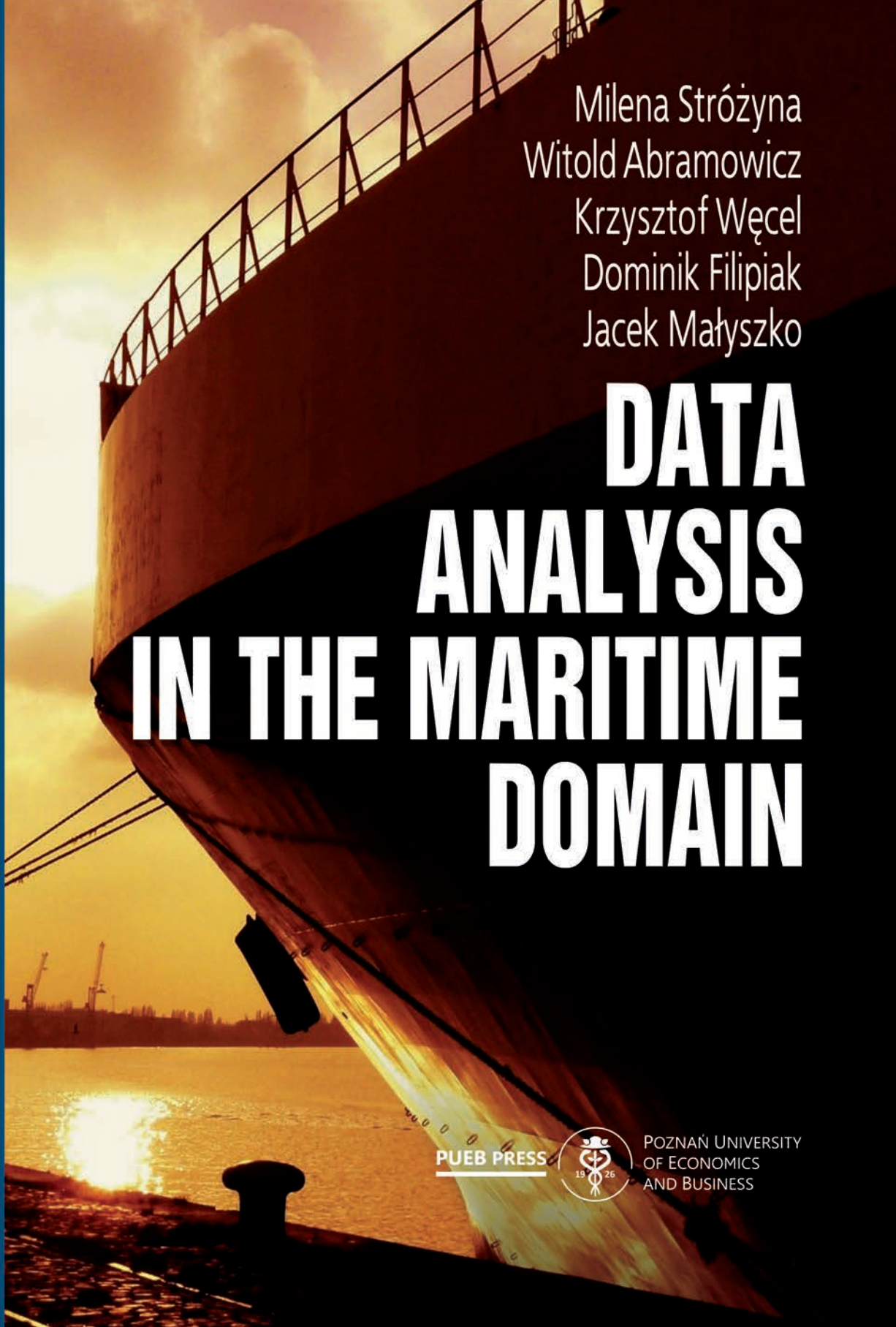


Milena Stróżyna
Witold Abramowicz
Krzysztof Węcel
Dominik Filipiak
Jacek Małyszko

DATA ANALYSIS IN THE MARITIME DOMAIN

PUEB PRESS
Poznań University of Economics and Business

DATA ANALYSIS IN THE MARITIME DOMAIN

Milena Stróżyna
Witold Abramowicz
Krzysztof Węcel
Dominik Filipiak
Jacek Małyszko

DATA ANALYSIS IN THE MARITIME DOMAIN

PUEB PRESS
POZNAŃ UNIVERSITY OF ECONOMICS AND BUSINESS
Poznań 2022

Publication co-financed within the Regional Initiative for Excellence programme of the Minister of Education and Science of Poland, years 2019–2023, grant no. 004/RID/2018/19, financing 3,000,000 PLN.

ISBN 978-83-8211-136-1
eISBN 978-83-8211-137-8
https://doi.org/10.18559/978-83-8211-137-8

POZNAŃ UNIVERSITY OF ECONOMICS AND BUSINESS PRESS
ul. Powstańców Wielkopolskich 16, 61-895 Poznań, Poland
phone: +48 61 854 31 54, 61 854 31 55
www.wydawnictwo.ue.poznan.pl, e-mail: wydawnictwo@ue.poznan.pl
postal address: al. Niepodległości 10, 61-875 Poznań, Poland
Printed and bounded by Poznań University of Economics and Business Print Shop

TABLE OF CONTENTS

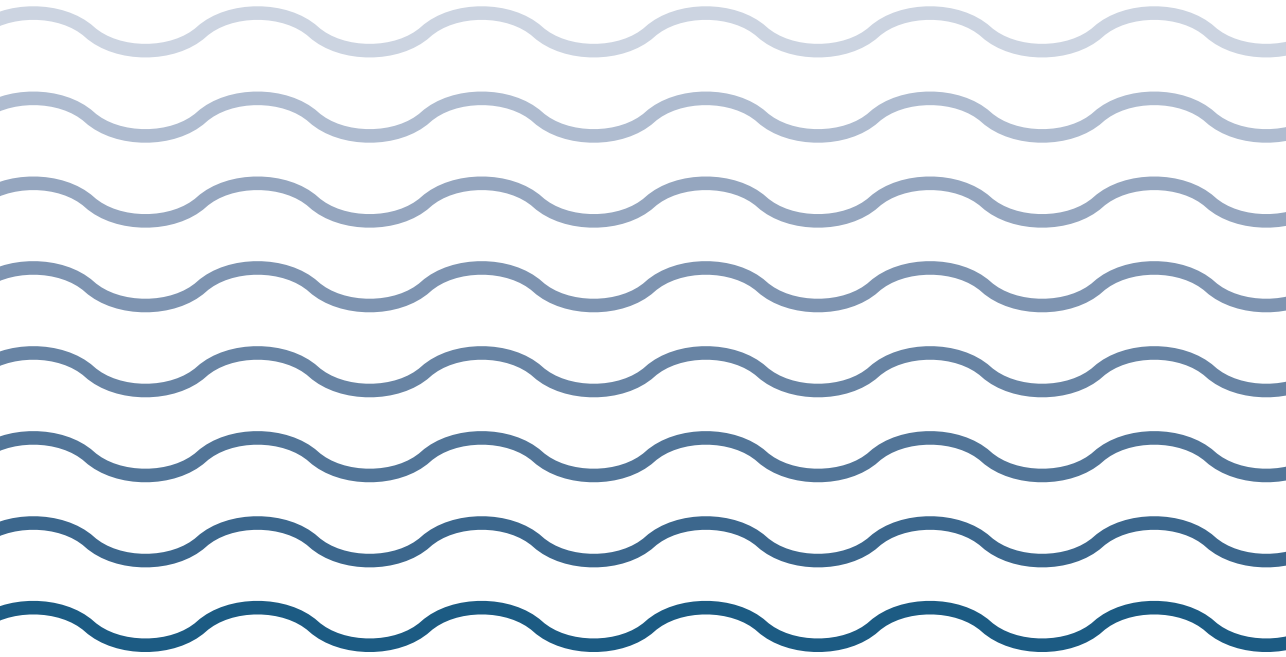
ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to the partners of our research projects for their expert advice and providing data for the research. Firstly, the representatives of LuxSpace Sarl: Gerd Eiden and Miguel Nuevo with whom we had the pleasure to run the SIMMO project. Secondly, our partners from the HANSA project: Axel Hahn, Matthias Steidel, Sebastian Feuerstack, and Arne Lamm from OFFIS, Ståle HØYLANDSKJÆR, Bjørn Åge HJØLLO, and Anders Holme from NAVTOR, Thomas Hahnel and Reinherd Zimmermann from in-innovative navigation GmbH, and Michał Burka, Paweł Kojkoł, Krzysztof Mendalka, and Joanna Haber from Sprint SA. Next, we would like to thank Błażej Lisiecki, Mateusz Jarmużek, and Tomasz Wagner for their technical and programming support. Finally, we would like to thank our colleagues, master's and bachelor's students, and interns from the Poznań University of Economics and Business, who over the years cooperated with us in various forms on the research presented in this book.

Many thanks to the following institutions whose financial support has made it possible to conduct the research: the European Defence Agency and the Contributing Members of the JIP ICET 2 Programme, the MarTERA programme partners: German Federal Ministry of Economic Affairs and Energy (BMWi), Polish National Centre for Research and Development (NCBR) and Research Council of Norway (RCN), and the Minister of Education and Science of Poland.

The research presented in Chapter 2, Chapter 3, Chapter 7, and Chapter 8 has been financed within the Regional Initiative for Excellence programme of the Minister of Education and Science of Poland, years 2019–2023, grant no. 004/RID/2018/19, financing PLN 3,000,000.

Chapter 1



1. INTRODUCTION

Maritime transport plays a key role in transporting goods in modern day economy. It is the backbone of international trade. Nowadays, around 80% of global trade by volume and 70% by value is carried out by sea (UNCTAD, 2017). In this context, the maintenance of high reliability of maritime transport is of prime importance. This, in turn, requires fast access to information about the current situation at sea and in ports as well as appropriate risk management with regard to the implementation of transport services.

With the growing seaborne trade, the usage of maritime areas increases. This leads to the rising number of various maritime threats and anomalous behaviours. The threats encompass behaviours deviating from what is usual, normal, expected, or what is not conforming to the rules and laws in force. We observe a number of dangerous behaviours, such as illegal activities at sea, pollution (oil spills, ballast water, solid-waste from ships), piracy and terrorism, or the trend to register merchant vessels under the "flag of convenience". As a result, the issues of monitoring maritime trade to provide security and safety of ships and cargo gain in importance.

Along with the technological development, new technologies are emerging in the maritime domain that allow to track ships and monitor what is happening on the seas in real time. At the moment, there is a variety of such technologies, starting from sensors, satellite and terrestrial systems that generate maritime data streams which end up in the registries and databases that store ship-related data. All these technologies generate huge amounts of data. Therefore, its analysis, extracting/deriving the relevant information, and finally timely reasoning on a situation based on the analysis results is required to support maritime actors in decision making and provide them with a real time assessment of the situation. Besides, the quality of the available maritime data in many cases is not sufficient and still requires improvement. This is another aspect that needs to be covered.

Having in mind these trends, it is necessary that maritime data is collected and analysed automatically. Thereby, there is a need for novel methods and systems for maritime monitoring, anomaly detection, as well as risk and reliability assessment for maritime transport. In the authors' opinion, the methods and approaches used right now in the maritime domain are not sufficient and do not assure proper effectiveness and efficiency when it comes to the analysis of huge amounts of maritime data. There is still a great potential for further improvement of maritime data analysis capabilities. Moreover, the books published in this domain cover either a single (specific) topic (e.g. ship routing problem) or focus on non-analytical approaches to maritime risk and vulnerability. Thus, the existing literature neither

covers all the aspects indicated above nor provides a comprehensive approach to deal with maritime data, starting with identification and selection of data sources, through data retrieval, fusion and disambiguation, up to state-of-the-art data science methods that would provide appropriate efficiency, effectiveness and quality when it comes to the analysis of big volumes of maritime data. The proposed book aims to explore this research gap.

The main objectives of the book are: (1) develop a theoretical background underlining the available maritime data sources as well as approaches used in data analysis in the areas of maritime stream data analysis, anomaly detection, maritime traffic analysis, and maritime risk assessment; (2) propose novel approaches, tools and methods for maritime data retrieval, fusion and analysis that might be used to detect maritime anomalies or conduct risk assessment as well as can deal with heterogeneous and big volumes of data; (3) verify the proposed methods based on examples and experiments conducted on real maritime data; (4) present real examples how the methods proposed in the book may be used for anomaly detection and risk assessment; (5) show the advantage of the application of various data science methods and big data technologies in different maritime scenarios.

The primary audience for the proposed monograph consists of researchers from the fields of computer science and maritime transport. To some extent it may also attract attention of organizations trying to develop or enhance their information systems with implemented methods for data retrieval and fusion, anomaly detection or maritime risk assessment.

The book is divided into two parts. The first part (non-original) is devoted to the motivation, analysis of literature and available data sources (Chapters 2, 3 and partially Chapters 4, 5, and 6). In the second part (original), the methods, algorithms and systems developed during the study, followed by examples, experiments and evaluation results, are presented (Chapters 7 to 9 and parts of Chapters 4, 5, and 6). The book ends with a summary and discussion (Chapter 10).

Chapter 2 introduces the topics of maritime transport, maritime logistic services, and maritime risk. The aim is to provide a theoretical background to the research and the methods presented in the other chapters. Firstly, the significance of maritime transport in the global economy is discussed, followed by a description of the identified challenges and trends in the maritime domain. Then, maritime logistic services, including aspects related to the service attributes and the service quality, are described, including the concept of reliability of the logistic service as one of the key service quality attributes. Moreover, actors related to the maritime domain, for whom the reliability of maritime transport services and maritime anomaly detection are of prime importance and who thus may be potentially interested in the methods proposed in the book, are characterized. In the last part of this chapter, the issue of maritime transport monitoring is discussed, including a characteristics of the existing systems and methods in this area.

Chapter 3 describes the theoretical background of maritime risk assessment. Here, various approaches, methods and systems for maritime risk are presented. The aim of that chapter is to provide an overview of those methods and systems as well as to analyse the shortcomings of the existing approaches and to present gaps and challenges that still need to be addressed. Based on the critical literature review, a typology of risk factors and variables in the existing methods used in maritime risk assessment is presented.

Chapter 4 elaborates on maritime data. It starts with a presentation of data sources available in the maritime domain, including sensor data as well as open and Internet data sources. Then, the issue of maritime data quality is discussed, followed by a proposal of a framework for the selection of data sources for different analytics purposes. The framework focuses on the process of data enhancement and provides information on how to identify, assess and select the data sources. Further on, proposals of methods for data extraction from the selected sources are presented, including such aspects as data retrieval, fusion, disambiguation and pre-processing. This part of the chapter is then summarized with a description of data sources that are used in the study presented in the following chapters. Finally, a case study of the System for Maritime Monitoring (SIMMO) is provided, what is a real example of how the framework and methods presented earlier in the chapter might be applied in a maritime surveillance system.

Chapter 5 presents the problem of maritime routing and traffic networks. First, an introduction to this topic is provided, followed by a review of the methods that have been developed so far in this research area. The chapter ends with the case study of the HANSA system—an application developed in the project under the same name that extracts maritime traffic patterns based on historical ship movement data and finds an optimal route for a ship's voyage.

Chapter 6 is dedicated to the topic of maritime anomaly detection. It starts with an overview of threats and anomalies that are observed in the maritime domain. Then, based on the conducted literature review, a typology of maritime anomalies is presented and discussed. Finally, a review of the existing approaches and methods for maritime anomaly detection is provided. The chapter ends with a proposal of methods for the detection of anomalies related to loitering on the sea.

The next chapters (7 to 9) present solely the methods developed by the authors for maritime data analysis.

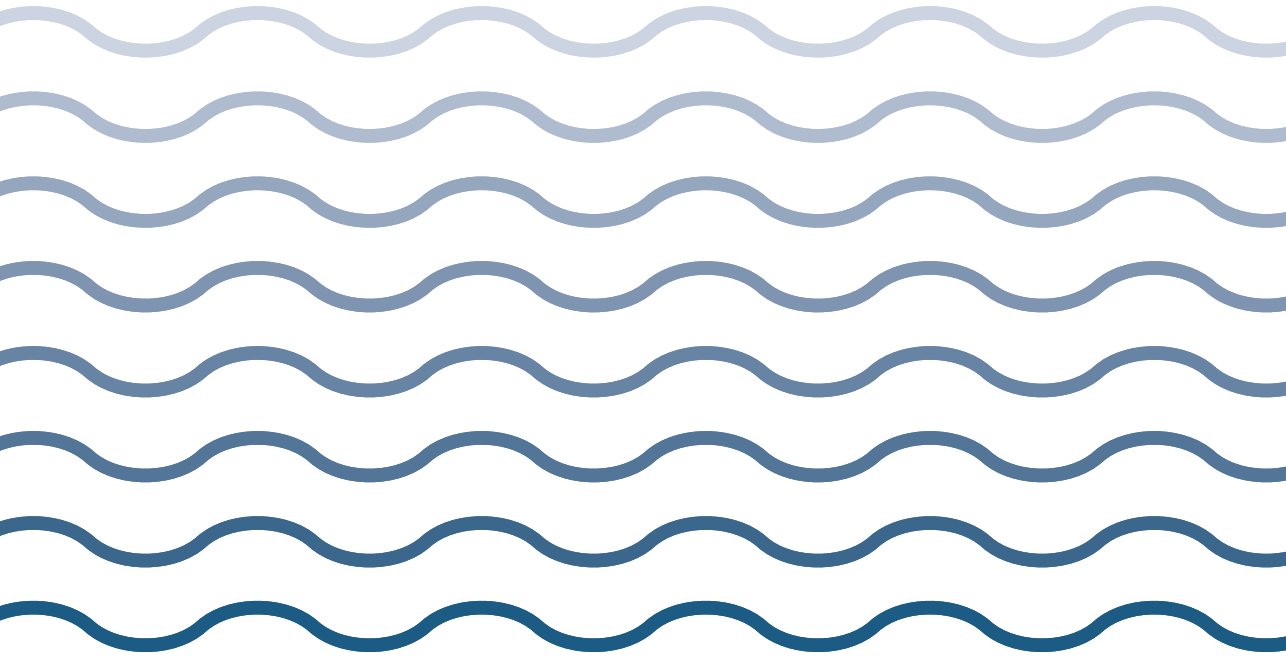
Chapter 7 presents the Short-term Maritime Reliability and Risk Assessment Method (MRRAM). First, the assumptions and concept of the method are provided. The MRRAM method consists of three classifiers that include different variables that may influence the reliability of a delivery being carried out by a given ship. These classifiers are characterized in detail in that chapter. For each classifier, its risk variables are discussed, with a justification why they are significant for reliability assessment. Finally, examples and experiments are presented, showing how the method may be applied, including the results of the evaluation of the method.

In Chapter 8, the method of Ship's Punctuality Prediction (SPP) is presented. It is a method that might be used separately as well as a component of the MRRAM method from the previous chapter. The chapter starts with an outline of the method. Then, in the following sections, the components of the method are described (route prediction, travel time profile determination, influence of additional variables on the punctuality of ships and the final prediction of the punctuality of ships). The chapter ends with a presentation of some applications of the method, including the results of the method evaluation.

The Chapter 9 focuses on the methods for maritime data analysis. The foundation for these methods are big data technologies and state-of-the-art data science algorithms, including machine learning. The presented methods are divided into two categories. Firstly, methods for the detection of static anomalies and loitering-related anomalies are presented. The chapter presents the results of the application of the proposed approach to anomaly detection through AIS data analysis using big data technologies. Then, a comparison of a traditional (SQL-based) approach with a big data-based approach to AIS data analysis is presented to show the advantages of the latter in the process of maritime anomaly detection. Secondly, the developed approach for maritime traffic networks generation, based on historical movement data, is elaborated upon. The approach consists of four methods: the CUSUM algorithm, spatial partitioning of data, the genetic algorithm and finally the method of mesh generation. Details on the implementation of each method are also described along with the obtained results of the evaluation of the methods.

The last part is a summary of the methods and approaches presented in the book, including their results, the final conclusions, the relevance of the study, and suggestions for future work.

Chapter 2



2. MARITIME TRANSPORT AND LOGISTIC SERVICES

2.1. Maritime transport

The development of civilization has always been associated with the movement of people—societies and their members—and goods. The factor that enables this movement is transport.

Transport, in terms of a function, is a technological process that concerns movement of people and goods (cargo). This process can either have a form of a dedicated service, performed for a fee and by specialized companies, or can be carried out as an ancillary activity in relation to other processes (e.g., internal or individual transport) (Rydzkowski & Wojewódzka-Król, 2007, p. 1–4). In a nutshell, transport concerns performing a paid service which results in movement of goods and people from point A to point B, and the supporting services that are connected with the movement (e.g., cargo preparation). The movement itself may consist of carriage/transportation, loading, unloading, and (short-term) storage (Miler, 2015). In vertical terms, transport is divided into road, rail, water, air, and other (e.g., pipeline, cable, spaceflight). In this research, we focus only on a single type of transport—maritime transport.

The development of shipping and maritime means of transport is strictly connected with the development of human civilization and the technological progress. From the dawn of history, maritime transport was used for different purposes, such as personal, economic, and social. Ships have been used for movement of people and goods also for military purposes—as a platform for transporting soldiers, equipment, and weapons. Maritime transport is divided into inland shipping and sea shipping.

