime behavior. *Proceedings of the Twelfth International Conference Information Visualisation, UK,* 273-279. doi:10.1109/IV.2008.25

Riveiro, M., Falkman, G., & Ziemke, T. (2008b). Improving maritime anomaly detection and situation awareness through interactive visualization. *Proceedings of the Eleventh International Conference on Information Fusion, Germany.*

Riveiro, M., Johansson, F., Falkman, G., & Ziemke, T. (2008). Supporting maritime situation awareness using self organizing maps and gaussian mixture models. *Proceedings of the Tenth Scandinavian Conference on Artificial Intelligence,* 84–91. Retrieved from www.his.se/PageFiles/34259/SCAI2008.pdf

Roy, J. (2008). Anomaly detection in the maritime domain. *Proceedings of SPIE - The International Society for Optical Engineering, USA, 6945,* 69450W-1-69450W-14. doi:10.1117/12.776230

Roy, J. (2010). Rule-based expert system for maritime anomaly detection. *Proceedings of SPIE - The International Society for Optical Engineering, USA, 7666*, 76662N-1-76662N-12. doi:10.1117/12.849131

Roy, J., & Davenport, M. (2010, November). *Exploitation of maritime domain ontologies for anomaly detection and threat analysis*. Paper presented at the Waterside Security Conference, Carrara, Italy. doi:10.1109/WSSC.2010.5730278

Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

van Laere, J., & Nilsson, M. (2009). Evaluation of a workshop to capture knowledge from subject matter experts in maritime surveillance. *Proceedings of the Twelfth International Conference on Information Fusion,USA*, 171-178.

Vespe, M., Sciotti, M., & Battistello, G. (2008). *Multi-sensor autonomous tracking for maritime surveillance.* Paper presented at the International Conference on Radar, Adelaide, SA, Australia, 525-530. doi:10.1109/RADAR.2008.4653980

Vespe, M., Sciotti, M., Burro, F., Battistello, G., & Sorge, S. (2008). *Maritime multi-sensor data association based on geographic and navigational knowledge.* Paper pesented at the IEEE Radar Conference, Rome, Italy. doi:10.1109/RADAR.2008.4720782

Winkler, W. E. (1990). *String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage.* Retrieved from www.eric.ed.gov/PDFS /ED325505.pdf

Yin, R. K. (2009). *Case study research: Design and methods*. Sage Publications.

Zamora, E. M., Po llock, J. J., & Zamora, A. (1981). The use of Trigram Analysis for Spelling Error Detection. *Information Processing and Management* 305–316.