Maritime transport offers some advantages in comparison to other modes of transport. The main benefits include (Ficoń, 2010; Miler, 2015): (1) the lowest unit transport costs; (2) low energy consumption for a travel; (3) a high capacity; (4) a high versatility and susceptibility to load cargo; (5) a high specialization of transport means; (6) a high security and ecological standards; (7) easy access to the world's economic markets.

However, it also has some disadvantages. The main shortcomings of maritime transport include: (1) limited availability of some ports and maritime areas; (2) a relatively long travel time; (3) irregularity of cruises and supplies; (4) the need for

transshipment in ports and further transport using other modes; (5) dependency on climate and hydrological conditions; (6) relatively low punctuality.

The importance of the global seaborne trade continued to grow throughout the last century (Asariotis & Benamara, 2012). Since 1945, seaborne trade doubled every decade (el Pozo, Dymock, Feldt, Hebrard, & di Monteforte, 2010). The development of technology made shipping an increasingly efficient and swift mode of transport. It placed it as a leader not only in the transport economics, due to the significant drop in unit costs, but also in terms of safety and reliability of supply, meeting the requirements of ecological standards as well as versatility and comprehensiveness of transport services (Ficoń & Sokołowski, 2012).

In the study, we focus only on the maritime transport that is performed by merchant cargo vessels (seagoing merchant vessels). It includes only ships that transport cargo for hire, such as general cargo vessels, tankers, bulk carriers, and container vessels. This category excludes pleasure craft and boats, warships, passenger and fishing vessels, off shore vessels, high-speed craft, support vessels as well as boats designed for inland and coastal waterways.

The main task of merchant cargo vessels is to provide a transport service that consists in carriage of goods on a specified route, between the place (port) of origin and the place (port) of destination. For this service, a carrier (shipowner) receives a remuneration (freight). For each ship's voyage it is in the interests of the carrier to carry as much cargo as possible, in order to maximize the usage of the vessel's capacity but at the same time to avoid any loss or damage to the transported cargo.

Nowadays, maritime transport is the backbone of international trade and the global economy. The increasing globalization, industrialization, and liberalization of national economies have driven free trade and the growing demand for consumer products. As a result, around 80% of global trade by volume and over 70% by value are carried by sea. In 2016, about 10.6 billion tons of goods were loaded and traveled between 8,000 ports worldwide (UNCTAD, 2017). In 2015, there were more than 89 thousand merchant ships in service, with a combined tonnage of almost 1.7 million deadweight tonnage (DWT) (United Nations, 2015). The flow of dry cargo, including bulk commodities, containerized trade, and general cargo, accounted for 70.3% of total seaborne trade, while tanker trade (crude oil, petroleum products, and gas) was responsible for the remaining 29.7% (UNCTAD, 2017).

Interestingly, with the rising volume and value of the global goods transported by sea, the average distance traveled appears to have remained steady over time—between 1970 and 2008 it accounted on average for 4,100 nautical miles (UNCTAD, 2013). This trend reflects in particular the importance of intraregional trade.

The volume and value of the world maritime trade unambiguously indicates that maritime transport plays an important role in the global economy as well as in the world transportation and shipping system. Its dynamic development, expressed in the steady growth of the world fleet capacity and high adaptability to both the quantitative and qualitative requirements of the commodity markets,

shows that maritime transport keeps up with the needs and challenges of global trade by shipping efficiently and effectively huge streams of goods. Moreover, it creates and ensures a transportation and logistic potential for further growth of the global economy, in terms of technical, operational, economic, and financial dimensions.

Maritime transport actively affects the flow of goods on a global scale. The sphere of its influence is extensive, including not only freight markets and other related transport markets but also other segments of global supply chains. In many cases, the transport process accounts for up to 70–80% of all operations carried out within the supply chain (Grzelakowski, 2012).

Maritime transport is controlled by a set of regulations and legal principles. Historically, the law of the sea is *the freedom of seas*, meaning that "the high seas are open to all states, whether coastal or land-locked" (United Nations, 1982, Article 87). It is a principle that stresses the freedom to navigate the oceans. In addition to this basic rule, there exist various conventions that regulate behavior at sea. One of the basic document is "The Law of the Sea Convention (UNCLOS)", which regulates the rules of shipping, defines the rights and responsibilities of nations with respect to their use of the world's oceans, and establishes guidelines for businesses, environment, and management of marine natural resources.

Another important regulation is "The Convention on the High Seas" from 1958, which introduces the rule of the flag state saying that each commercial vessel must be registered or licensed under a flag. The flag state has the authority and responsibility to enforce regulations over vessels registered under its flag, including those relating to inspection, certification, issuance of safety, and pollution prevention. As a ship operates under the law of its flag state, this law is applicable when the ship is involved in an admiralty case.

Other international maritime regulations that concern the safety and security of maritime shipping are described in Section 3.1.

Although maritime transport has been accompanying people for hundreds of years, it is dynamically developing even now and its importance for the global economy is growing. Moreover, there still exist gaps and challenges that should be addressed. The most important ones from the point of view of this research are described in the next section.

2.2. Trends and challenges in the maritime domain

Along with the growing usage of maritime areas the seas have become a shared, common "good" for humanity that needs worldwide management and protection. The need for regulation and control of the seas has increased for environmental,

economic, safety, and security reasons (el Pozo et al., 2010). The progressive development of globalization and liberalization of the global economy promotes the development of global crime in international trade (T. T. Kaczmarek, 2010, p. 337) resulting in a growing number of threats and anomalies at sea. As indicated by the maritime experts, a key issue is building a better responses to these threats. Trafficking, piracy and terrorism are listed as the most serious areas for the coastguards or NAVY. The experts from the Polish Naval Academy indicate five key factors which are affecting maritime security in the coastal states: (1) port and anchorage crime; (2) domestic instability and civil unrest; (3) political violence; (4) territorial disputes; (5) migration.

They also list the most important risks that European states want to tackle: (1) illegal immigration; (2) smuggling and transnational crime at sea; (3) threats against the freedom of the seas and maritime trade, including energy security; (4) potential expressions of terrorism at sea; (5) degradation of the marine environment; (6) conflicts and crises in the periphery of Europe.

As we can see, a great variety of maritime threats and anomalies is a common problem in the maritime domain. Therefore, they are described in detail in a separate chapter (Chapter 6).

Due to the existence of maritime threats and anomalies Maritime Domain Awareness plays nowadays a critical role. Maritime Domain Awareness (MDA) is "the effective understanding of any activity associated with the maritime environment that could impact upon the security, safety, economy or environment" (International Maritime Organisation, 2013). According to el Pozo et al. (2010), MDA is the *sine qua non* of maritime security and depends on surveillance and exchange of information within the international maritime community. The current capabilities to achieve this awareness are developing but still remain inadequate and poorly coordinated. States are facing a challenge protecting their sovereignty and their infrastructure, countering terrorism and piracy, and detecting illegal activities happening at sea. Still, there also exist other phenomena that influence the security and the reliability of maritime trade.

El Pozo et al. (2010) paid attention to the fact that the majority of merchant ships are from the open registries—the so-called Flag of Convenience (FOC). FOC refers to countries that offer shipowners competitive costs of registration and ships service. They usually do not assure compliance with international safety and security standards, cursorily control the technical condition of ships, and allow hiring foreigners (Ficoń, 2010). Moreover, control by the flag states with the open registries is often ineffective or non-existent. This trend creates an issue for the international maritime community, since the FOC ships pose environmental threat and often are engaged in illegal or criminal activities. Moreover, such illegal activities at sea are not confined to territorial waters or Exclusive Economic Zones of an FOC but occur in international waters or waters belonging to other countries. As a result, the FOC ships present nowadays a significant problem when it comes to

providing protection and security at sea. In order to distinguish an FOC from other registries, flags have been categorized into three colors: black, grey, and white, where black flag concerns a particularly risky country from the point of view of maritime security.

The next problem is the quality of ship crews (qualifications, experience, etc.). In the pursuit of cost reduction, shipowners commonly decrease the number of crew members and the requirements regarding crew qualifications, which also influences the security of ships and the reliability of the services provided by them.

The next issue is operational productivity and effectiveness of the world fleet. Research by Grzelakowski (2009) showed that ship operators generate an overcapacity (tonnage oversupply) in order to accomplish a strategy of flexible and efficient demand fulfillment on the highly competitive freight markets. It means that they keep bigger fleet than they actually use and, as a result, an average ship sails not fully loaded. Moreover, ship operators, usually in response to high oil prices, are interested in reducing the service speed to save fuel. This phenomenon is called slow steaming. Another factor that negatively influences the fleet's productivity, is congestion at ports. As a result, ships' capacity is blocked while queuing.

Because of the existence of various maritime regulations, there is still a need to integrate maritime surveillance on the international level. The Directorate-General for Maritime Affairs and Fisheries (EC) (2010) indicates that in this area, enhancement of the present maritime awareness picture with additional, relevant cross--sectoral and cross-border surveillance data is important issue. This data concern for example illegal activities and threats, impacting both the internal and external EU security. Therefore, the exchange of information in case of imminent threat between various actors working in the maritime domain plays an important role. Another maritime challenges, that according to the Commission should be addressed and solved as soon as possible, include: supporting safe and efficient flow of vessel traffic, early warning and identification of maritime security threats, incidents, accidents as well as monitoring of compliance with regulations on the safety of navigation (vessel traffic safety). In the Commission's opinion, a priority here should be given to an application of a transparent system of main shipping routes, based on analysis of vessel traffic and planned investments in port infrastructure (Hajduk, 2009).

The maritime transport challenges concern also ports and port infrastructure matters. With the growing seaborne trade, ports need to be adapted to handle the increased vessel traffic. There exists a significant performance gap between different ports. There are few very large ports, which serve most of the maritime trade in a region (like Antwerp, Rotterdam, and Hamburg in Europe), while performance of smaller ports is insufficient due to lack of appropriate infrastructure. Such performance gaps produce huge inefficiencies—longer routes, major traffic detours, longer sea trips, and finally more transport emissions. Congestion in ports

or popular canals and a high density of ships at some maritime areas is also an important issue.

High density and congestion may lead in turn to a higher maritime accidents rate, delays of ships, and an increase in maritime risk. Punctuality of ships is another important issue. Statistics show that in practice only 52% of the vessels arrive to a port on time (Vernimmen, Dullaert, & Engelen, 2007) and that average schedule deviations amount to between one and one and a half day (Kim & Lee, 2015). Moreover, in the last year the punctuality further dropped by 8.4%.[1] This creates another challenge for port operators that need to deal with delayed vessels and re-scheduling of the planned port operations. These problems will still have to be faced in the future, since port cargo volumes are expected to rise by 50% by 2030 and even more for the fast growing traffic of containers.

Creation of MDA implies the collection, fusion and dissemination of enormous quantities of data in order to build intelligence and create a comprehensive Common Operating Picture (COP). However, current capabilities to achieve that awareness are still under development, what especially concerns the integration of data from different sources and increase of the quality of maritime-related data. Therefore, the current potential stemming from utilization of this data is not yet fully exploited, particularly in view of data fusion and the use of intelligent data analysis tools.

To fulfill this potential, methods and systems for creating a complete maritime situation picture are required. This includes for examples systems, which integrate static and dynamic data about vessels with information from external sources (further called as ancillary information). Such systems would support operators in charge in the process of monitoring and controlling of the maritime traffic as well as in the OODA loop (Angerman, 2004):

- Observe: to know what is going on;
- Orient: to understand what is going on;
- Decide: to weight the options and their impact;
- Act: to carry out the decision.

According to the maritime experts, there is a number of directions in which the existing surveillance systems should be extended/improved to create a comprehensive maritime picture:

- extension of coverage of the current surveillance and monitoring systems so they would include not only coastline, but also the high sea;
- increase of a frequency of data update;
- provision of other information about ships and marine environment, including vessel's routes;

1. http://www.gospodarkamorska.pl/Porty,Transport/punktualnosc-kontenerowcow-spadla-w-zeszlym-roku.html

- extension of vessel's voyage history to one year;
- inclusion of data and information from additional sources, such as routine surveillance operations and sensors, cued intelligence sensors, open source publications, archived databases, reports published by the maritime community;
- detection of maritime threats and anomalies.

This list can be further extended based on interviews conducted by Riveiro (2011), which revealed a demand for some further improvements:

- integration of data from different sources;
- detection of standard ships routes and distribution of traffic at different times of a day, month, year;
- visualization of typical sea lanes for different types of ships;
- marking ships that require operator's attention;
- listing of vessels on watch (currently suspicious vessels) and their priority.

The trends and challenges presented in this section do not exhaust the catalog of phenomena observed in the maritime domain. Due to limited space, we focused only on the most important ones from the point of view of this research.

The methods presented further in this study aim at addressing some of the issues described in this section, such as FOC, congestion, service speed (including slow steaming), punctuality, maritime threats (like piracy, maritime accidents), anomalies in ship's behavior and, indirectly, an increase of MDA and improvement of the quality of transport services.

2.3. Maritime logistic services

The development of maritime transport services was initiated by the processes of globalization and liberalization that further led to the rise of international trade and the emergence of international supply chains. These processes have required efficient movement of goods between different locations in the world. As a result, transport has become an important aspect. However, we need to bear in mind that transport itself is an element of a wider system—a logistics system.

In a wide context, logistics is defined as a process of management of the whole supply chain of goods or services, starting from the primary resources up to the final customer, in which the standard of customer service is of highest importance (Ficoń, 2010). The core of the logistic activities is to manage and streamline the flow of goods between four basic elements of the supply chain: supply, production, distribution and recycling.

In a traditional approach, the main goal of the supply chain is to achieve a high level of customer service. In order to ensure this the supply chain needs to work efficiently. The main factors that influence this efficiency are Čepinskis and Masteika (2015) and Janvier-James (2012):

- reliability and certainty;
- speed (especially important in the case of seasonal or cyclical demand and perishable or expensive goods);
- delays;
- costs associated with flows within the supply chain, including transport costs;
- relationship with the customers;
- strategic co-operation with partners in the supply chain;
- level and quality of information being shared with the partners.

On the one hand, smooth operation and efficient logistics processes within the supply chain are now being seen as an opportunity to offer added value to customers. Therefore, speed and efficiency of the exchange of goods and information in the supply chain has become a key factor for success and development of markets and entities operating on these markets. A basic element for the efficient flow of physical goods, apart from warehouse management and advanced IT solutions, is transport.

On the other hand, the traditional elements of supply chains, like transport, warehousing and, production, are changing along with the prevalence of digital technologies, which force organizations to re-imagine the way they work—they need to be more flexible and agile in their business operations (Kowalkiewicz, Safrudin, & Schulze, 2017). Enterprises are pressured to provide real-time business and, as a result, aspects such as punctuality, reliability, fast adaptation to changes, customer-centricity, and an opportunity-driven approach gain in importance. Along with the digitalization supply chains adapt to the new trends. In maritime transport it concerns, for example, hyper-connectivity (utilization of various sensors, continuous data exchange through localization and tracking capabilities, the Internet of Things), and cloud-computing (Kowalkiewicz et al., 2017).

Maritime transport is a logistic service and, as other logistics processes in the supply chain, it must be realized effectively and at the right quality level. This quality can be assessed taking into account various measures and attributes that are described in the next section.

2.3.1. Quality of a maritime logistic service

The quality of a logistic service (including maritime transport service) can be assessed taking into account various aspects. Over the years there have been many empirical studies conducted that, based on an analysis of data from firms

in supply chains, presented attributes that significantly influenced the quality of logistic services and proposed models for assessment of this quality. Sahay, Seth, Deshmukh, and Vrat (2006) provided a survey of these studies that showed that there was no agreement on what attributes should be used to measure the service quality. Researchers proposed different attributes for different applications. Examples are presented below.

One of the first studies on the logistic service quality was performed by Matear and Gray (McGinnis, 1989, cited in Matear & Gray 1993). They conducted 11 empirical studies and concluded that reliability was consistently the most important variable in the freight service choice decision. That study showed also that the punctuality aspect of the service was of prime importance in the purchase of both sea and air transport services. Moreover, transit time was frequently more important than freight rates.

Further research was conducted by Matear and Gray (1993), who presented an analysis of factors that influenced the choice of a transport service provider. Based on responses of freight suppliers that purchased sea services, the service factors that influenced the most the choice of a carrier included: punctuality of sea service, availability of freight space, high frequency of sea service, fast response to any problems, value for money, freight rate, arrival and departure time, good relationships with sea carrier. According to Matear and Gray (1993), the three most important service attributes are: 1) fast response to problems; 2) avoidance of loss or damage; and 3) on-time collection and delivery.

Another evaluation of the quality of logistic services was conducted by Franceschini and Rafele (2000). They proposed a list of the main indicators used for evaluation of a logistic service. Among them were:

- lead time: time between the arrival of a customer order and the reception of goods;
- regularity: dispersion around the mean value for the delivery lead time;
- reliability: ratio of the number of orders delivered on time to the total number of orders;
- completeness: ratio of the number of orders delivered in a period of time to the total number of orders delivered in the same period;
- flexibility: ratio of the number of accepted special/urgent/unexpected orders to the total number of special/urgent/unexpected orders;
- correctness: ratio of the number of mistake orders to the total number of orders;
- harmfulness: ratio of the number of 'damaged' orders to the total number of orders.

At the same time, Lu (2000) performed a survey among logistic firms to analyze 33 service attributes. The obtained results distinguish eight important attributes of a maritime service provider: speed and reliability, value-added services (e.g., ability

to provide consolidation/door-to-door/just-in-time service), long-term contractual relationship with other firms working in the maritime domain (e.g., container depots, linear shipping operators), freight rate, equipment and facilities, corporate image, and promotion.

A similar survey was conducted later by Paixão Casaca and Marlow (2005), who performed a study among short sea shipping service customers. Based on the empirical data they analyzed in total 61 service attributes. Then, they prepared a ranking of the six most identifiable attributes of a short sea shipping service. These variables include: carrier's technical capabilities, service quality, the carrier's information technology, innovativeness, the pricing policy and marketing activities. Besides, they indicated the three most important attributes that influenced the final service quality:

- punctuality: notice of cargo availability or delivery to the agent by agreed time;
- regularity: frequency of service;
- safety: provision of safe transport of goods including dangerous ones.

Danielis, Marcucci, and Rotaris (2005) analysed how shippers evaluated and selected maritime transport services. They proposed a model for such evaluation which assumed that decisions were based on four attributes: cost, travel time, reliability (punctuality, risk of delay), and damage and loss. Nowakowski (2011), in turn, divided quality measures of the logistic service into two groups:

- operative measures, such as punctuality of delivery, returns rate, process efficiency, stock turnover ratio, finished products turnover ratio, ROI, total values of inventories, operative costs;
- economic measures, such as materials' costs (price/cost of purchased goods), production costs, inventory costs, transport costs.

The exploratory qualitative study conducted by Sahay et al. (2006), undertaken to investigate the concept of service quality in a supply chain, not the service market, showed that the measures with the largest influence are: percentage of orders delivered in time, net profit in comparison to the productivity ratio, and percentage payments received in time.

Besides, various quality attributes of transport service providers are measured and monitored on a regular basis. Then, a service profile for a given provider, which captures the quality of its service over a certain time-span, may be created (Mutke, Augenstein, Roth, Ludwig, & Franczyk, 2015).

To sum up, the above survey of studies available in the literature, shows that, in fact, there is no standard list of attributes that could be used to assess the quality of a maritime logistic service. Therefore, for the purpose of this study, a coherent list of the attributes proposed by other researchers was defined. The selected attributes are these that were indicated as important factors by most of the studies and include:

- reliability: indicated as one of the most important service attribute by Danielis et al. (2005) and Lu (2000), McGinnis (1989, cited in Matear & Gray 1993);
- punctuality: indicated by Matear and Gray (1993), Paixão Casaca and Marlow (2005), and Sahay et al. (2006);
- travel time: indicated by Danielis et al. (2005), Franceschini and Rafele (2000), and Lu (2000);
- security and safety: indicated by Franceschini and Rafele (2000), Matear and Gray (1993), and Paixão Casaca and Marlow (2005).

Moreover, punctuality, travel time, and security and safety can be an element of reliability, which is defined as the ability to perform the promised service dependably and accurately (Franceschini & Rafele, 2000). The concept of maritime transport service reliability and its relation to punctuality and travel time is presented in the next section. Besides, security and safety may also relate to safe transport of goods without dangerous activities and anomalous behavior that may lead to potential damages or loses of a cargo. This concept is further elaborated in Chapter 6.

2.3.2. Reliability of a maritime logistic service

In general, reliability is a statistical prediction of a desired performance over time. In logistics reliability is referred to the problem of providing delivery of ordered products in a timely and uninterrupted way. It may also be related to the ability of a supply chain to meet customer's requirements, i.e., an uninterrupted flow of goods through the whole supply chain. The reliability of a logistic service includes assurance of the 7R rule (Nowakowski, 2011): Right products, Right quantity (completeness of orders), Right quality (no damage in delivered goods), Right place of delivery, Right time (punctuality of delivery), Right customer (accurate order fulfillment), and Right price (accurate invoicing).

The reliability of the logistic service can be related to (Nowakowski, 2011):

- reliability of the delivery process: defined in relation to punctuality and completeness of deliveries according to customer's requirements;
- reliability of transport: defined as the probability that during a shipment no damage occurs to cargo;
- reliability of logistics infrastructure: defined as the probability of valid support, for example, from a working staff or supporting devices.

Since this study deals with maritime transport services, we will focus only on the reliability of a transport process. This type of reliability may be assessed based on analytical models combined with reliability data, derived from historical performance records (Cross & Ballesio, 2003).

Reliability can be seen in two ways: as a property, described in probabilistic terms (which include a chance of events and processes) or in deterministic terms. For the former, the reliability of an object is understood as its ability to successfully perform a specified task under certain operating conditions and at a given time. The measure here is the probability of performing a task in an assumed time period. In the literature, this measure is called the reliability function (Nowakowski, 2011) and is defined as follows:

$$R(t) = P(T < t_0),$$

where t—time of performing the task (e.g., a supply), T—random time of performing the task (e.g., a supply), t_0—assessed time limit to complete the task.

If we assume that a supply should be performed in a defined time interval (not too soon and not too late), the reliability of the service can be defined as:

$$R(t) = P\left(t_0 - \frac{\triangle t}{2} \leq T < t_0 + \frac{\triangle t}{2}\right),$$

where: $\triangle t$—a defined time interval for performing a task.

Deterministic models are also applied for reliability assessment. In this case reliability may be calculated for example as:

- a ratio of shipments realized with the right fulfillment of the 7R rule to the total number of shipments;
- a ratio of shipments carried out on time to the total number of shipments (punctual share);
- a ratio of damaged transport units to the total number of shipped transport units (damage share).

As indicated in the previous section, reliability is the main attribute that is considered in this research for the assessment of service quality. Another factors, which are parts of reliability assessment, are travel time, punctuality, and transport security.

In the case of a maritime transport service, travel time is defined as the amount of time needed to transport a cargo from port A to port B. In general, the shorter the travel time, the better. It also means *ceteris paribus* a higher chance that a customer will select a transport service provider that offers a shorter travel time.

Punctuality of a maritime transport service is defined as the probability that a ship will complete a delivery at a previously designated time. Punctual is often used synonymously with "on time". Thus, arriving too early or too late may both be perceived as unpunctual.

Determination of a ship's punctuality is connected with estimating its arrival time to a destination port. This is called Estimated Time of Arrival (ETA). Nowadays, estimation of ETA is provided in two ways—either it is a human-based, when a captain or an agent provide this information based on their experience or schedules,

or automatically by applying methods and tools for ETA calculation. A review of methods used to estimate ETA is presented in Section 3.2.3.

The maritime security is frequently defined as the protection from threats at sea. This threats may include crimes such as piracy, armed robbery at sea, trafficking of people and illicit goods, illegal fishing or pollution as well as anomalous behavior of ships that may lead to a threat to health, life, property and marine environment. The security is the status of the sea conditions under which this threat does not exceed the acceptable risk level (Urbański, Morgaś, & Specht, 2008).

The security of the maritime transport occurs when there are three prerequisites met (Abramowicz-Gerigk, Burciu, & Kamiński, 2013): 1) freedom from danger; 2) freedom from unacceptable risk or personal harm; 3) not losing money.

Maritime security can be also considered from the point of view of the supply security. It is defined as the level of guarantee that a cargo shipped by a vessel will be successfully delivered to a customer (recipient of the cargo).

This aspect of the service reliability requires methods for anomalies detection, which are presented in Chapter 6.

The concept of maritime transport service reliability, presented in this section, is also the foundation of supply reliability and a ship's assessment method presented in details in Chapter 7.

2.4. Actors in the maritime supply chains

Maritime transport services are much more complex and complicated than other modes of transport, due to the fact that there are a lot of actors and entities involved. The common actors participating in maritime transport are (Ficoń, 2010): shipowners, forwarding agents, carriers, receivers of goods (consignee, end customers), senders of goods (shipper, loader), agents, ship brokers. In the following paragraphs, all actors are shortly characterized:

- **Shipowner** is responsible for keeping its ships in the required technical and operational condition in order to provide transport services at sea. They gain economic benefits from the exploitation of their vessels from the entities that buy the transport services. They may be interested in performing comparisons with other shipowners (e.g., with regard to the quality and reliability of transport services perceived by customers) in order to determine some attributes of their shipping service (the level of freight, routes, determination of linear shipping services etc.).
- **Forwarding agent** is an entity that professionally and for a fee organizes the whole flow of goods between senders and receivers. They are responsible

for preparing cargo for carriage, transport documentation, cargo insurance, customs formalities etc. In their work, they are interested in knowing the reliability and quality of logistic services while deciding which ship or carrier to choose for cargo shipping.

- **Carrier** is an entity that carries out transport services for a fee. Very often, in maritime transport shipowners are at the same time carriers. Also, chartering is common practice—leasing a ship from the shipowner and its commercial exploitation. In this case, the carrier provides the transport service, not having their own modes of transport. The chartering or leasing of a ship can cover the whole ship or only part of its loading capabilities.
- **Sender** (shipper, loader) is an entity that orders maritime transport services and delivers cargo to a carrier or directly to a shipowner. The sender might be either an exporter of goods, importer of goods, or a forwarding agent itself (if they comprehensively carry out physical logistics activities as well). When a logistic service is being conducted, the sender is interested in knowing whether the goods will be delivered on time and without disruptions, and whether the selected ship is a safe means of transport.
- **Receiver** (consignee) is an entity that is entitled to receive cargo in the destination port. Most often, the receiver of goods is indicated by the sender, and to this end, they have the relevant documents. Upon presentation of these documents, they can pick up the cargo from the ship. The receiver of goods can be any business entity, such as a private or a state-owned enterprise, a public or non-public company, an international corporation, a national institution for foreign trade or an individual customer. Similarly to the sender, they want to know whether the goods will be delivered without disruptions and on time.
- **Logistics companies** are entities that use transport services offered by a carrier of a shipowner when they need to organize a transport of a cargo for their clients and do not own their own ships. They are interested in knowing the delivery time, the punctuality and safety of the selected mean of transport.
- **Maritime authorities** (maritime office, customs services, SAR and other) are responsible for maritime safety and security in a defined maritime area (e.g., port, Exclusive Economic Zone of a country). They are interested in quick identification of suspicious ships that pose a potential threat to the port security (critical infrastructure, continuity of supply etc.) or might be engaged in illegal activities.
- **State authorities** uses the maritime transport to ensure domestic supply of critical and key resources (e.g., oil or gas) and guarantee of security and safety of the country. They want to know whether there are any threats to the supply of these key resources or if there happen any events that may endanger the state security.
- **Ship's crew** is responsible for conducting the shipping from a departure port to a destination port. Their work and the decisions made by a captain may greatly

influence the reliability of the logistic service and its attributes. The crew requires up-to-date information about the current situation at sea (weather conditions, possible maritime threats etc.) in the planned route to make appropriate decisions on the ships' voyage.

The separate group is regulators for the maritime shipping industry. A key regulator for maritime shipping is the International Maritime Organization (IMO) that is a specialized agency of the United Nations (UN) responsible for providing regulations to improve safety and security of international shipping. The sources of maritime regulations are also agreements, resolutions and conventions made by various UN agencies, like the International Labour Organization, EU legislation and the national legislation.

In conclusion, we can see that the list of entities that operate in the maritime domain is quite long. Moreover, each entity has its own mission and goals. As a result, different service attributes can also have a different value for them. Service reliability, including travel time and punctuality, seem to be particularly important for senders and receivers of goods. But since low reliability, a long travel time, and often delays may negatively influence the reputation of a service provider, shipowners and carriers, as well as agents and brokers, are interested in providing a transport service of the best possible quality. The maritime and state authorities in turn are particularly interested in aspects related to maritime threats and detection of anomalies in ships' behavior.

Having this in mind, the methods proposed in the study, that allow for evaluation of a ship's reliability, determination of a ship's short-term punctuality and detection of maritime anomalies, are addressed to all maritime entities mentioned in this section. We believe that these methods can be applied, e.g., in decision support systems dedicated to different maritime users and used in different contexts.

2.5. Maritime transport monitoring

Taking into account trends and challenges presented earlier in this chapter, it is of prime importance to be able to generate a Recognized Maritime Picture (RMP), which is a composite picture of activities in a given maritime area for a given time (Vespe, Sciotti, & Battistello, 2008). Generation of RMP requires timely input from many data sources to determine location, identity, and activity of ships, in order to provide sufficient information to decision makers. As a result, there is a need for development of maritime surveillance systems, which would collect, fuse and analyze various maritime data.

Along with the publication of the IMO regulations and development of sensors and technologies for collecting information about maritime traffic and ship

movement, systems for the marine traffic monitoring started to emerge. Today the basic pillar for these systems are coastal radars, data provided by marine patrol boats or aircrafts, cameras located in ports, VHF radio, meteo and hydro information, databases with historical information (e.g., Lloyds, SafeSeaNet, Equasis) and internal databases with historical comments and alerts on ships (Riveiro, 2011).

Along with the development of satellite technologies and location-based services, most surveillance systems use also two other sources of vessel tracking data: 1) Long Range Identification and Tracking (LRIT), 2) Automatic Identification System (both described in details in Chapter 4).

At the beginning, there were no coordinated activities related to development of such systems on the international level. As a result, various systems have been developed, with different architectures, sensors, and application areas. However, in the recent years, a trend to unify and standardize the systems emerged, resulting in three basic types of the marine traffic monitoring systems (Miler, 2015):

- Vessel Traffic Monitoring and Information System (VTMIS), which has been developed at a national and a regional (EU) level;
- integrated information systems that concern the security and safety of shipping, e.g., SWIBŻ in Poland, BRITE in NATO, IMDatE (Integrated Maritime Data Environment in EMSA);
- commercial information systems, whose aim is to support maritime operators in management of the sea-land logistic chains.

The national VTMIS include:

- Automatic Identification System (AIS), described in details in Chapter 4;
- Long Range Identification and Tracking (LRIT), described in details in Chapter 4;
- Vessel Traffic Services (VTS), including radars, closed-circuit television (CCTV), VHF radio telephony for communication with vessels, and management center of VTS; some VTS include also additional modules for analytics;
- Ship Reporting Systems (SRS), which concerns the rules for ships regarding the reporting and notification procedure; it defines when, to whom, what kind of information, and in what format should be sent by a ship; the typical reports/notifications include: itinerary, position report, report on dangerous cargo/harmful or hazardous substances or pollution;
- Maritime Assistance Services (MAS), which supports activities at sea in case of collision or accident;
- National SafeSeaNet (SSN) and CleanSeaNet (CSN) systems, the European platforms for maritime data exchange on ships (SSN) and oil spills (CSN).

There are also systems that support VTMIS, such as: DGPS, satellite telecommunication systems like INMARSAT, VHF communication system.

There are 2 types of VTS: Coastal VTS and Port VTS (Riveiro, 2011). Main services provided by VTS include: information, traffic organization, and navigational assistance services. However, each VTS can have a little bit different tasks and can use different systems. In general, many of them use only basic data sources and maritime surveillance systems, like LRIT or AIS. The work of maritime operators in these VTS consist mainly of two types of activities: reactive (providing information requested by a ship) and proactive (searching for deviations from normal behavior and contacting the ship, if something suspicious is detected). However, discovering of anomalous behavior of ships is performed mainly manually (continuous observation of the map), and it relies mainly on the operator's experience. Both types of activities performed by VTS operators require appropriate information, which sometimes is missing or need to be searched for in different sources. Moreover, not all systems include a module or tool for automatic detection of anomalies provided. Some examples of such systems are described below in this section.

Despite the fact that the existing system already provide some essential information about ships, there is a lot more relevant information necessary to generate a comprehensive maritime picture. This information is available, but often diffused across different systems and sources. One of such data source is the Internet, where relevant maritime information is available on various web pages. As a result, this additional information is currently rarely exploited due to lack of integration. This is a weak point of the existing systems. In the Internet, there is a lot of maritime--related data freely available and accessible, which can be extracted, integrated and used in analysis and decision making.

In this regard, maritime operators complain about some features (or lack thereof) in the existing surveillance systems, such as lack of additional information about ships (e.g., owner, flag, historical data), or no integration between different system used (e.g., radar, AIS, meteo).

There are also commercial systems for monitoring the maritime traffic, e.g., SARGOS, GreenLine, CATE, SADV. Some of them can be regarded as a Decision Support System (DSS). DSS is a computer-based information system that supports business or organizational decision-making activities. DSS can be either fully computerized, human, or a combination of both.

SARGOS[2] (Système d'Alerte et de Réponse Graduée OffShore/Graduated Offshore Response Alert System) was developed to protect offshore infrastructures against threats (Giraud et al., 2011). It has been supported by the French National Research Agency in the frame of global safety programme and aims at threats detection, threats evaluation, displaying risks, and response planning. SARGOS operates on the innovative Frequency Modulated Continuous Wave (FMCW). It is based on AIS, navigation radars and infra-red sensors, and provides inference pro-

2. http://en.sofresud.com/Maritime-Surveillance/SARGOS

cess using Bayesian Networks in risk assessment, which was described in (Bouejla, Chaze, Guarnieri, & Napoli, 2014; Chaze et al., 2012). The Detection module is responsible for information collection, which is used for calculating potential risks. A possible avoidance of these risks is suggested by Reactions module along with Means Management. Visualization and Record & Replay modules provide access to graphical analysis of the maritime situation and can be used for distinguishing real threats from false alarms.

GreenLine Vessel Selection System[3] (VSS) is a commercial system, which aims at supporting decision-making process in the maritime domain. It combines data from multiple sources. An automatic risk assessment is determined by a rule-based scoring system, which is highly customizable; each threat can be prioritized manually and there is a possibility to create new rules.

CATE[4] (Computer Assisted Threat Evaluation) is a Maritime Domain Awareness system and consists of four components: Threat Evaluation, Sense Making, Situational Awareness, Knowledge Management. It combines radar, AIS, LRIT, imagery and open data, and provides rule-based real-time risk/threat analysis. CATE is designed in Service Oriented Architecture.

Statistical Anomaly Detection and Visualisation (SADV) for Maritime Domain Awareness[5] is a Swedish project, which aims at providing an advanced anomaly detection.

Another examples are Maritime Situational Awareness (MSA)[6] developed for NATO nations or GeMASS (GEnetic algorithm knowledge discovery for MAritime Security System) (C.-H. Chen, Khoo, Chong, & Yin, 2014). GeMASS consists of modules for data pre-processing (raw AIS data translation), real-time ship analysis and components for decision/result update (for obtaining training datasets), knowledge discovery and data post-processing (for data accumulation).

The analysis of the existing systems shows that there is a growing number of maritime surveillance systems that offer functions for threats/anomalies detection and risk assessment. An important success factor for all described maritime surveillance systems is merging data from many different sources. This concept is called data fusion. Data fusion is a challenging task, since there are many issues arising from the data to be fused, such as data imperfection, correlation, inconsistency, disparateness and ambiguity. The solutions described above focus on providing data fusion techniques, which combine mainly sensor data such as AIS, VTS, radars or video cameras (Kazemi, Abghari, Lavesson, Johnson, & Ryman, 2013). A more sophisticated approach, which assumes enrichment of sensor data with open data, data available in various databases or data stored in structured or unstructured

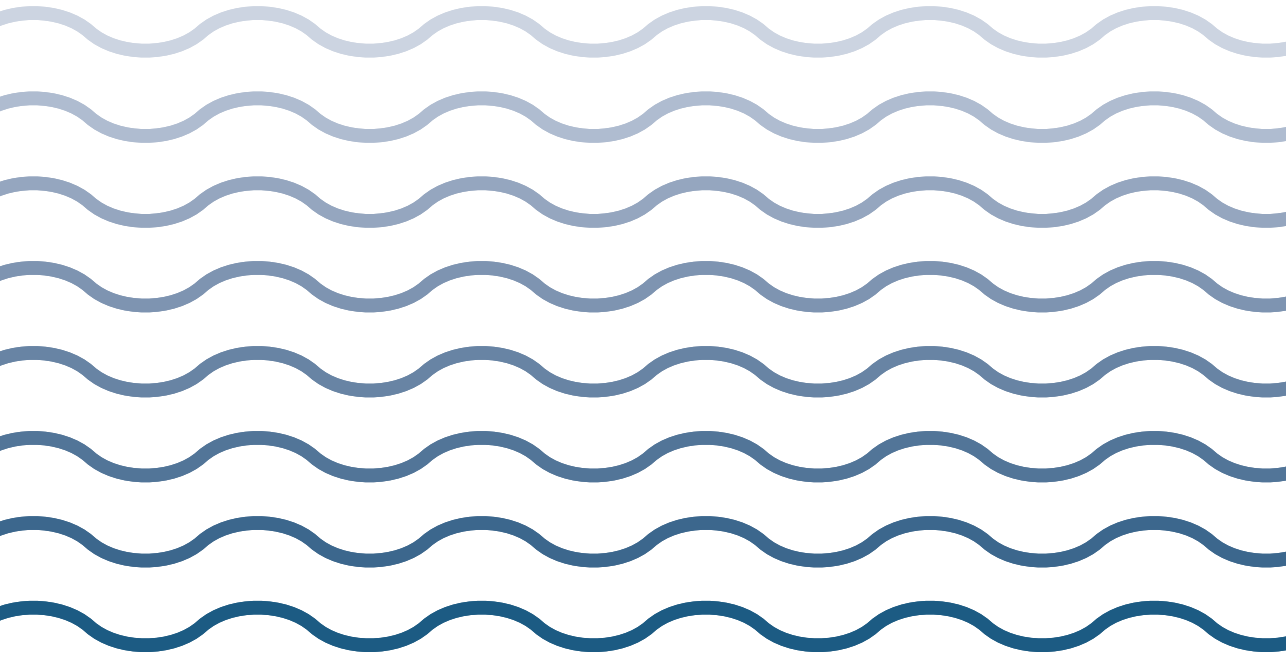
3. http://www.greenlinesystems.com/vessel-risk-targeting/
4. http://www.channellogistics.com/images/mhsSummit2009.pdf
5. https://www.sics.se/projects/sadv
6. http://www.cmre.nato.int/research/maritime-situational-awareness

documents (e.g., Web pages, historical reports and comments on ships behaviors) is still missing (Brax, 2011).

Besides all presented solutions are either dedicated for the maritime authorities (e.g., port authorities), without access to the data for external entities, or commercial. To the best of our knowledge, there is no open system for maritime data gathering and analysis. Besides, only some of the systems offer functionality for risk assessment; the main goal of such systems is rather to provide information about current situation at sea and to monitor traffic/movements of ships.

There is also shortage of the maritime DSS. In the maritime domain, mainly human DSS have been developed. It means that the results of information and knowledge discovery provided by application of various analytics methods (e.g., risk assessment or anomaly detection), are intended as a basis for human decision-making (Riveiro, 2011). The maritime DSS concern above all ship routing and scheduling, or navigational matters (see e.g., Fagerholt, 2004; Fagerholt & Lindstad, 2007). The latter is a component of the maritime intelligent transport system, which supports the process of ship conduct. According to Pietrzykowski (2011) development of such systems will be going towards DSS—intelligent navigational advisory systems, which apart from information functions would provide hazard identification in ship's movement, warning against hazards and generation of recommendations. This would result in supporting maritime users in the risk analysis process and maritime surveillance.

Chapter 3



3. MARITIME RISK ASSESSMENT

3.1. Maritime risk and reliability

In general, a risk can be perceived as a potential harm from an unforeseen event. It is a threat that something might happen to disrupt normal activities or stop things happening as planned (Waters, 2011). For example, there is a risk that a new product will not sell as well as expected, that a delivery to a customer will be delayed, or a supplier will go bankrupt.

Risk itself can also be perceived twofold. When speaking about risk most people think about a danger, a harm, a threat, or a state that can lead to a loss. (Stemmler, 2007) says that "risk denotes the chance of danger, loss or injury". Similarly, the Royal Society (Royal Society Study Group, 1983) describes the risk as "the probability that a particular adverse event occurs during a stated period of time".

However, there is a group of people, who say that risk can also be positively beneficial. Since the classic principle of economics says that profit is a reward for taking risks, then the greater the risk, the greater the potential profit. As a result, risk management should not necessarily try to eliminate or minimize risk, but it can also search for opportunities offered by uncertainty (Waters, 2011).

In this study we focus solely on the first approach to risk—risk is perceived as a potential harm due to unforeseen events.

When analyzing risk, some of its features shall be taken into account. Risk:

- is heterogeneous;
- may be objective and subjective;
- depends on the context;
- is dynamic and is influenced by many factors (which are dependent or independent);
- is a process rather than a state.

The general definition of risk says that it is based on the probability of an undesired event, where probability is a measure of likelihood, relative frequency or proportion of times this event occurs. Risk (R) might be calculated as a product of the value of the probability (P) of an event, its duration (E) and its effects/severity (S):

$$R = P \times E \times S$$

Risk takes a value in the range from 0 to 1.

There are three ways of finding the probability of an event:

(1) estimation—expert knowledge is used to calculate theoretical (*a priori*) probability:

$$\text{Probability of an event} = \frac{\text{number of ways the event can occur}}{\text{total number of possible outcomes}};$$

(2) observation—historical data is used to see how often an event actually happened in the past, and use this information to give an experimental or empirical probability:

$$\text{Probability of an event} = \frac{\text{number of occurrences of the event}}{\text{total number of observations}};$$

(3) subjective estimates—the likelihood of an event is based purely on a researcher's opinion. This method is, however, not recommended since the results are notoriously unreliable as they rely solely on the researcher's judgment and opinion.

In this research, the second approach to finding the probability of an undesired event and risk calculation is adopted.

3.1.1. Risk management

Risk management is a broad function for dealing with risks (Waters, 2011). The three core activities within risk management are:

(1) risk identification: finding events that may occur;
(2) analysis of consequences: finding the likelihood of events and possible harm (or benefit);
(3) designing appropriate responses: defining alternatives and assessing their relative merits.

According to ISO 31000:2009, risk management is an important element of the decision making process, since it creates a foundation for collecting appropriate data and information required to make a decision. Access to this data enables to estimate and plan future activities and define expected results. As a result, the process of risk management provides support for decision makers.

There are three steps of risk management (Szymanek, 2008):

(1) risk analysis: identification of threats, frequency estimation, and risk assessment;
(2) risk evaluation: establishing an acceptable risk level;

(3) risk control: achievement of an acceptable risk level under the prevailing economic and social constraints.

Risk analysis is about understanding the essence of risk and providing required data and information for risk estimation and decision making. This step is indispensable and crucial for risk analysis (Jarysz-Kamińska, 2013, p. 241). Assessment of risk allows organizations to carry out preventive actions and implement contingency procedures. Risk analysis includes identification of risk events, definition of their results and probability of their occurrence (using qualitative or quantitative measures).

When the possible risk factors are identified and analyzed, then the risk evaluation step can take place. Here the main issue is to select an appropriate risk evaluation criterion and define the acceptable level of risk. The identified risk factors can be divided into three categories:

- factors which the entity can influence (e.g., operational errors of employees);
- factors which the entity can influence only partially (e.g., the risk of a supplier);
- factors which the entity practically cannot influence (e.g., changes in commodity prices).

Then, for each risk factor one of the three strategies can be adopted: passive (acceptance of the risk without any further activities to reduce it), active (prevention and monitoring), and reactive (impact on effects of an adverse event that has happened).

In the evaluation step an estimated risk is being related to a defined risk threshold—the acceptable risk level (Szymanek, 2008). In this approach, the risk level is divided into three regions: 1) acceptable risk; 2) tolerable risk; 3) unacceptable risk. The level of estimated risk is located between acceptable risk and unacceptable risk, which should not be exceeded. Between these two regions there is tolerable risk, which is often called ALARP—As Low As Reasonably Practicable (Health and Safety Executive, 2015).

The last step of risk management is risk control. It is about taking actions in order to mitigate the risk and to monitor the current risk level for various scenarios on a regular basis. The actions can be preventive (with the aim of reducing the probability of an adverse event), corrective (fixing the results), prescriptive (aiming at avoiding an adverse event), and detecting (identification of negative effects which have already happened).

The relevant point for risk management is that our knowledge of a situation changes over time and the level of uncertainty changes as well. Typically, we may have very little information about a problem in advance; then as time passes we learn more and get new information. As a result, the level of risk and the probability of an undesirable event is updated. Therefore, the Bayes' theorem and Bayesian inference might be used here, which allow for an estimation of the probability of

an event based on prior knowledge of conditions that might be related to the event. It also means that risk is not constant but changes over time. This, in turn, leads to the conclusion that risk management is never completed but is a continuous process.

Having this in mind, a dynamic approach for assessing risk and for evaluation of service reliability is adopted in this study. It assumes that the level of risk for a given transport service (ship voyage) can be updated as new information appears.

3.1.2. Transport risk

Research on risk assessment has a long history in various domains, like health, banking, finance, insurance, or project management. Nowadays, risk is the subject of research in management, economy, marketing, but also logistics.

Logistic risk, in a broad perspective, occurs as a result of mistakes or errors in the area of supply, production, and distribution, including transport. The transportation process is exposed to different types of risk due to the occurrence of various threats and undesired events that may happen. There are basically two kinds of risk to a transportation process (and a supply chain in a broader perspective) (Waters, 2011):

- internal risk that appears in normal operations, such as late deliveries, minor accidents, human errors, etc.;
- external risk that comes from outside, such as hurricanes, wars, terrorist attacks, price rises, etc.

Transport risk appears with any event that might disrupt the planned transportation process on its journey from a supplier to a final customer. These events may happen at any point in a supply chain from the initial supplier to the final customers and can interrupt the supply of materials or the demand for products. Moreover, their effects might be localized in one part of a supply chain, or be passed on as a threat to the whole supply chain (Waters, 2011). Because all members of the supply chain are linked together, they might be affected by events that happened far away and over which they have no control. Risk to one member is automatically transferred to all other members. Transport risk might prevent deliveries, cause delays, damage goods, or somehow affect smooth operations. However, the consequences are generally much broader. A late delivery of raw materials might halt production; it might raise costs by forcing a move to alternative transport, materials, or operations; it might make partners reconsider their trading relationships and lead to a loss of customer trust. In the case of maritime deliveries a late delivery might enforce re-planning of all further port- or terminal-related activities, such as berth assignment or finding storage location for the cargo. Therefore, each organization should define a mechanism for dealing

with unforeseen events, and when something unexpected actually occurs have alternative plans.

As a result, the logistic and transport risk management is particularly important. The appropriate risk assessment and control can lead to the reduction of probability of occurring adverse events and to increase of the processes reliability in the supply chain.

In general, there is a relationship between risk and reliability of a transport process. As defined in the previous section, risk relates to the probability of occurrence of the undesired event. If this event can further involve the unsuccessful performance of a "part" of a supply chain (a transport process), then the reliability of such a "part" may suffer. As a result, the high risk level of a transport process may influence the level of its reliability (Cross & Ballesio, 2003).

Transport risk is multidimensional—it means that we have to take into account many factors that are involved in this process and can influence its realization, such as infrastructure, means of transport, people. Field research conducted among companies (Wieteska, 2011) showed that during risk analysis a special attention is paid to the following risk factors: late deliveries (indicated by 91.63% of the surveyed companies), inadequate technical quality of supply (89.47%), lack of flexibility of supply (71.77%), failure to meet the technical quality requirements (52.87%), random and non-random adverse events (destruction or loss of goods) (57.66%), chance events (fire, storm) (58.61%).

A study conducted by Ferrer, Karlberg, and Hintlian (2007), in turn, showed which risk factors actually impacted business and the supply process. Among the factors, which companies cannot control or can control only partially the authors mentioned: natural disasters (35% of the surveyed companies), political instability (20%), terrorism (13%), port operations and customs clearance delays (23%).

In case of transport services, the main four types of potential risks are (T. T. Kaczmarek, 2012, p. 161–165): (1) risk of delay; (2) local transport risk (deviation from the initially planned mode of delivery or transport route due to some local incidents); (3) risk of cargo loss (total or partial loss); and (4) risk of quality deterioration (e.g., decrease of products' quality due to changes in their physical characteristics).

3.1.3. Maritime risk

Since in this study we focus solely on one mode of transport—maritime transport——the risk related to maritime transport services should be considered.

The concept of maritime risk is defined variously in the literature. Goerlandt and Montewka (2015) presented an overview of risk definitions and various approaches to risk analysis applied in the maritime domain. According to

this review, the risk can be defined as: 1) the expected value of the probability of an event occurrence and the utility of the consequences; 2) the probability of an undesirable event, or the chance of a loss; 3) the uncertainty that is understood either as a probability distribution over an outcome range or as a statistical variation compared with an average value; 4) the possibility of an unfortunate occurrence, which can be further; 5) combined with consequences and their severity.

There are also definitions suggested by relevant authorities or standardization organizations. The International Maritime Organization (IMO), which provided the Formal Safety Assessment methodology, defines maritime risk as "the combination of the frequency and the severity of the consequence." The ISO definition sees risk as "the effect of uncertainty on objectives," however, this approach is not applied in the maritime domain (Goerlandt & Montewka, 2015).

In this work we limit our discussion to the risk of an undesirable event which may threaten the safety of individuals, the environment, or physical assets. As a result, we define maritime risk as the probability of occurrence of an undesirable event or the chance of a loss. This undesirable event may be caused by a ship, due to its features or its behavior as well as due to events that are happening in the operational environment of the ship.

In the classical approach (ABS, 2020; Szymanek, 2008), risk assessment in the maritime domain consists of four basic steps:

(1) definition of undesirable events (threats identification and scenarios definition);
(2) calculation of a probability for each scenario;
(3) calculation of consequences;
(4) risk evaluation—determining whether the level of risk is acceptable.

Examples of potential threats and risk factors that are connected with maritime transport, and which can negatively influence the reliability of a logistic service, are (Wieteska, 2011): errors in transport orders, improper transport conditions (temperature, humidity), failures of control and measurement systems/GPS, improper packaging or protection of goods against mechanical damage, water absorption, improper placement of cargo on a ship, errors in shipping documents, improper technical condition of a ship, failures of a ship, peak seasons resulted in a reduced transport capacity of a carrier and/or increase of a transportation fee, no insurance of cargo, random events (fire, storm, explosions), traffic accidents as well as non-random events encompassing human errors or intentional actions, theft and acts of terror. These threats concern especially the cargo transported, due to a long transportation time.

Along with the increase in seaborne trade and the volume of cargo carried by sea, the need to provide security of ships and cargo has gained in importance. As

a result, a number of codes, conventions, and various acts of international law have emerged which regulate the security issues as well as the performance of maritime transport and the efficiency of transport capabilities usage (some of them were already presented in Section 2.1).

The most active organization in this matter is the International Maritime Organization (IMO). IMO released the main conventions regarding maritime transport and the security of the transported cargo, including SOLAS 74, MARPOL 73/78, COLREGS, STCW, SAR (Miler, 2015).

Another important organization that takes care of maritime security and safety is the International Association of Lighthouse Authorities (IALA).[1] In 2005, IALA published two methods for assessment and control of maritime risk (Miler, 2015):

- Ports and Waterways Safety Assessment (PAWSA): a qualitative method for analysis, assessment and control of risk in ports and waterways;
- IALA Waterways Risk Assessment Programme (IWRAP): a quantitative method for analysis, assessment and control of risk in waterways and straits.

IMO, in turn, has developed one of the most commonly used methodology for maritime risk assessment—the Formal Safety Assessment (FSA). This method is described in detail in Section 3.2.1. Another is the International Ship and Facility Security Code (IPSP Code), which obliges ships "to detect security threats and take preventative measures against security incidents affecting ships" (International Maritime Organisation, 1974). According to the IPSP, ships are obliged to develop and implement an individual risk plan. The plan should be developed by a shipowner and include analysis and assessment of risk for various categories of threats as well as mitigation measures and a plan of actions for the threats which are most likely to happen.

Such a plan is a confidential document and must be accepted by the administration of the flag the ship is flying. It means that other entities or actors working in the maritime domain do not have access to the results of the risk assessment of a particular ship. As a result, if one (e.g., sender or receiver of goods) wants to assess the risk of a given ship, they need to conduct such an assessment on their own or buy such information. This is one reason why access to appropriate information plays an important role in the process of risk assessment. Moreover, various approaches, methods, and tools for maritime risk assessment have been developed that are used by external entities interested in knowing the risk of a given ship or a given event. The research presented in this book is also about providing a solution in this matter.

1. http://www.iala-aism.org/

3.2. Maritime risk assessment systems and methods

3.2.1. Formal Safety Assessment

In the maritime context, there is a rational and systematic risk-based approach for safety assessment—Formal Safety Assessment (FSA) (Berle, Asbjørnslett, & Rice, 2011; Trucco, Cagno, Ruggeri, & Grande, 2008). FSA was developed by the International Maritime Organization, which is the basic international institution responsible for developing and maintaining a comprehensive regulatory framework for shipping, and thus for providing maritime security and safety.

FSA can be applied to specific maritime safety issues in order to identify cost-effective risk reduction options. The FSA process consists of five steps (Berle et al., 2011; Ellis et al., 2008):

(1) hazard identification: identification of all hazards related to the activity / ship;
(2) risk assessment: building a risk model and determining probabilities and consequences for all branches of the risk model;
(3) risk control options: identification of measures to control and reduce the identified risks;
(4) cost benefit assessment: determining cost effectiveness of each risk mitigation option and preparing a ranking for them;
(5) recommendations for decision making: deciding and making a plan of future activities, based on the results of previous steps.

FSA is commonly seen as the premier scientific method for maritime risk analysis and for formulating maritime regulatory policy (Goerlandt & Montewka, 2015). Therefore it was selected as a foundation for the risk and reliability assessment method that will be presented in Chapter 7.

3.2.2. Maritime risk assessment approaches

In the literature, there are many different analysis techniques and models that have been developed to aid in conducting risk assessments in the maritime domain and which are dedicated to the different steps of FSA.

With regard to the first step of FSA—identification of threats and risk variables—the commonly used methods are: literature review, brainstorming, methods for analysis of possible threats, and unwanted events (e.g., Hazard Identification Study, Hazard and Operability Study Failure Mode and Effect Analysis) (ABS, 2020; Ellis et al., 2008).

The second step—risk assessment—concerns mainly building a risk model. The methods here can be divided into qualitative and quantitative ones. The quantitative methods include: statistical analysis (based on historical records) (Blaich, Köhler, Reuter, & Hahn, 2015; Gerigk, 2012; Soares & Teixeira, 2001), Bayesian Networks (Berle et al., 2011; Gyftakis et al., 2018; Trucco et al., 2008), correlation analysis, Fuzzy Logic (Balmat, Lafont, Maifret, & Pessel, 2009; Elsayed, 2009; Johansson & Falkman, 2007), simulation-based methods (Blaich et al., 2015), or a combination of several methods (Eleye-Datubo, Wall, & Wang, 2008; Tu, Zhang, Rachmawati, Rajabally, & Huang, 2017).

With regard to qualitative risk assessment, the common methods are: Fault Tree Analysis (Hahn, 2014), Event Tree Analysis (Berle et al., 2011), risk matrixes, risk profiles, F-N curves, and relative ranking/risk indexes (ABS, 2020).

There are also risk assessment methods with a differentiation of critical factors which influence the overall risk level more heavily. They include either weights (Balmat et al., 2009; J. Liu, Yang, Wang, & Sii, 2005) or assume that only these risk variables are taken into account for which the probability of their occurrence is above a defined threshold (Trucco et al., 2008).

From the point of view of information systems, risk models are developed based on various artificial intelligence and machine learning methods. They focus mainly on modeling a "normal behavior of a ship by application of supervised and unsupervised techniques, such as classification, SVM, clustering, neural networks, or rule-based systems" (Chandola, Banerjee, & Kumar, 2009; Laxhammar & Falkman, 2010; Lee & Lee, 2006). Besides, test beds for assessment of new safety and risk applications are used (Hahn, 2014).

Table 3.1 presents a summary of popular methods for risk assessment, which are applied in the maritime domain.

The presented summary shows that there is a number of methods that can be applied to conduct maritime risk assessment. Therefore, the key issue is to choose the right method (or a combination of methods) which best matches the analyzed situation. The selected approach must also take into consideration that estimation of the probability of an adverse event and its effects. In relation to maritime transport this estimation may depend on various factors such as: itinerary, cargo size and volume, type of cargo and its properties (see Section 3.3 for a detailed overview of risk factors). One of the methods presented in this research (Chapter 7) assumes utilization of Bayesian Network (BN). The method for punctuality prediction, in turn (Chapter 8), uses concepts of a route prediction, ETA estimation, ship's density and various hazard in the maritime operational environment, including geopolitical risk. Therefore, these methods are presented in more detail in the next section.

Table 3.1. The selected risk analysis methods used in the maritime domain

Category	Method	Description and application	References
Qualitative	Expert-based analysis	Method in which risk factors, risk scenarios, their probability and impact are determined by subject matter experts; examples are brainstorming, interviews and Delphi method; used as a descriptive risk assessment; applicable to any type of risk	(Başhan, Demirel, & Gul, 2020; Ellis et al., 2008; Riveiro, 2011; van Laere & Nilsson, 2009; Wan, Yan, Zhang, & Yang, 2019b)
	Risk catalog	Collection of risks, defined using the common language; generic in nature; all items are potential risks that have been identified; each item is defined by risk type, scope and risk factor; used to generate possible risk scenarios	(Choi, Pelinovsky, Lee, & Woo, 2005; Miler, 2015).
	Fault Tree Analysis	Graphical, deductive technique for identification and analysis of risk; it starts from an undesired event and shows logical relationships between equipment failures, human errors and external events which might cause a specific event; applicable for almost every type of analysis, mostly used to address the fundamental causes of specific system failures; often used for complex electronic, control and communication systems	(ABS, 2020; Arici, Akyuz, & Arslan, 2020; Berle, Asbjørnslett, & Rice, 2011; Cem Kuzu, Akyuz, & Arslan, 2019)
	Event Tree Analysis	Graphical, deductive technique for identification and analysis of risk; it analyzes possible outcomes of an initiating event, capable of producing a mishap; applicable for almost every type of analysis, mostly used to address possible outcomes of an initiating event; often used for analysis of vessel movement mishaps and propagation of fire/explosion	(ABS, 2020; Arici, Akyuz, & Arslan, 2020; Berle, Asbjørnslett, & Rice, 2011; Ellis et al., 2008; Endrina, Rasero, & Konovessis, 2018; Hahn, 2014)
	Root cause analysis	Set of analysis tools, such as Event charting, 5 Whys technique, Root Cause Map, used to systematically discover how a mishap has occurred and the underlying root causes of the key contributors; applicable to any type of risk	(ABS, 2020)
	Risk matrix	Method used to rank the risk criticality of the failure modes; each risk scenario is evaluated taking into account its likelihood and consequences using a matrix; for each combination of likelihood-consequence, a level of risk is defined; used for evaluation of the impact of different scenarios with respect to different consequences	(ABS, 2020; Elsayed, 2009; Endrina, Rasero, & Konovessis, 2018; Wan, Yan, Zhang, & Yang, 2019b)

Category	Method	Description and application	References
Qualitative	Risk matrix	Method used to rank the risk criticality of the failure modes; each risk scenario is evaluated taking into account its likelihood and consequences using a matrix; for each combination of likelihood-consequence, a level of risk is defined; used for evaluation of the impact of different scenarios with respect to different consequences	(ABS, 2020; Elsayed, 2009; Endrina, Rasero, & Konovessis, 2018; Wan, Yan, Zhang, & Yang, 2019b)
	Preliminary risk assessment / risk profiles	Mashup-based approach used to characterize the risk associated with significant loss scenarios; relies on subject matter experts; used for generating risk profiles across broad range of activities (e.g., port-wide risk assessment)	(ABS, 2020)
	Influence diagrams	Graphical technique used to show interrelations between regulatory, operational and organizational factors; it models series of possible events and allows to define the critical events; applicable for various hazard categories like grounding, collision, fire, loss of propulsion and steering	(Berle, Asbjørnslett, & Rice, 2011; Cross & Ballesio, 2003)
Quantitative	Risk index	Method that uses various attributes (e.g., features of vessel/port) to calculate indexes that are further useful for making relative comparisons of various alternatives; generally applicable to any type of analysis situation (especially when only relative priorities are needed); extensively used to establish priorities for boarding and inspecting vessels	(ABS, 2020; Balmat, Lafont, Maifret, & Pessel, 2009; Ellis et al., 2008)
	Statistical analysis	Statistical methods like correlation, analysis of histograms, sensitivity, standard deviation; based mainly on historical data (e.g., accident statistics); used to estimate the probability of an event or risk scenario	(Blaich, Köhler, Reuter, & Hahn, 2015; Eiden & Martinsen, 2010; Endrina, Rasero, & Konovessis, 2018; Gerigk, 2012; Li, Meng, & Qu, 2012; Soares & Teixeira, 2001)

Category	Method	Description and application	References
Quantitative	Bayesian Networks	Graph-based technique, where nodes are risk variables with defined probabilities; these probabilities can depend on other node(s), through connections made by arcs; used for probability estimation of various risk scenarios, often in the causation analysis	(Berle, Asbjørnslett, & Rice, 2011; Gyftakis et al., 2018; Trucco, Cagno, Ruggeri, & Grande, 2008; Tu, Zhang, Rachmawati, Rajabally, & Huang, 2017; Wan, Yan, Zhang, Qu, & Yang, 2019a)
	Fuzzy Logic, IF-THEN rules	Fuzziness is a type of deterministic uncertainty that describes the event class ambiguity (i.e., outcomes that belong to several event classes at the same time but to different degrees); it measures the degree to which an event occurs, not whether it occurs; used when there exists an uncertainty for a risk factor (factor is vague, ambiguous, or fuzzy) and thus cannot be represented precisely by a probability distribution; allows for incorporation of human factors in risk analysis; often used with if-then rules which map probability, consequences and risk value	(Arici, Akyuz, & Arslan, 2020; Balmat, Lafont, Maifret, & Pessel, 2009; Eleye-Datubo, Wall, & Wang, 2008; Elsayed, 2009; Gaonkar et al., 2011; J. Liu, Yang, Wang, & Sii, 2005; Markowski, Mannan, & Bigoszewska, 2009; Tu, Zhang, Rachmawati, Rajabally, & Huang, 2017; Wan, Yan, Zhang, Qu, & Yang, 2019a)
	Geometrical estimation	Mathematical methods used to estimate a probability of an event; applicable for modeling risk of grounding or collision	(Li, Meng, & Qu, 2012)

Source: Own work.

3.2.3. Other methods used in the maritime domain

Apart from the methods that strictly concern maritime risk analysis, in the literature other groups of methods and techniques used in the maritime domain are also presented. These methods are not to be directly used as tools for risk assessment, but the results of these methods can be used as inputs to the risk assessment (i.e., as risk factors). From all the existing approaches we present here only these that are further used in the proposed methods, namely: Bayesian Networks, ship's route prediction, ships density calculation, and determination of geopolitical risk factors in the risk assessment. The methods dedicated for these issues are shortly characterized below.

Besides the methods described below, a special attention is paid to the process of detection of maritime threats and anomalies. The methods related to this topic is presented in detail in Chapter 6.

3.2.3.1. Bayesian Networks

The Bayesian Network (BN) is a directed acyclic graph consisting of nodes which represent a set of variables and edges which represent conditional dependencies between these variables. Each node takes as an input a set of values for the node's parent variables and gives as an output the probability (or probability distribution) of the variable represented by the node. Nodes that are not connected represent variables that are conditionally independent of each other. A BN could represent cause-effect relationships of a modeled phenomenon. It is a popular tool used to represent knowledge when there exists uncertainty.

Johansson and Falkman (2007) observed that BNs offer two interesting advantages over other approaches: 1) Bayesian models are easily understood by non-specialists; and 2) they allow for a straightforward incorporation of expert knowledge.

This approach has also some advantages in comparison to other methods which are important from the point of view of risk analysis. These are: the ability to model cause-effect and casual relations between variables and their probability distribution (inclusion of uncertainty); tolerance for missing data and imprecision on parameters; possibility to include prior information (expert knowledge); ability to model changes in time (by using the dynamic BNs). They can also be updated in real-time as soon as new information appears (Fooladvandi, Brax, Gustavsson, & Fredin, 2009). BNs are also easier to validate and evaluate (Mascaro, Nicholso, & Korb, 2014) by using, for example, a technique called sensitivity analysis. They are a good solution for supporting decision making process (Weber, Medina-Oliva, Simon, & Iung, 2012).

The learning of BN consists in modeling its structure and discovering which variables depend on each other as well as learning its content (which variables

should be included in the network). The structure of BNs can be created manually or automatically based on data analysis. For an automatic learning of BN various algorithms can be used, such as constraint-based methods (e.g., PC algorithm and its modifications or inductive causation algorithm) and search-and-score methods (used for small data samples). Conditional probabilities in BNs can be quantified using the Expectation-Maximization algorithm.

BNs are not free from defects and shortcomings: utilization of BNs requires providing information about *a priori* probability distribution of variables, which are not always known or can be calculated. Moreover, it is not obvious how the discretization of variables should be done, and significant information may be lost in the process of creating the structure of a BN (Laxhammar & Falkman, 2010).

The conducted literature review shows an increasing interest in the use of BN to estimate and improve reliability of complex systems over the last decade. The first utilization of BN for risk analysis is dated in 2001, when risk assessments using BN were conducted in military decision support systems, fire protection systems and in analyzing critical system failures due to human factors (Weber et al., 2012). Further on, BNs were successfully used in other domains: ecology (Pollino, Woodberry, Nicholson, Korb, & Hart, 2007), natural hazards (Grêt-Regamey & Straub, 2006; Straub, 2005), project and enterprise risk management (Lee & Lee, 2006) and health (K. F.-R. Liu, Lu, Chen, & Shen, 2012). Finally, BN models are used in dependability analyses to support such aspects as reliability, availability and maintainability (Weber et al., 2012).

In the maritime domain, BNs were used to detect anomalies in a ship's behavior. In 2007, a detection of point vessel anomalies (like speeding) with a BN approach was presented by Johansson and Falkman (2007). Fooladvandi et al. (2009) proposed signature-based activity detection using BNs, based on knowledge acquired from experts.

Nevertheless, in the review provided by Weber et al. (2012) it is concluded that the research, which solely focuses on risk analysis in technical systems, is no longer valid. Further research should: also take into account organizational and human factors contributions, include temporal dimensions of the system (system dynamics) as well as integrate qualitative information with quantitative knowledge on different abstraction levels.

As it was mentioned before, in this research the BN approach is used as one of the basic methods for the estimation of reliability and risk of a maritime transport service (presented in Chapter 7). The proposed method addresses the above requirements indicated by Weber et al. (2012), especially when it comes to the inclusion of changes in time (dynamic approach) and combining of various risk variables, both qualitative and quantitative. Initially, an application of BN was conducted for the static characteristic of ships. The result of this experiment was presented in (Stróżyna, 2017a). This concept was further developed for other characteristics of ships, which is presented in this work.

3.2.3.2. Travel time estimation

Travel time estimation concerns above all the calculation of an Estimated Time of Arrival (ETA) to a destination. There are three general methods for ETA determination:

(1) captains estimation: based on their experience;
(2) agent estimation: based on planning of the voyage and existing schedules;
(3) data-based estimation: based on data from various sources, both historical and forecasts.

In this research, we focus solely on the third method. Therefore, the below analysis concerns only the data-based estimation. In order to provide such an estimation, information about a ship's position is required. Nowadays almost all ships are equipped with an Automatic Identification System (AIS). Having information from an AIS about current and historical positions of a ship it is possible to determine whether a ship will be punctual and to estimate its travel time and the ETA.

The first group of methods that are used here are based on historical data. They assume that by collecting data regarding travel between two points, the time spent on covering the distance between these points can be predicted. To this end, both data stored in port systems regarding the ETA and the ETD (Estimated Time of Departure) can be used, as well as AIS data that allows for determination of times at various positions of an individual ship.

Veldhuis (2015) proposed a method for definition of an ETA in long distance shipping based on historical AIS data. He divided the ship's voyage into smaller segments, the defined points (ports) being the start and the end of each segment, and calculated the ETA based on historical travel times between these points. He also included information about the time spent at a given port.

An ETA can also be calculated based on a predicted average speed of a vessel over ground and distance to travel (Wielgosz, Wiśniewski, & Korwin-Piotrowski, 2012). Thus, the travel time and the ETA depend mainly on the speed. The speed, in turn, can depend on many factors, such as meteorological conditions (currents, waves, wind) or marine regulations (minimum or maximum speed for different areas).[2]

Apart from only position data, other data types are also taken into account. For example, there is research that considers the influence of certain weather data on the arrival time or a ship's speed (Calkoen & Santbergen, 2016; Szelangiewicz, Wiśniewski, & Żelazny, 2014; Wielgosz et al., 2012). Other research indicates that the credibility of the predicted ETA depends also on a ship's characteristics, like the ability of a ship to maintain speed on calm water both when the ship is fully

2. http://www.adrenaship.com/products/eta.html

loaded as well as in ballast condition. Wielgosz et al. (2012) claimed that, in order to accurately predict an ETA, it is required to use speed curves that depend on wind speed and direction as well as waves height and direction. However, the main issue here is the lack of accurate speed curves that would take into account all important factors like loading or ballast condition of a ship, fuel consumption, and weather conditions. Although these authors proposed a method for ETA calculation that takes into account predicted ship positions in the future and speed curves in different weather conditions, it can be used for predictions for the maximum of 9 days (due to access to weather forecasts data); for longer periods it uses average climatic data. As stated by the authors themselves, due to this fact, the results are probably influenced by seasonal average weather data.

In the Melodies project (Calkoen & Santbergen, 2016), a method for ETA estimation based on extrapolation was proposed. The method assumes that engine power is constant and thus the travel time depends on the ship's speed. The ship's speed can be estimated from (forecasted) weather conditions (a model of speed variations depending on wind, waves, and currents is provided). Then a relation between speed, time, and location is formulated and solved numerically. As a result, an ETA is provided, which can be updated every 3 hours.

Szelangiewicz et al. (2014), in turn, proposed a model based on relatively simple relations of speed, the basic parameters of the vessel and average statistical parameters of waves, wind, and surface currents. These relations are presented as nomograms or simple formulas prepared on the basis of measurements or calculations performed for many ships. The speed characteristic is then the basis for calculating a vessel's speed in the assumed statistical average weather conditions.

Summarizing, the main drawback of the presented ETA calculation methods is that they require a large variety of ship-related data, which might not always be present, like engine power, fuel consumption, etc. Moreover, they are only valid given a number of assumptions, like maintenance of constant engine power, constant specific fuel consumption, or constant ship speed in changing weather conditions. These assumptions may not always be met, especially on long routes.

3.2.3.3. Ships' density

As indicated by Wu, Xu, Wang, Wang, and Xu (2017), the problem of calculating vessel density has so far been addressed by few researchers. The existing approaches are based mainly on using positions of ships acquired from the AIS. The density analysis of the AIS is done either for traffic management and identification of low- and high-density regions (Chen, Xu, & Li, 2017; Eiden & Martinsen, 2010) or to build AIS receiving frequency maps to find areas with weak coverage (Wu et al., 2017).

There are two types of methods for analyzing ships' traffic: grid-based and vector-based. The first one (and the most widely used) divide the area into grid,

and then properties (like density) are calculated for each grid's cell (Marine Management Organisation, 2014; Shelmerdine, 2015). However, when it comes strictly to ships' density, so far, it was simply represented as a number of AIS messages per grid (Greidanus et al., 2013). Still, the existing methods do not consider consecutive AIS records for a given ship. As a result, the obtained maps are rather AIS message density maps, not traffic density maps. An example of such message-density maps that were created also in this research are presented in Chapter 9.

A more advanced grid-based method for calculating ships' density, that has been tested on a global scale, is presented in (Wu et al., 2017). They developed one month vessel density maps for 15 ship types in three spatial resolutions: 1 degree latitude by 1 degree longitude, 10 minutes latitude by 10 minutes longitude, and 1 minute latitude by 1 minute longitude. Moreover, the motion between two successive AIS messages was used to compute a ship's distance and time spent in each grid. Then, the traffic density was calculated as the average number of vessels that cross this region per unit area per unit time.

The topic of ships' density is also covered in research by LuxSpace (Eiden & Martinsen, 2010). They defined the density as "the average abundance of vessels within a defined geographical area". They assumed some ship population (62,000 ships) and lack of seasonality in the global vessel movement patterns. The latter assumption means that there are no daily/monthly changes in vessel traffic, which is, as even the authors indicated, only partially true. In order to calculate the density, they generated vessel position 'snapshots' or subsets, covering an 8 days time window with elimination of duplicated ships (each ship is considered only once in a given period) based on AIS data from 3 months. This assumption seems to be oversimplified since 8 days is a rather long period and at least some of the ships change a geographical area during this time. The analysis was conducted for a $1^{\circ} \times 1^{\circ}$ spatial grid, where for each subset the number of ships in each cell was calculated.

3.2.3.4. Geopolitical risk

The next issue that must be taken into account in maritime risk assessment is a geopolitical risk. This type of risk results from the route a ship follows, including the visited or passed countries. There are various geopolitical issues that may influence the security and safety of maritime traffic. The factors that may be taken into account include, inter alia, political conflicts, unrest, piracy, hijacking, armed robbery, terrorism, corruption, and civil disorders. There is some research that includes these aspects in risk assessment.

Lam (2012) proposed a rough-set approach to marine cargo risk analyses and have identified influential risk factors that affect shipping operations. In their research, they considered the geopolitical factor that involves the relationships among politics, geography, demography, and economy. In particular, two

sub-factors were included: 1) piracy, calculated on the basis of analysis of piracy hijacking incidents, and 2) political conflicts, such as wars and terrorist attacks, including analysis of location of major terrorist hubs. These hazards were also stressed as one of the potential threats for transported cargo.

Another research where geopolitical factors were taken into account is the model for assessment of operational reliability of the maritime transport system proposed by Gaonkar et al. (2011). In their approach, apart from factors like congestion and weather conditions on the route, ships characteristics (age, crew, technological advancement, maintenance, and past operational history), and the probability of unforeseen events on route, were taken into account. The authors analyzed whether a ship is sailing through areas which are prone to danger events, such as ship hijacking or capturing, looting, pirate attacks, or armed robbery. The area the ship sails through is an important factor influencing the safety of a ship's exploitation, especially the areas threatened by piracy. The problem of piracy was addressed also by other research, for example (Andler et al., 2009; Balmat et al., 2009; Bouejla et al., 2014) and European project PROMERC (Patrick, Davies, Baldacci, & den Breejen, 2015). The PROMERC project developed a solution for route planning that allows for reduction of piracy threat. They analyzed the historical piracy attacks and identified key parameters (ships' and environmental characteristics) that influence the probability of an attack. These parameters are used to calculate the risk of being successfully attacked.

Abramowicz-Gerigk et al. (2013) indicated that one of important hazards is the phenomenon of registering ships under the so-called flags of convenience (FOC). This risk factor was stressed also by others, e.g., el Pozo et al. (2010) (see Section 2.2 where the FOC problem is discussed).

Determination of the country risk is a subject of research of various international institutions. They publish their results on a regular basis and include a wide scope of information for each analyzed country. The examples of such reports are:

- **INFORM**[3]: a risk assessment for humanitarian crises and disasters. It is a transparent tool for understanding the risk and how it affects sustainable development. This measure is a global index, calculated for 191 countries, and takes into account open data published by international organizations.
  INFORM takes under consideration a wide range of indicators (approximately 50) to measure hazards and people's exposures to them. It creates a risk profile for each country and rates them between 0 (low risk) to 10 (high risk).
- **Basel AML Index**[4]: a risk index regarding money laundering and terrorism financing that takes also into account other related factors, such as financial and public transparency, and judicial strength. It is published by the Basel Institute, affiliated with the University of Basel.

3. http://www.inform-index.org/
4. https://index.baselgovernance.org

Basel index is calculated for 149 countries and the overall score is aggregated from 14 indicators divided into 5 weighted categories (Money laundering / Terrorist Financial Risk: 65%, Financial Transparency & Standards: 15%, Corruption Risk: 10%, Public Transparency & Accountability: 5%, Political & Legal Risk: 5%) and rated between 0 (low risk) to 10 (high risk).

- **World Risk Index**[5]: an index that measures the risk of disaster as a consequence of extreme natural events. It is created by the United Nations University's Institute for Environment and Human Security and is calculated for 171 countries.
  It consists of four components: exposure to natural hazards, susceptibility, coping capacities and adaptive capacities, which further include 28 indicators. The results are presented as percentage.

These three risk indicators (INFORM, BASEL, and World Risk Index) are further included in the proposed methods for punctuality prediction (Chapter 8) and reliability and risk assessment (Chapter 7).

3.3. Maritime risk variables

As presented in the previous sections, there already exist various methods for maritime risk assessment. Since one of the aims of this study is to identify what variables may influence the reliability and risk posed by individual ships, the existing methods and approaches were analyzed from the point of view of risk variables that they use. The variables were then consolidated and categorized. Finally, a typology of the ships' characteristics and attributes of their operational environments (risk variables) used in the maritime risk research is presented in this section.

The identified maritime risk variables were assigned to eight categories, proposed by the authors:

- Ship-related: it includes variables that relate to ships' characteristics; these are rather static attributes of ships that do not change very often, or not at all.
- Voyage-related: variables that concern a specific voyage of a ship (e.g., from port A to port B); they are voyage-specific (might change for each voyage) and relate to the ship itself (e.g., transit time, or transported cargo), or to the environment the ship operates in during the voyage (e.g., characteristics of areas a ship is sailing through); it is assumed that these attributes are stable within a given voyage.

5. http://collections.unu.edu/view/UNU:5763

- Dynamic: variables that concern a specific voyage, but may change underway; they are related to the movement of a ship over time.
- History-related: variables that concern the behavior of a ship in the past.
- Crew-related: variables that concern the crew of a ship and their competences as well as human errors and work conditions on a ship that might influence the number of errors.
- Environment-related: variables that characterize the environment a ship operates in during the voyage; they concern especially the meteorological conditions.
- Port-related: variables that relate to characteristics of ports a given ship is visiting; it includes both departure and destination ports.
- Other: variables not classified to any of the above categories, for example maritime regulations.

Table 3.2 presents the typology of risk variables used in other research and methods. The column "How often used" provides information in how many methods/studies a given variable was used, followed by a listing of the research in the column "References".

Based on the created typology it was possible to identify variables that, to the best of our knowledge, so far have not been used in the assessment of reliability and risk in the maritime domain, as well as variables that were used only in few of the analyzed research studies. Both types of variables have been considered as ones that might or should be included in the actual research, also in the methods presented in this study.

Moreover, the analysis of the typology allowed for an identification of a set of variables, which so far have not been used together in a single research, and which, in our opinion, might be linked together to discover some unknown correlations and their influence on the reliability and risk of a maritime transport service.

3.4. Shortcomings and gaps in the existing risk assessment methods

The conducted literature review as well as the results of the survey among the maritime experts have allowed us to identify gaps and shortcoming of the existing solutions and, thus, reveal important barriers limiting the increase of the quality of maritime transport services, especially their reliability. The main barrier, which has also been confirmed by other researchers (Wieteska, 2011, p. 176), is still a lack of comprehensive and efficient IT tools with implemented functionalities for maritime reliability and risk assessment. Despite the added value of using

Table 3.2. Typology of maritime risk variables

Category	Variable	How often used	References
Ship-related	Capacity / gross tonnage / size	11	(Berle, Asbjørnslett, & Rice, 2011; Cross & Ballesio, 2003; Eiden & Martinsen, 2010; Gilberg, Kleiven, & Bye, 2016; Mazaheri, 2017; Mazaheri, Montewka, & Kujala, 2013; Pedersen, 1995; Soares & Teixeira, 2001), (Samuelides, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013), (Kite-Powell et al., 1999, as cited in Mazaheri, Montewka, and Kujala, 2013), (Jebsen & Papakonstantinou, 1997, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Dimensions (e.g., length, draft)	7	(Gilberg, Kleiven, & Bye, 2016; Mazaheri, 2017; Pedersen, 1995), (Macduff, 1974, as cited in Mazaheri, Montewka, and Kujala, 2013), (Ramboll, 2006, as cited in Mazaheri, Montewka, and Kujala, 2013), (Uluscu, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013), (van Dorp, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Owner	2	(Berle, Asbjørnslett, & Rice, 2011; Gilberg, Kleiven, & Bye, 2016)
	Age	8	(Abramowicz-Gerigk, Burciu, & Kamiński, 2013; Balmat, Lafont, Maifret, & Pessel, 2009; Eiden & Martinsen, 2010; Gaonkar et al., 2011; Gilberg, Kleiven, & Bye, 2016; Lam, 2012), (Samuelides, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013), (Uluscu, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Type	9	(Balmat, Lafont, Maifret, & Pessel, 2009; Gaonkar et al., 2011; Gilberg, Kleiven, & Bye, 2016; Pedersen, 1995; Soares & Teixeira, 2001), (Uluscu, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013), (van Dorp, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013), (Kite-Powell et al., 1999, as cited in Mazaheri, Montewka, and Kujala, 2013), (Jebsen & Papakonstantinou, 1997, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Flag	4	(Abramowicz-Gerigk, Burciu, & Kamiński, 2013; Balmat, Lafont, Maifret, & Pessel, 2009; Gilberg, Kleiven, & Bye, 2016), (Uluscu, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Reputation	1	(Elsayed, 2009)
	Maintenance program	2	(Gaonkar et al., 2011; Trucco, Cagno, Ruggeri, & Grande, 2008)
	Emergency system on the ship	2	(Gaonkar et al., 2011; Gerigk, 2012)

Category	Variable	How often used	References
Ship-related	Technological features (hull, machinery, state, stability, reserve buoyancy, equipment, innovations, upgrade)	13	(Abramowicz-Gerigk, Burciu, & Kamiński, 2013; Balmat, Lafont, Maifret, & Pessel, 2009; Başhan, Demirel, & Gul, 2020; Cross & Ballesio, 2003; Eiden & Martinsen, 2010; Gaonkar et al., 2011; Gerigk, 2012; Gilberg, Kleiven, & Bye, 2016; Lam, 2012; Mazaheri, 2017; Soares & Teixeira, 2001; Trucco, Cagno, Ruggeri, & Grande, 2008), (DNV, 2003, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Management rules	4	(Gerigk, 2012; Soares & Teixeira, 2001; Trucco, Cagno, Ruggeri, & Grande, 2008), (DNV, 2003, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Classification society, status, advices	2	(Gilberg, Kleiven, & Bye, 2016; Trucco, Cagno, Ruggeri, & Grande, 2008)
Dynamic	Speed / heading / course	13	(Balmat, Lafont, Maifret, & Pessel, 2009; Berle, Asbjørnslett, & Rice, 2011; Endrina, Rasero, & Konovessis, 2018; Gilberg, Kleiven, & Bye, 2016; Mazaheri, 2017; Mazaheri, Montewka, & Kujala, 2013; Soares & Teixeira, 2001; Trucco, Cagno, Ruggeri, & Grande, 2008), (Macduff, 1974, as cited in Mazaheri, Montewka, and Kujala, 2013), (Ramboll, 2006, as cited in Mazaheri, Montewka, and Kujala, 2013), (Fujii et al., 1974, as cited in Mazaheri, Montewka, and Kujala, 2013), (Quy et al., 2006, as cited in Mazaheri, Montewka, and Kujala, 2013), (COWI, 2008, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Location / position; distance evolution; ship's trajectory	3	(Balmat, Lafont, Maifret, & Pessel, 2009; Eiden & Martinsen, 2010; Trucco, Cagno, Ruggeri, & Grande, 2008)
	Time to shore	1	(Eiden & Martinsen, 2010)
	Anomalous route/trajectory	1	(Balmat, Lafont, Maifret, & Pessel, 2009)
	Time	2	(Eiden et al., 2007, as cited in Mazaheri, Montewka, and Kujala, 2013), (van Dorp, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013)
Voyage--related	Transit time	1	(Berle, Asbjørnslett, & Rice, 2011)

Category	Variable	How often used	References
Voyage--related	Geopolitical issues (e.g., political conflicts/unrest, piracy, hijacking, armed robbery, terrorism, crimes, corruption, civil disorders)	6	(Abramowicz-Gerigk, Burciu, & Kamiński, 2013; Başhan, Demirel, & Gul, 2020; Berle, Asbjørnslett, & Rice, 2011; Gaonkar et al., 2011; Jarysz-Kamińska, 2013; Lam, 2012)
	Cargo type / vulnerability (e.g., dangerous)	6	(Eiden & Martinsen, 2010; Gilberg, Kleiven, & Bye, 2016; Jarysz-Kamińska, 2013; Lam, 2012; Trucco, Cagno, Ruggeri, & Grande, 2008; Wan, Yan, Zhang, Qu, & Yang, 2019a)
	Congestion (at source harbor, destination harbor, sea) / traffic density	9	(Başhan, Demirel, & Gul, 2020; Endrina, Rasero, & Konovessis, 2018; Gaonkar et al., 2011; Pedersen, 1995; Soares & Teixeira, 2001; Trucco, Cagno, Ruggeri, & Grande, 2008), (Fujii et al., 1974, as cited in Mazaheri, Montewka, and Kujala, 2013) (COWI, 2008, as cited in Mazaheri, Montewka, and Kujala, 2013), (Ramboll, 2006, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Area (incl. characteristics of waterway, tug response time, self repair time)	15	(Abramowicz-Gerigk, Burciu, & Kamiński, 2013; Başhan, Demirel, & Gul, 2020; Cross & Ballesio, 2003; Eiden & Martinsen, 2010; Mazaheri, 2017; Pedersen, 1995; Soares & Teixeira, 2001), (Fujii et al., 1974, as cited in Mazaheri, Montewka, and Kujala, 2013), (Ramboll, 2006, as cited in Mazaheri, Montewka, and Kujala, 2013), (Quy et al., 2006, as cited in Mazaheri, Montewka, and Kujala, 2013), (COWI, 2008, as cited in Mazaheri, Montewka, and Kujala, 2013), (Kite-Powell et al., 1999, as cited in Mazaheri, Montewka, and Kujala, 2013), (Jebsen and Papakonstantinou, 1997, as cited in Mazaheri, Montewka, and Kujala, 2013), (Briggs et al., 2003, as cited in Mazaheri, Montewka, and Kujala, 2013), (Lin and all 1998, as cited in Mazaheri, Montewka, and Kujala, 2013)
History--related	Delays / downtime / timely delivery	4	(Berle, Asbjørnslett, & Rice, 2011; Cross & Ballesio, 2003; Elsayed, 2009; Gaonkar et al., 2011)
	Past incidents / accidents / ship damage	13	(Başhan, Demirel, & Gul, 2020; Berle, Asbjørnslett, & Rice, 2011; Cross & Ballesio, 2003; Eiden & Martinsen, 2010; Elsayed, 2009; Endrina, Rasero, & Konovessis, 2018; Gaonkar et al., 2011; Gilberg, Kleiven, & Bye, 2016; Jarysz-Kamińska, 2013; Lam, 2012; Mazaheri, 2017; Trucco, Cagno, Ruggeri, & Grande, 2008), (Kristiansen, 2005, as cited in Mazaheri, Montewka, and Kujala, 2013)

Category	Variable	How often used	References
History--related	Number of casualties/fatalities	2	(Ellis et al., 2008; Elsayed, 2009)
	Pollution events	5	(Cross & Ballesio, 2003; Eiden & Martinsen, 2010; Ellis et al., 2008; Elsayed, 2009; Jarysz-Kamińska, 2013)
	Cargo loss / cargo damage / excessive loads	5	(Abramowicz-Gerigk, Burciu, & Kamiński, 2013; Ellis et al., 2008; Gaonkar et al., 2011; Jarysz-Kamińska, 2013; Soares & Teixeira, 2001)
	Number of owners	1	(Balmat, Lafont, Maifret, & Pessel, 2009)
	Duration of detentions	1	(Balmat, Lafont, Maifret, & Pessel, 2009)
	Failure frequency (e.g., technical failure)	13	(Arici, Akyuz, & Arslan, 2020; Başhan, Demirel, & Gul, 2020; Cross & Ballesio, 2003; J. Liu, Yang, Wang, & Sii, 2005; Mazaheri, 2017; Trucco, Cagno, Ruggeri, & Grande, 2008), (Amrozowicz et al., 1997, as cited in Mazaheri, 2017), (Fowler and Sorgard, 2000, as cited in Mazaheri, Montewka, and Kujala, 2013), (Ramboll, 2006, as cited in Mazaheri, Montewka, and Kujala, 2013), (Uluscu, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013), (van Dorp, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013), (DNV, 2003, as cited in Mazaheri, Montewka, and Kujala, 2013), (Eiden et al., 2007, as cited in Mazaheri, Montewka, and Kujala, 2013)
Crew, human errors	Crew characteristics (including knowledge, skills, experience, training, size, rotation scheme, language, culture, morale, motivation, performance, fitness for duty)	256	(Abramowicz-Gerigk, Burciu, & Kamiński, 2013; Arici, Akyuz, & Arslan, 2020; Cem Kuzu, Akyuz, & Arslan, 2019; Eleye-Datubo, Wall, & Wang, 2008; Ellis et al., 2008; Gaonkar et al., 2011; Gerigk, 2012; Gilberg, Kleiven, & Bye, 2016; Lam, 2012; Mazaheri, 2017; Pedersen, 1995; Soares & Teixeira, 2001; Trucco, Cagno, Ruggeri, & Grande, 2008; Wan, Yan, Zhang, & Yang, 2019b), (Amrozowicz et al., 1997, as cited in Mazaheri, 2017), (DNV, 2003, as cited in Mazaheri, Montewka, and Kujala, 2013), (Ramboll, 2006, as cited in Mazaheri, Montewka, and Kujala, 2013), (Uluscu, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013), (COWI, 2008, as cited in Mazaheri, Montewka, and Kujala, 2013), (Lin et al., 1998, as cited in Mazaheri, Montewka, and Kujala, 2013), (van Dorp, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013), (Martins and Maturana, 2010, as cited in Mazaheri, Montewka, and Kujala, 2013), (Brown and Haugene, 1998, as cited in Mazaheri, Montewka, and Kujala, 2013), (Kite-Powell et al., 1999, as cited in Mazaheri, Montewka, and Kujala, 2013), (Jebsen & Papakonstantinou, 1997, as cited in Mazaheri, Montewka, and Kujala, 2013), (Akhtar & Utne, 2014, as cited in Mazaheri, 2017)

Category	Variable	How often used	References
Crew, human errors	Work conditions (complexity, workload, ergonomics, work process, stress, duty scheme)	7	(Cross & Ballesio, 2003; Eleye-Datubo, Wall, & Wang, 2008; Endrina, Rasero, & Konovessis, 2018; Gilberg, Kleiven, & Bye, 2016; Lam, 2012; Wan, Yan, Zhang, & Yang, 2019b), (DNV, 2003, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Human errors	10	(Abramowicz-Gerigk, Burciu, & Kamiński, 2013; Arici, Akyuz, & Arslan, 2020; Cem Kuzu, Akyuz, & Arslan, 2019; Cross & Ballesio, 2003; Endrina, Rasero, & Konovessis, 2018; Gerigk, 2012; Läsche, Pinkowski, Gerwinn, Droste, & Hahn, 2014; Soares & Teixeira, 2001; Trucco, Cagno, Ruggeri, & Grande, 2008), (Kristiansen, 2005, as cited in Mazaheri, Montewka, and Kujala, 2013)
Environment, weather	Meteorological conditions (weather) and climate	19	(Abramowicz-Gerigk, Burciu, & Kamiński, 2013; Arici, Akyuz, & Arslan, 2020; Balmat, Lafont, Maifret, & Pessel, 2009; Cross & Ballesio, 2003; Eiden & Martinsen, 2010; Gaonkar et al., 2011; Gerigk, 2012; Gilberg, Kleiven, & Bye, 2016; Soares & Teixeira, 2001; Trucco, Cagno, Ruggeri, & Grande, 2008), (DNV, 2003, as cited in Mazaheri, Montewka, and Kujala, 2013), (Ramboll, 2006, as cited in Mazaheri, Montewka, and Kujala, 2013), (Uluscu, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013), (van Dorp, 2009, as cited in Mazaheri, Montewka, and Kujala, 2013), (Kite-Powell et al., 1999, as cited in Mazaheri, Montewka, and Kujala, 2013), (Jebsen & Papakonstantinou, 1997, as cited in Mazaheri, Montewka, and Kujala, 2013), (Fowler & Sorgard, 2000, as cited in Mazaheri, Montewka, and Kujala, 2013), (Eiden et al., 2007, as cited in Mazaheri, Montewka, and Kujala, 2013), (Briggs et al., 2003, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Weather extremes (wind, fog, rain, snow, clouds) and natural hazards (earthquake, tsunami, volcano eruption)	7	(Başhan, Demirel, & Gul, 2020; Cross & Ballesio, 2003; Endrina, Rasero, & Konovessis, 2018; Jarysz-Kamińska, 2013; Lam, 2012; Soares & Teixeira, 2001; Wan, Yan, Zhang, & Yang, 2019b)
	Sea state (sea current, waves)	8	(Abramowicz-Gerigk, Burciu, & Kamiński, 2013; Arici, Akyuz, & Arslan, 2020; Balmat, Lafont, Maifret, & Pessel, 2009; Gerigk, 2012; Gilberg, Kleiven, & Bye, 2016; Trucco, Cagno, Ruggeri, & Grande, 2008), (Briggs et al., 2003, as cited in Mazaheri, Montewka, and Kujala, 2013), (Lin et al., 1998, as cited in Mazaheri, Montewka, and Kujala, 2013)

Category	Variable	How often used	References
	Visibility	5	(Balmat, Lafont, Maifret, & Pessel, 2009; Gilberg, Kleiven, & Bye, 2016; Trucco, Cagno, Ruggeri, & Grande, 2008), (Kite-Powell et al., 1999, as cited in Mazaheri, Montewka, and Kujala, 2013), (Jebsen & Papakonstantinou, 1997, as cited in Mazaheri, Montewka, and Kujala, 2013)
	Time of day (Day/night, darkness)	6	(Balmat, Lafont, Maifret, & Pessel, 2009; Başhan, Demirel, & Gul, 2020; Mazaheri, 2017; Mazaheri, Montewka, & Kujala, 2013), (Lin et al., 1998, as cited in Mazaheri, Montewka, and Kujala, 2013), (Fujii et al., 1974, as cited in Mazaheri, Montewka, and Kujala, 2013)
Port-related	Port operations (e.g., berthing/unberthing procedure)	4	(Arici, Akyuz, & Arslan, 2020; Başhan, Demirel, & Gul, 2020; Lam, 2012; Trucco, Cagno, Ruggeri, & Grande, 2008)
	Port capacity (water depth, shoreline length, number of berths, cranes)	3	(Lam, 2012; Trucco, Cagno, Ruggeri, & Grande, 2008; Wan, Yan, Zhang, & Yang, 2019b)
	Cargo loading/unloading procedure	4	(Abramowicz-Gerigk, Burciu, & Kamiński, 2013; Lam, 2012; Trucco, Cagno, Ruggeri, & Grande, 2008; Wan, Yan, Zhang, & Yang, 2019b)
	Traffic management	3	(Trucco, Cagno, Ruggeri, & Grande, 2008; Wan, Yan, Zhang, & Yang, 2019b), (Praetorius, 2012, as cited in Mazaheri, 2017)
	Personnel (competence, training)	3	(Arici, Akyuz, & Arslan, 2020; Cem Kuzu, Akyuz, & Arslan, 2019; Trucco, Cagno, Ruggeri, & Grande, 2008)
	Geopolitical (political unrest, corruption, civil disorders, terrorism, crimes)	3	(Lloyd, 2013; Wan, Yan, Zhang, Qu, & Yang, 2019a; Wan, Yan, Zhang, & Yang, 2019b)
Other	Maritime regulations (conventions, recommendations, rules)	4	(Gerigk, 2012; Kangsvik, Størkersen, & Antonsen, 2017; Trucco, Cagno, Ruggeri, & Grande, 2008; Wan, Yan, Zhang, & Yang, 2019b)

Source: Own work.

multiple sources of information and implementation of some analytic methods, there still exist deficits and gaps in the available maritime systems when it comes to conducting risk analysis. There is room for improvement in this area. Moreover, it is exacerbated by challenges identified in other maritime-related areas, which also may influence the process of reliability and risk assessment.

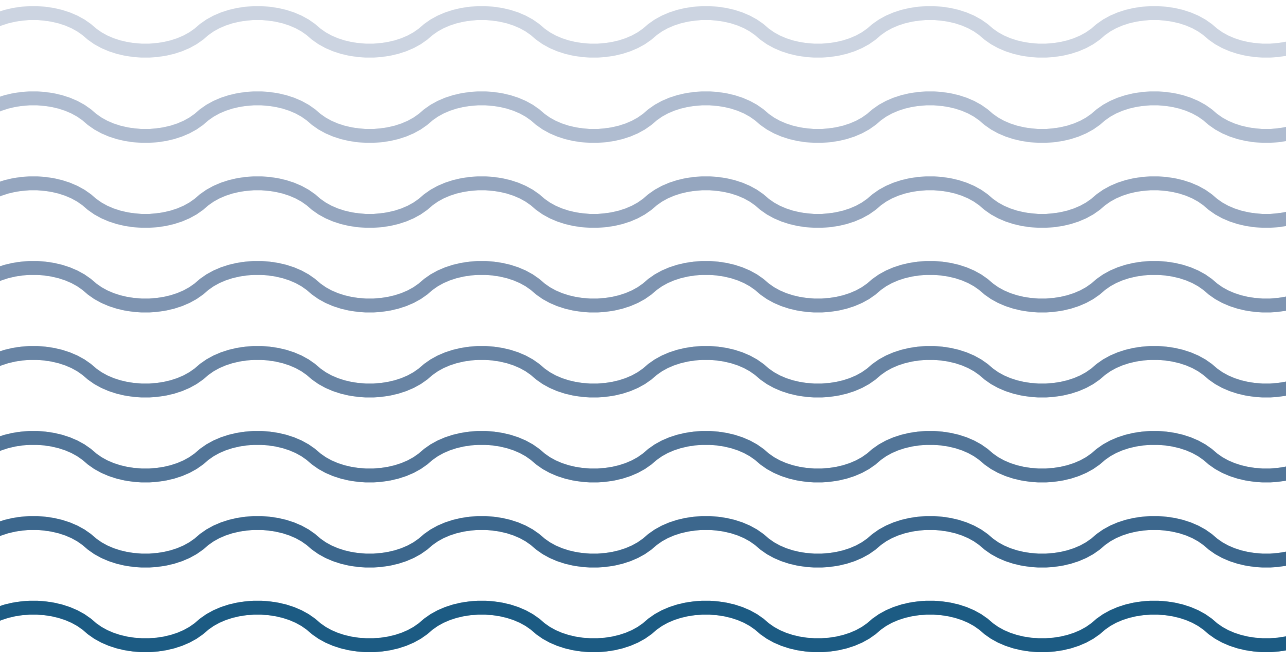
The conducted literature review, presented in the previous sections, has revealed some shortcomings and disadvantages of the existing methods for maritime risk assessment. In general, the existing methods focus either on modeling of a licit behavior of ships or identification of an abnormal behavior. Both approaches have some limitations.

First, the existing methods focus mainly on estimating a risk of one of the three (separate) types of situations: 1) either a specified type of hazard (e.g., collision, oil spill); or 2) an undesired event in relation to a ship's technical attributes (e.g., an engine problem); or 3) a human error. Therefore, they take into account only selected risk factors, strictly connected with a given type of hazard. Besides, the estimated risk concerns a particular group of ships (e.g., tankers with the same or similar characteristics), instead of an individual ship. Thus, their results may be used only in a given context, i.e., in a selected risk scenario.

Second, the analyzed methods use a limited set of factors or only factors of a given category, like technical characteristics of a ship, experience of the crew, history of accidents. Many of them do not include such important aspects as: factors that may change in time (e.g., flag, owner, classification status, congestion on the route), current and historical anomalies in a ship's behavior, past routes, characteristics of the current route, and further risks that are related to the ship's localization (e.g., geopolitical risk, congestion, weather). The maritime experts indicated that more attention should be paid to analysis of several factors (categories of factors) combined, instead of considering them separately.

Final, the existing methods provide above all information about the long- or mid-term risk level (especially when it comes to risk of supply performed by merchant ships). Very few methods (e.g., Balmat et al., 2009; Blaich et al., 2015; Hornauer & Hahn, 2013), focus on the short-term risk, which concerns a given voyage, a given ship, and which additionally may change during this voyage.

Chapter 4



4. MARITIME DATA

Businesses operating in the era of the digital economy need to be supplied with appropriate data in order to make informed decisions and remain competitive. This also concerns entities working in the maritime domain. The necessary data can be acquired from various sources, depending on the system, its purposes, and its operating context. One source of data can be internal (e.g., transactional data) or external (e.g., sensors, external systems and databases, or the Internet). Irrespective of the type of data, each data source to be used for decision-making needs to be appropriately defined and assessed in order to assure delivery of high-quality results (Robey & Markus, 1984). It also has to be relevant to the entity and must fulfill certain criteria concerning its purposes and the domain in which it operates. Only after high-quality data sources are selected, can data be retrieved, integrated, and finally analyzed for use in decision-making. In the maritime domain there are various sources that may provide the data required for analyses. In this chapter, selected examples of maritime-related data sources are presented, followed by a methodology for selecting sources and methods for extracting data. Finally, a case study of a system that extracts, fuses, and analyzes data from two sources is presented.

4.1. Data sources used in the maritime domain

The data sources that are applicable to the maritime domain can be divided into three categories.

The first category is sensors. Sensors provide kinematic data for objects observed in their coverage area and can be further split into active (e.g., radar or sonar) and passive, which rely on data broadcast intentionally by an entity (e.g., AIS or LRIT). Sensors generate data streams,that is, sequences of digital signals in the form of data packets, which are used to transmit information. Depending on the source, data streams may contain different sets of data, such as timestamps, object IDs (e.g., the vessel identification number), and various attributes (e.g., geodata or ship characteristics).

The second category of sources is authorized databases. They include information about vessels, cargo, crew, etc. Examples are port notifications sent by ships,

HAZMAT reports, and the Long Range Identification System (LRIT),[1] SafeSeaNet,[2] or West European Tanker Reporting System.[3]

The third category includes data that are publicly available on the Internet (hereinafter referred to as Internet data sources). These sources include vessel traffic data, reports, and news, among others. The data are provided online by various organizations and communities. For example, some ports publish vessel traffic data or facilities information. There are also dedicated services that provide ship-related information on a regular basis, such as MarineTraffic,[4] FleetMon,[5] VesselFinder.[6]

In general, data sources from the first and second category—sensors and authorized databases—are not publicly available. They are accessible primarily for the owners of the sensors, the maritime authorities, and other authorized entities. The exception is real-time Automatic Identification System (AIS) data, which can be collected by anyone who has an appropriate receiver. However, historical AIS data are provided commercially. A similar situation can be observed with commercial (paid) Internet data sources. These sources are referred to as closed data sources.

Taking this into account, further in this study we focus mainly on data from open Internet sources and the AIS. They are listed in Section 4.5. The popular maritime data sources are described in the following sections.

4.1.1. Sensor data

Automatic Identification System

The AIS isa system used for identifying and locating vessels in real time (a tracking system). It was created as a tool for avoiding collisions at sea and is based on the automatic exchange of data about a ship and its position between it and other nearby ships, AIS base stations, and satellites. To exchange the data, a VHF maritime mobile band is used, with a range of a 20 to 30 nautical miles (nmi). Since 2008, satellites have also been able to receive AIS signals. Thanks to this, the system has global coverage and creates the possibility to track ships on a worldwide scale. Satellites operating in Low Earth Orbits, at an altitude of around 500 km, provide a field of view of more than 3,000 km (1,620 nmi) in diameter. Nowadays, the AIS is used as one of the main data sources in maritime surveillance, since it is required to be used by all vessels above 300 GRT (Gross Register Tonnage) and

1. https://www.imo.org/en/OurWork/Safety/Pages/LRIT.aspx
2. http://www.emsa.europa.eu/ssn-main.html
3. Introduced by IMO Resolution MSC.190(79), Amended by MSC.301(87).
4. https://www.marinetraffic.com/
5. http://www.fleetmon.net/
6. https://www.vesselfinder.com/

because it has a high update frequency. Depending on the ship's speed over ground, the dynamic data should be sent every 2 to 10 seconds and the static data every 6 minutes.

The AIS data from individual ships can be received not only by other ships, but also by land-based stations. All ships with VHF communication in range of coastal receiver stations can be seen by them, thus creating a solution for monitoring and controlling vessel traffic close to the coastal station. Coastal countries have established shore-based AIS receiving station networks for surveying the vessel traffic within particular areas. However, the range of these stations is limited to more or less 30 or 40 nautical miles along coast lines, which narrows the surveillance area significantly. Therefore, satellite AIS data are fused with AIS data captured by the global network of terrestrial stations in order to enhance this coverage.

Compared to other data sources, the AIS provides a significant amount of data about the movement of vessels. The information exchanged includes navigational data (location, course, speed, and navigational status), static data (identification numbers, type, name, and call sign), and voyage data (destination port and estimated time of arrival). Tables 4.1 and 4.2 present the dynamic and static AIS messages. Figures 4.1 and 4.2 present the coverage of satellite and terrestrial AIS.

Table 4.1. Dynamic AIS message information

AIS dynamic information	
Maritime Mobile Service Identity (MMSI)	a unique nine-digit identification number of a vessel
Navigation status	e.g., "at anchor", "under way using engine(s)", "not under command", etc.
Rate of turn	right or left, in degrees per minute
Speed over ground (SOG)	vessel's speed in 1/10 knots
Position	latitude/longitude of a vessel
Course over ground (COG)	vessel's course relative to true North
True heading	vessel's course in degrees (for example from a gyro compass)

Source: Own work based on (International Telecommunication Union, 2010).

In general, the AIS offers some unique benefits:

- Coverage area of the space-borne AIS reaches far beyond the range of coastal AIS stations, enabling global coverage.
- Terrestrial-based AIS systems, which are already used in many countries, include data distribution mechanisms. Thus, space-based AIS data can be a complementary data source for existing AIS systems.

Table 4.2. Static AIS message information

AIS static information	
IMO	vessel identification number—a seven digit number that remains unchanged upon transfer of the vessel's registration to another country
Callsign	international radio call sign, up to seven characters, assigned to the vessel by its country of registry
Name	max 20 characters to represent the name of a vessel
Type	type of ship/cargo
Dimensions	dimensions of a vessel, to nearest meter
Location of antenna	location of positioning system's (e.g., GPS) antenna on board the vessel
Positioning system	type of positioning system, such as GPS, DGPS
Draught	draught of ship: 0.1 meter to 25.5 meters
Destination	where a vessel is heading, max 20 characters
ETA	Estimated Time of Arrival at destination—UTC month/date hour:minute

Source: Own work based on (International Telecommunication Union, 2010).

- In regions without coastal AIS stations, a space-based AIS could be a cost--effective alternative for monitoring vessel traffic.
- It has more information in comparison to other ship reporting systems, like the LRIT. Thus, it may be used to complement information from other systems, such as radar.
- The system is popular—every ship over 300 GRT is equipped with AIS, but there are also many smaller ships that use AIS voluntarily. It is estimated that AIS is currently used by more than 200,000 ships.
- AIS messages are automatically transmitted by ships every few seconds. Thus, a near real-time maritime picture can be created.
- Thanks to the high update frequency, AIS can provide full tracking of ships.

Although it offers many benefits, the AIS is still an imperfect solution and requires further improvement of data quality. There are several reasons for that.

Firstly, the AIS is the foundation for generating the maritime picture, but often some information in AIS messages is missing. This is due to the fact that the AIS is a "cooperative" system, meaning both that information sent in the messages must be provided by a ship master and that whether an AIS transponder is switched on or off is also the ship's master decision (although continous AIS transponder operator is recommended by the IMO). As a consequence, vessel tracks are often incomplete or AIS messages are missing important parts. Thus, the maritime situation appears

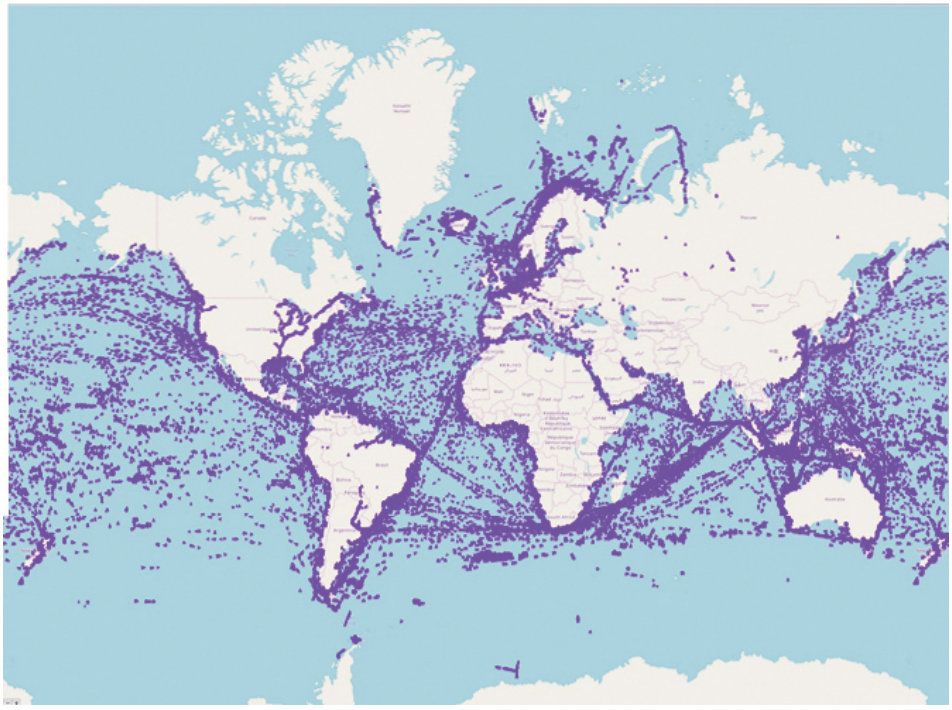


Figure 4.1. Satellite AIS coverage

Source: Materials developed within the SIMMO project.

incomplete, which makes it difficult for a maritime entity to monitor and control the current situation at sea.

Secondly, the AIS is vulnerable to different threats, such as spoofing, hijacking, and availability disruption (Coleman, Kandah, & Huber, 2020).

The third important issue is the limited scope of information provided by the AIS about a ship. AIS messages include only some essential, basic information about ships (see Table 4.1 and Table 4.2). However, in order to perform a comprehensive risk assessment or detect an anomaly, more relevant pieces of information are necessary. Therefore, additional data sources need to be used to complement the information provided by the AIS (e.g., from other data sources groups described at the beginning of the chapter). These especially encompass the following:

- general ship data: flag, detailed type, length, gross tonnage, capacity, technical specifications, and construction details;
- ownership data (current and past owners), classification status, classification history;
- former ship's name, flag history;
- security related information: bans, detentions, port state controls;

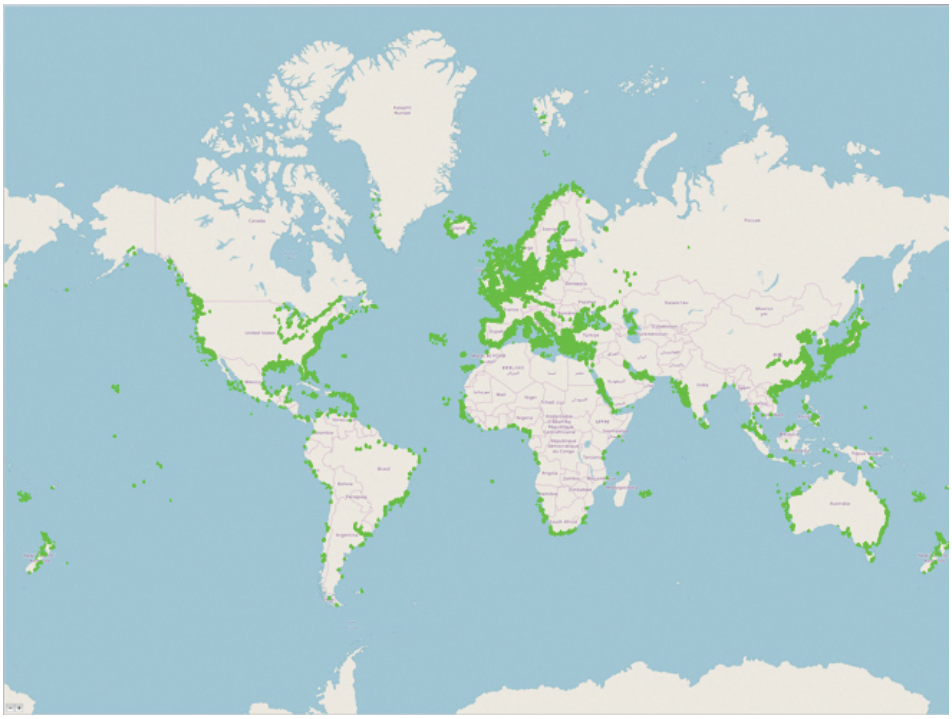


Figure 4.2. Terrestrial AIS coverage

Source: Materials developed within the SIMMO project.

- validated voyage information: current destination, expected time of arrival (ETA);
- voyage history: last port calls, average speed over ground, drought.

All the above elements are relevant in describing a ship and its movement, conducting risk assessments, or detecting anomalies. However, additional information is rarely exploited currently, mainly due to the lack of integration between different data sources. This is a weak point of the existing systems. On the Internet there are a number of open data sources that provide such information.

Moreover, due to the fact that the AIS messages are transmitted on a constant basis and taking into account the fact that the AIS is now being used by more than 200,000 ships, every day the system generates a huge amount of data. As a consequence, it is a challenge to efficiently process such big data sets. This is another shortcoming of the existing methods, because many of them suffer from computational inefficiency (Marine Management Organisation, 2014; Shelmerdine, 2015). Therefore, they offer data analysis at regional and national scales and for short time periods (Wu et al., 2017). However, generating a Recognized Maritime Picture requires requires a huge amount of data records be analyzed and

interpreted. Therefore, information systems with advanced processing, analysis, and reasoning capabilities are required, which would provide a fast assessment of the situation and support users in decision-making (Pallotta, Vespe, & Bryan, 2013). It concerns real-time identification of potential maritime threats in particular.

Other reasons why the AIS suffers from some data quality problems and needs further improvements are as follows:

- Along coastal regions, ships are tracked using the terrestrial station network, offering update frequency of ship positions within 15 minutes.
- Satellite AIS reception in coastal regions, especially in areas of high vessel density such as the North Sea or the Baltic Sea, is relatively poor due to the limited storage capacity of satellites. Still, there are maritime regions where AIS coverage is limited (see an example of the Baltic Sea with poor AIS coverage, i.e., in Bothnian Bay, the East Gotland Basin, and the Bornholm Basin) (Figure 4.3).
- Despite the growing satellite constellation (totaling to 60 at the end of 2020, provided by different companies like ORBCOMM, exactEarth, or Spire Global), AIS reception on the open seas (outside of terrestrial coverage) may still be

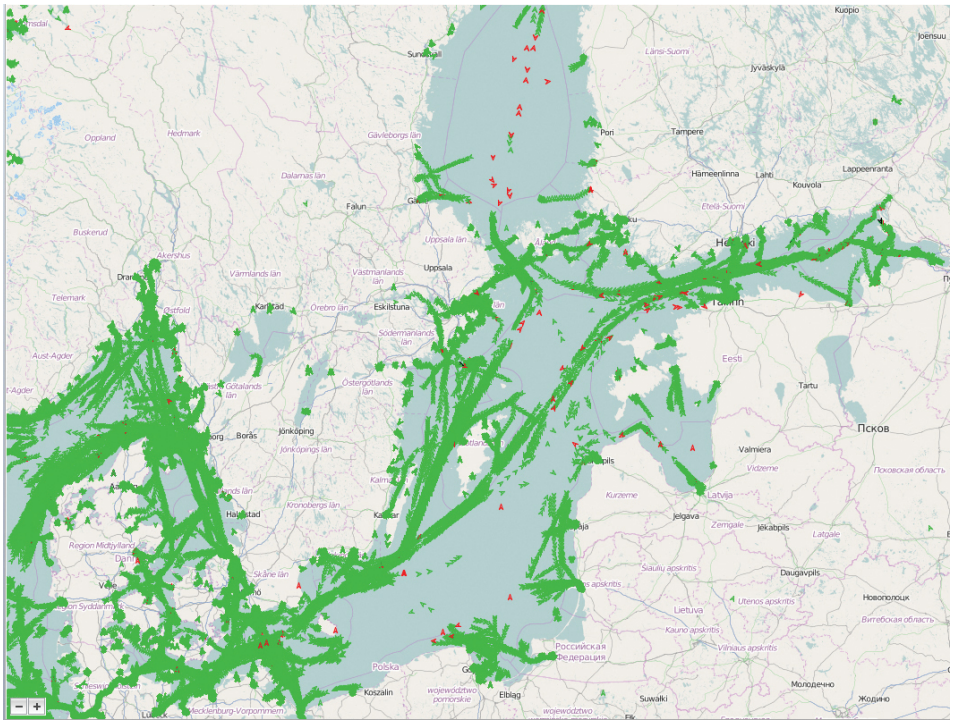


Figure 4.3. An example of AIS coverage on the Baltic Sea

Source: The SimmoViewer application developed within the SIMMO project.

limited. As a result, access gaps, i.e., time periods when a ship is not in view of an AIS satellite and no vessel position can be acquired, still happen.

To sum up, the current capabilities in the area of AIS data provision and utilization are still under development. This especially concerns the integration of data about ships from various sources and the use of intelligent data analysis tools. Even when it comes to AIS data, the usage of terrestrial and satellite-based AIS has not yet been fully exploited. As a result, there are some challenges with regard to the capabilities of maritime surveillance systems.

Long Range Identification System (LRIT)

The Long Range Identification System (LRIT) is another international tracking and identification system incorporated by the IMO under its SOLAS convention to ensure a monitoring system for ships across the world. The LRIT is required of all passenger ships, cargo ships of 300 gross tonnage and above engaged in international voyages, and mobile offshore drilling units. These ships must send reports to their flag administration at least four times a day (i.e., every 6 hours). A vessel transmits its identity, position (latitude and longitude), and the date and time of the position. The system consists of shipborne satellite communications equipment, like INMARSAT or IRIDIUM, and is a point-to-point communication system. The data transmitted within the LRIT is stored in national or regional LRIT Data Centers, which are managed by contracting governments. In the case of the European Union (EU), LRIT data are stored in the EU LRIT Data Centre and are managed by the European Maritime Safety Agency. The data are available only to authorized entities of the Member States.

Vessel Monitoring System (VMS)

Vessel Monitoring Systems (VMS) are tracking systems that are used to track and monitor the activities of commercial fishing vessels. They are mainly used for fisheries management by ensuring proper fishing practices to prevent illegal fishing. A VMS usually covers the territorial waters of a country or Exclusive Economic Zone. Unlike the AIS, it is not standardized globally. Therefore, the functionality of a VMS varies according to the requirements of the nation to which vessel is registered and the regional or national waters on which the vessel is operating. However, the ships under EU flags must send reports based on the EU standard, EU-VMS. The main disadvantages of VMS is that a VMS data are private and not publicly available.

Multi-sensor contact data

A multi-sensor signal generates contact-level data for all available sensors, such as coastal radar, SAR, video, IR, etc.

Coastal High Frequency Radar (HFR) provides regular, high-quality information on ocean surface currents. The HF-Radar provides real-time observational data of the surface currents via coastal stations. Understanding marine currents is of great importance for the development of activities related to maritime transport, since it provides information about the trajectory of a vessel or drifting object. Thus, it allows vessels in the radar range to be tracked.

Synthetic-aperture radar (SAR) is a form of radar that is used to create two- or three-dimensional images of objects, such as maritime areas. SAR uses the motion of the radar antenna over a target region to provide better spatial resolution than conventional radars. An SAR is typically mounted on a moving platform, such as an aircraft or spacecraft.

SAR images have a wide scope of applications in remote sensing and mapping the surface of the Earth. It is also a useful technology in environmental monitoring, foer example oil spills, flooding, urban growth, and global change. Measurements that cover an ocean area can be used to deduce surface waves or to detect and analyze surface features such as fronts, eddies, and oil slicks. SAR can also be implemented as inverse SAR in order to observe moving targets over time (e.g., ships). In the maritime domain, apart from mapping the surface of the sea and oceanography, it is used to detect objects in open seas. Some SAR images are published by the European Space Agency, but access to the data requires prior registration and the submission and approval of a proposal.[7]

Signal Intelligence refers to the capability to detect, characterize, and geolocate various types of radio frequency emitters. Specifically, in the context of maritime surveillance and the detection of non-cooperative ships, signal intelligence data are key. Signal intelligence data are commonly collected by various military stakeholders, but recently private entities are also offering such capabilities, for example HawkEye 360 (US) or Kleos (UK).

Other sensor data include cameras, closed circuit television (CCTV), infra red imaging, and underwater sensors.

Geographic Information System data

A Geographic Information System (GIS) is a system designed to capture, store, manage and analyze spatial and geographic data. GIS datasets can be used in various applications, especially for locating all kinds of phenomena, especially those which vary over time, and for further visualizing them on maps. In the

7. https://earth.esa.int/web/guest/data-access/products-typology/radar-imagery/

maritime domain examples of GIS data are port locations, maritime protected areas, ocean fishing regions, fish species habitat distribution, political national borders and Exclusive Economic Zones, bathymetry, etc. Much of the data is freely available for potential users.

4.1.2. Weather data

There are several sources that provides weather data for maritime areas on a regular basis. They can be grouped into two categories:

- sources providing only forecast data, for example, windy.com, predictwind.com, NOAA;
- sources providing forecast data and historical weather data, for example, yr.no, Copernicus, or the European Centre for Medium-Range Weather Forecasts (ECMWF).

The first group provides only forecast data for a defined number of days in advance (e.g., 5- or 10-day), while the second one additionally offers information about actual weather in the past in the form of daily, monthly, or yearly means. The available weather data sources also differ with respect to the area covered (global or selected local areas), data resolution (from 30 km up to 7 km), update frequency (once or several times a day), the forecast model used, and the scope of the data (the set of weather parameters that can be observed). Moreover, the technical parameters of the available data may vary with regard to the data format (the most popular are grib or NetCDF files, though JSON/XLM formats are also supported), how the data are shared (via API, a webservice, or ftp), and data accessibility (there are fully open and free data sources, such as Copernicus, yr.no, or NOAA National Weather Service), commercial sources with free and paid options available (e.g., windy.com, or predictwind.com), as well as sources available only to authorized users (e.g., ECMWF).

In the study presented in this book historical weather data from Copernicus were used. Therefore, this data source is described in more detail.

Copernicus[8] is the European Union program aimed at developing European information services based on satellite Earth observation. It is managed by the EU and the European Space Agency (ESA). Within this program vast amounts of global data from satellites and seaborne measurement systems are provided. The content is freely and openly accessible to users.

The information services offered by Copernicus can be grouped into six main themes: land, ocean, emergency response, atmosphere, security, and climate

8. http://www.copernicus.eu

change. For the scope of this research, the ocean topic is highly relevant. Copernicus offers sea status observation and forecast information for various parameters like wind, temperature, ice cover, salinity, or chlorophyll. These datasets can be downloaded in an automatically from the data hub.[9]

The main source of maritime weather data is the Copernicus Marine Environment Monitoring Service (CMEMS).[10] The service provides information from both satellite and in situ observations, daily state-of-the-art analyses and forecasts daily, and historical weather data for different maritime areas. The data are available through the CMEMS services that are open, free, reliable, and sustainable.[11]

The Copernicus weather data are stored in NetCDF files—the Network Common Data Form. This is a file format dedicated to sharing array-oriented scientific data. It is also the standard of the Open Geospatial Consortium (Opengeospatial.org, 2018). Version 4.0 (released in 2008) allows for the HDF5 data file format. Hierarchical Data Format (HDF) is a file format designed to store and organize large amounts of data.

The characteristic thing about NetCDF is its capability of self-description. The header of the file describes the layout of the rest of the file, in particular the data arrays. It can also provide arbitrary file metadata in the form of name-value attributes. The NetCDF format is platform independent and there libraries available for all major programming languages.

For the research presented further in Chapter 9, from all available Copernicus services we used only those that provide parameters of interest to our analysis, that is, data about wind (speed and direction), wave height, sea currents and tides, ice coverage, and covering selected maritime areas (i.e., the Baltic Sea, the North Sea, and the Norwegian Sea in the Arctic Ocean). The process of acquiring and extracting weather data from Copernicus is elaborated in Section 4.6.3.

4.1.3. Internet sources

Sensor data, like the AIS, provide only basic information about a given ship. In order to complement the sensor data with relevant information about ships, external sources and databases can be used. A great example might be various Internet sources that publish maritime-related data.

Open Internet data sources can provide general ship data (flag, detailed type, length, gross tonnage, capacity, technical specifications, and construction details),

9. http://marine.copernicus.eu/services-portfolio/access-to-products/
10. http://marine.copernicus.eu
11. The detailed catalogue of services is available at http://marine.copernicus.eu/wp-content/uploads/2016/06/r2421_9_catalogue_services.pdf

ownership data (current and past owners), classification status, classification history, former ship name, flag history, or security related information (bans, detentions, or port state controls). All this information is relevant to characterize a ship and its movement—it gives some context that might be important in detecting anomalies and assessing risk. However, currently such additional information is rarely exploited in analysis.

On the Internet, there are a number of data sources which can be used in order to provide this information. They can be divided into four groups (Kazemi et al., 2013; Stróżyna et al., 2016):

- data sources, in which data is available online and freely accessible to and reusable by the public (no authorization required): open data;
- data sources with authorization required: they provide information to registered users (e.g., Equasis);
- data sources with partially paid access: they provide basic information for free, though access to a wider scope of information requires a fee (e.g., MarineTraffic);
- commercial (paid) data sources: websites with only paid access to the data (a fee or subscription is required).

Further on we focus basically on sources from the first group—open data—by presenting its definition and the methodology for selecting open data sources (Section 4.3) and extracting open data (Section 4.4).

Open data

According to a widely accepted definition "open data and content can be freely used, modified, and shared by anyone for any purpose".[12] The concept is not new but it was popularized by open-data government initiatives such as Data.gov and Data.gov.uk. It was later regulated by European Commission in Directive 2003/98/EC on the re-use of public sector information (European Union, 2003). This directive has an economical goal of facilitating the development of innovative services and the free exchange of market information.

Open data is a movement that is raising interest with its potential to improve the delivery of public services by changing how the government works. It can also empower citizens and create added value for businesses. The reports suggest that open data can unlock $3–5 trillion anually in economic value (Manyika et al., 2013). Further potential can be released by applying advanced analytics to combined proprietary and open knowledge.

Open data is also crucial for a European Single Digital Market. Internet and digital technologies offer new possibilities, which so far have not been fully exploited

12. http://opendefinition.org/

by governments and companies. There is a very strong economic motivation, since "tearing down regulatory walls and moving from 28 national markets to a single one (...) could contribute 415 billion euros per year to the [European] economy and create hundreds of thousands of new jobs."[13]

The term open data refers to the idea of making data freely available to use, reuse, or redistribute without any restriction (Alonso et al., 2009). In the maritime context, there are organisations and communities that provide their maritime related data on-line and make it accessible for the public. Examples include ports and publishing vessel traffic data, as well as blogs, forums, and social networks which share information about maritime events (Kazemi et al., 2013).

The main advantage of the Internet data sources is that they are relatively easily accessible to users (in comparison to sensors or authorized databases). They may reveal also facts which are not reported to the maritime authorities or made available in their databases, providing a global context for data and guaranteeing a lack of legitimate limitations on exchanging data between different countries.

In most cases open data is crowdsourced data, that is, provided by a community of users. This results in certain disadvantages: quality is mentioned as one of the challenges (Węcel & Lewoniewski, 2015). Data may be incomplete, not up-to-date, inaccurate, or incorrect. This problem also concerns the maritime--related sources, since some of them suffer from insufficient quality. One of the approaches to mitigate these deficiencies is to use several sources and then verify the information.

The possibility of using open data in the maritime domain has already been mentioned by Kazemi et al. (2013). They studied the potential to use open data as a complementary resource for detecting anomalies in maritime surveillance. As it was an initial idea and realized in the form of a case study, the scope of the research was limited. In Section 4.3 we present a framework that assumes that open data will be used in the maritime domain, but on a much larger scale: more geographical coverage, longer data collection, and more data sources.

The shallow and deep web

Internet data sources can be also classified according to how easily data can be found on the web. In this regard, we distinguish the shallow and the deep web. The former is that portion of the Internet which can be indexable by conventional search engines and which links billions of HTML pages. The latter consists of online databases that are accessible via web interface to humans, but poorly indexed by regular search engines and consequently unavailable through regular web searches (T. Kaczmarek & Węckowski, 2013). However, it is estimated that the

13. https://ec.europa.eu/priorities/digital-single-market_en

deep web contains far more significant information, which is 'hidden' behind the query forms of searchable databases. Thus, the shallow and deep webs differ in how they structure information. While information on the shallow web is mostly unstructured (HTML text and images), web databases can be both unstructured and structured (K. C.-C. Chang, He, Li, Patel, & Zhang, 2004). Using traditional access techniques designed for the shallow web (e.g., keyword-based indexing) may not be appropriate for the deep web and it is crucial to develop more effective techniques for online databases (He, Patel, Zhang, & Chang, 2007). Such webpages are not directly accessible through static URL links, but are instead dynamically generated as a response to queries submitted through the query interface of an underlying database (K. C.-C. Chang et al., 2004). Moreover, this data remains largely hidden from users, because current search engines are not able to effectively retrieve information (crawl) from these databases. As a result, when it comes to using the Internet as a data source, a few aspects and challenges need to be considered. Getting the data from the deep web is a complex process, which requires an understanding of website navigation and the application of appropriate data extraction and integration techniques (T. Kaczmarek & Węckowski, 2013). Due to the lack of fully automated tools, it has to be carried out manually to a large extent.

Moreover, unstructured data often takes the form of natural language text. An analysis of such text is much more difficult than extracting data from structured documents, due to linguistic ambiguities and oother reasons. To be able to correctly understand such ambiguities, a very broad knowledge about the real world is required and advanced techniques must be utilized (Jackson & Moulinier, 2002). Therefore, Natural Language Processing (NLP) is currently a very active area of research and development, focusing on providing algorithms and techniques of even higher quality.

The shallow and deep web can be interesting and valuable sources of information for maritime information systems. Analysis revealed that there are a number of online databases that contain valuable information on various maritime entities, such as vessels, ports, ship owners, etc. Likewise, on the Internet there are sites and services that publish information about ships and other maritime-related data like reports, statistics, or news. In many cases, the information is published by various maritime organizations, such as the IMO, Memoranda of Understanding, port authorities, coastguards, and private companies. Some examples of data that are available online are below:

- data about ship accidents and reported piracy and terrorist attacks published by the IMO in the GISIS database;[14]
- data about detentions and inspections of ships in different regions of the world.

14. https://gisis.imo.org

The data are published by Memoranda of Understanding, e.g., Tokyo,[15] the Indian Ocean,[16] the Mediterranean,[17] the Black Sea,[18] or the US Coast Guard;[19]
- data about ships and their characteristics available in services like MarineTraffic,[20] or FleetMon,[21] or published by ITU in the MARS database;[22]
- data about the classification of ships and their belonging to classification societies, published by the IACS;[23]
- data about risk indexes of various regions of the world, or countries, published by Inter-Agency Standing Committee (IASC) and the European Commission,[24] the Basel Institute on Governance,[25] and the United Nations University (UNU).[26]

4.2. Maritime data quality

There is no agreed approach among academics for the assessment of data quality in general. In the information systems literature, a lot of various data quality attributes can be found, such as completeness, accuracy, timeliness, precision, reliability, currency, and relevancy (R. Y. Wang, Reddy, & Kon, 1995). Other attributes such as accessibility and interpretability, are also used. R. Y. Wang and Strong (1996) identified nearly 200 such quality attributes. Batini, Cappiello, Francalanci, and Maurino (2009) presented different definitions of popular quality attributes provided in the literature. Heinrich and Klier explained the quality of data as a multidimensional construct embracing multiple dimensions, for example precision, completeness, timeliness, and consistency (Heinrich & Klier, 2015). In terms of the quality of information, a good summary was presented by Eppler, who proposed 70 attributes and then narrowed the list down to the 16 most important (Eppler, 2006).

The methods and criteria for quality evaluation also differ in various domains, such as business, medical, or technical information. For example, the Commission of the European Communities has established dedicated quality criteria for

15. http://www.tokyo-mou.org/
16. http://www.iomou.org/
17. http://www.medmou.org/
18. http://www.bsmou.org/
19. http://cgmix.uscg.mil/PSIX/PSIXSearch.aspx
20. https://www.marinetraffic.com
21. http://fleetmon.com
22. http://www.itu.int/en/ITU-R/terrestrial/mars/Pages/MARS.aspx
23. http://www.iacs.org.uk/shipdata
24. http://www.inform-index.org/
25. https://baselgovernance.org/basel-aml-index
26. http://collections.unu.edu/view/UNU:5763

webpages related to health care. In this case, the quality of a webpage (effectively, of its information) is measured by the following criteria: transparency, honesty, authority, privacy and data protection, updating of information, accountability, and accessibility (Commission of the European Communities, 2002).

Taking into account the fact that there is little agreement on the nature, definition, measure, and meaning of data quality attributes, the European Parliament decided to propose its own uniform standards for guaranteeing the quality of results for the purposes of public statistics, described in the European Statistical System (ESS) Quality Assurance Framework (European Statistical System, 2014). In this standard, seven quality criteria were defined (European Parliament, 2009):

(1) *relevance* (the degree to which the data meet the current and potential needs of the users);
(2) *accuracy* (the closeness of estimates to the unknown true values);
(3) *timeliness* (the period between the availability of the information and the event or phenomenon it describes);
(4) *punctuality* (the delay between the date the data are released and the target date);
(5) *accessibility and clarity* (the conditions and modalities by which users can obtain, use, and interpret the data);
(6) *comparability* (the measurement of the impact of differences in applied measurement tools and procedures where data are compared between geographical areas, sectoral domains, or over time);
(7) *coherence* (the adequacy of the data to be reliably combined in different ways and for various uses).

The quality report according to the ESS should also include additional aspects, such as *cost and burden* (the cost associated with producing the statistical product and the burden on the respondents), *confidentiality* (which concerns unauthorized disclosure of the data) and *statistical processing* (operations and steps performed by a statistical system to derive new information). However, these additional elements may not be included in the case of open Internet sources since they seem to be irrelevant.

Once the quality attributes are defined, the next step is data quality assessment. On this matter as well the literature provides a wide range of techniques to assess and improve the quality of data. In general, the assessment consists of several steps (Batini et al., 2009):

(1) *data analysis* (examining of data schemas, completly understanding the data and related architectural and management rules);
(2) *data quality requirements analysis* (surveying the opinions of users and experts to identify quality issues and to set quality targets);
(3) *identification of critical areas* (selecting databases and data flows);

(4) *process modeling* (a model of the processes that produce or update data);
(5) *measurements of quality* (selecting quality attributes and defining corresponding metrics).

The measurement of quality can be based on quantitative metrics, or qualitative evaluations by data experts or users.

In an approach by Dorofeyuk, Pokrovskaya, and Chernyavkii (2004) a data source is described by three qualities: understandability (a subjective criterion), extent (an objective criterion), and availability (an objective criterion), whereas the efficiency of a given data source is the weighted sum of its quality scores. Weights are calculated using linear programming. An important feature of this method is the fact that it focuses on each data source selectively. The step of quality measurement can be performed with different approaches, such as questionnaires, statistical analysis, and the involvement of experts on the subject (expert or heuristic techniques).

In the context of open data, it is also useful to look into the quality of linked open data sources. The field of linked data quality is relatively new in comparison to well-established publications about data quality, but it has the advantage of the structured approach inherent in linked data. Researchers are concerned not only with regards to the quality of the data sources, but also to the corresponding metadata, which can compromise the searchability, discoverability, and usability of resources.

In the following sections we present our approach to assessing maritime data quality that focus on two aspects: firstly, the quality of AIS data itself (the paragraphs below) and secondly, the quality assessment of open data sources that provide maritime-related information (Section 4.3).

AIS data quality. As indicated in Section 4.1, AIS is currently one of the most commonly used systems in the maritime domain for locating and identifying nearby vessels in real time (a tracking system) as well as for maritime surveillance due to its high update frequency. However, one of the main issues with AIS is the fact that, although it is required of all vessels above 300 GT, the actual use of AIS is at the crew's discretion. The ship's captain may decide whether to switch on or off an AIS transponder and is responsible for providing and updating some of the actual data being sent by the AIS. This, in turn, may negatively influence the data quality.

Data quality covers a broad range of concepts and has multiple dimensions. It is often defined as *fitness for use* with respect to a particular application (Nahari, Ghadiri, Jafarifard, Dastjerdi, & Sack, 2017). In the case of AIS data, the quality assessment can yield information on the reliability and integrity of the AIS data. Thanks to that, the users responsible for surveillance of maritime traffic (e.g., officers in the Vessel Traffic Services) or the crew of other ships may make more

informed decisions. The knowledge of the quality of information provided by ships is of prime importance in situational awareness in the maritime domain. Iphar, Napoli, and Ray (2015a) noticed that although most AIS users do not falsify their data, a certain amount of AIS messages are false and vessels emit or receive messages that are not true. This, in turn, may lead to a lack of trust in the AIS system or incorrect decisions of various maritime actors. Therefore, the maritime society stresses the importance of assessing the quality of AIS data.

Due to its basic characteristics, the AIS system is vulnerable to various quality issues. These problems may result, first of all, from the improper installation of an AIS device. The static information, which is entered manually, is not supervised by any authority and thus may be inaccurate. The dynamic information, in turn, depends on proper communication between an AIS device and other sensors on board (e.g., GPS antenna). Other problems may arise due to human errors and the behavior of a ship's crew. Since the AIS is a self-reporting system, some actors may intentionally provide erroneous data in order to hide their activities, may unintentionally make errors when manually entering the data into the device, or may spoof the AIS signal to mislead other actors (Iphar et al., 2015a). Another group of quality issues is related to the AIS system itself, namely, the limited coverage of the AIS in some areas due to weak satellite reception in areas of high vessel density, such as the North Sea or the Baltic Sea. Moreover, AIS reception on the open sea (outside the terrestrial coverage) is also limited due to access gaps—i.e., time periods when a ship is not in view of any AIS satellite—and consequently no vessel position can be acquired. These can last even a few hours.

The existing studies show that the quality issue in AIS data is a common problem in all three mentioned aspects. The analysis conducted by (Harati-Mokhtari, Wall, Brooks, & Wang, 2007) indicated that errors most often concerned the unique identification number (MMSI), defined vessel type (undeclared or default type), ship's name and call sign (no name provided, or abbreviations), navigational status (incorrect status), incorrect vessel length, reported draught (non-availability, draught greater then length, or inaccurate value), destination, and ETA (vague or incorrect entries). Iphar et al. (2015a) stressed also intentional falsification of AIS signal, identity theft (duplication of MMSI number), and destination masking. They also indicated the problem of switching off the AIS transponder in order to hide certain activities.

The quality analysis conducted by Tu et al. (2017) focused on the completeness and resolution of AIS data. They concentrated on four aspects: position precision, the time interval between two consecutive AIS messages, data completeness, and erroneous/corrupted entries. Their results indicate that ships' positions are rarely invalid, but errors in heading and status are quite common. The SOG and COG data are also sometimes wrong. With regard to the completeness of dynamic information, most data contained the necessary kinematic information (i.e., position, time,

speed, and course), but the remaining dynamic information was very often missing. They also identified three types of errors: 1) infeasibly large / small SOG values; 2) duplicated AIS messages; and 3) missing AIS messages due to errors in data broadcasting or VHF transmission. AIS is also vulnerable to different threats, such as spoofing, hijacking, and availability disruption (Coleman et al., 2020). Balduzzi, Wilhoit, and Pasta (2014) indicated three types of AIS spoofing: falsification of the closest point of approach alert, imitation of a fake ship which follows a given path, and simulation of a search and rescue alert.

In our study we conducted a quantitative assessment of AIS data quality to investigate whether the real data meets data quality standards and to identify the most common quality issues. We also tried to assess the scale of the issue where ships do not provide data of proper quality. Our approach was essentially based on statistical analysis of the data and an assessment of its reliability based on a set of quality attributes. To this end, a sample of real AIS data was retrieved and analyzed. These quality attributes included the following (Pipino, Lee, & Wang, 2002):

- completeness: the extent to which data are not missing and are sufficient for the task at hand;
- free-of-error: the extent to which the data are correct and reliable;
- ease of manipulation: the extent to which the data are easy to manipulate;
- timeliness: the extent to which the data are sufficiently up-to-date for the task at hand;
- reliability: the extent to which the data are regarded as true and credible.

In order to assess the quality of available AIS data with regard to the above attributes, AIS data from two data sources were used:

- AIS data received January-December 2015 from Orbcomm satellites (for the whole globe). The dataset contained 1,390,219,742 messages from 425,166 unique vessels and the analysis focused on vessel type, draught, dimensions, and destination.
- AIS data set covering weeks 33–35 of 2018 from the whole globe. The dataset contained 65,896,367 messages. It was used to analyze the following AIS attributes: navigational status, speed over ground, course over ground, true heading, IMO number, call sign, and name.

In order to process such a vast dataset, we used Apache Spark, a popular data processing engine, which can take advantage of in-memory computation.

The analysis concerned both static and dynamic AIS parameters. Among the static AIS parameters, vessel identification data (IMO number, call sign, and name), vessel dimensions, and vessel type were analyzed. With regards to the dynamic AIS parameters, vessel draught, navigational status, destination, speed over ground (SOG), course over ground (COG), destination, and location were analyzed. The results are presented in the following paragraphs.

Vessel identification. There are several ways of identifying ships: MMSI, IMO, call sign, and name; the last one is non-unique. Our analysis found many incorrect values for ship identification number (IMO). In the data, there were 47,791 unique IMO numbers, out of which only 45,598 values were 7-digit numbers (as the standard requires). The rest—almost 2,200 identifiers—were definitely incorrect. Moreover, a value of zero is suspiciously frequent.

Regarding the call sign, there were 103,268 unique values. Call signs were missing in 9,463,167 messages, i.e., 14.4% of the dataset. The most popular call signs were numerical ones (e.g., '700', '300', '0', or '200'). Normally, call signs for larger vessels should consist of the national prefix plus three letters. Among the most popular ones, there were none that met this requirement. Instead, there were obviously incorrect values, including dashes, NONE, or CH.16, for example.

The name of the vessel, as with the call sign, is not a unique value. We identified 152,473 various names in the data set. The name was provided more often than the call sign—almost 4 million messages did not contain a vessel name, representing 6.1% of all messages.

Vessel type. The next attribute analyzed was the type of ship. There are several classes allowed to be used in AIS messages (a two-digit value). Nevertheless, in the sample data we identified 211 different values for the type of ship. There were no missing rows—even if there was no IMO or call sign, the type of ship was always filled in. Apart from the types agreed for the AIS standard, there were also unknown three-digit types. Moreover, almost one fourth of the ships (23.07%) had provided a default value for ship type, making it impossible to specify what kind of ship it is. The overall distribution of vessel types, calculated for all vessels that sent an AIS message in 2015, is presented in Table 4.3. If we analyze the number of vessels, cargo vessels prevailed in the ranking (33.83% vessels), followed by fishing vessels (18.77%) and tankers (5.74%).

Vessel dimensions. Vessel dimensions (length and width) are another static parameter that should be provided in AIS. There are four vessel dimensions available in AIS data: to bow (A), to stern (B), to port (C), and to starboard (D). The default value for all of them is 0. Vessel length can be calculated from $A + B$, and vessel width from $C + D$. The analysis revealed that only 70% of the vessels provided their dimensions. For the rest, the values were missing. The default value is 0, which means it has not been set by the operator. Some operators set only some of these values, leaving the rest as default. In general, our analysis showed that tankers and cargo vessels provided the most reliable results.

Location. Ship location is an item that should be reported regularly in the AIS. In our study, vessel locations were analyzed based on geographical coordinates.

Table 4.3. Vessel types recorded in AIS between January and December 2015

Type	Number of AIS messages (%)	Number of vessels	Number of vessels (%)
Anti-pollution equipment	0.16	615	0.14
Cargo	36.83	141, 580	33.30
Diving ops	0.16	417	0.10
Dredging or underwater op	1.66	2, 383	0.56
Fishing	6.58	79, 783	18.77
High speed craft (HSC)	1.15	2503	0.59
Law enforce-ment	0.76	2, 029	0.48
Medical transport	0.02	179	0.04
Military ops	0.37	1, 456	0.34
Non-combatant ship	0.02	132	0.03
Not available (default)	7.59	98, 101	23.07
Other type	6.43	13, 817	3.25
Passenger	7.16	10, 743	2.53
Pilot Vessel	1.07	1, 709	0.40
Pleasure Craft	2.25	5, 564	1.31
Port Tender	0.23	881	0.21
Reserved	0.27	2, 948	0.69
Sailing	0.70	3, 490	0.82
Search and Rescue Vessel	0.86	2, 322	0.55
Spare—Local Vessel	0.06	343	0.08
Tanker	13.99	24, 389	5.74
Towing	2.33	5, 229	1.23
Tug	7.97	11, 663	2.74
Undefined or empty	0.94	10, 588	2.49
Wing in ground (WIG)	0.41	2, 302	0.54

Source: (Stróżyna, Eiden, Filipiak, Małyszko, & Węcel, 2016a).

Basically, as expected, the coverage of the analyzed data is worldwide (see Figure 4.4, which presents vessel traffic in weeks 33–35 of 2018). The data cover almost every point on the globe; however, such a distribution might be suspicious, as not every point on the earth is reachable by vessels, especially terrains close to the North or the South Poles. This might be explained by what is called AIS spoofing or an incorrect configured GPS device on board (Iphar et al., 2015a).

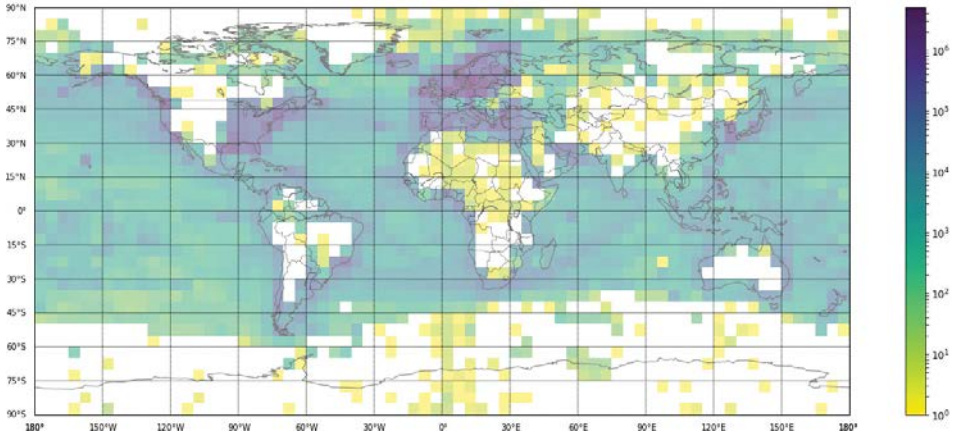


Figure 4.4. Frequency of messages visualized on the map of the whole world—logarithmic scale

Source: (Stróżyna et al., 2016a).

Vessel draught. Information about the current draught should be regularly updated by the captain. However, due to the fact that this information is often entered manually (set up statically), it is of poor quality. This was confirmed by our analysis. The vast majority of vessels (79.13%) reported only one draught value in 2015, a fact which can be interpreted as the value not being updated. The rest of the vessels (20.87%) updated this value more or less regularly. The average value of draught for these vessels was 14.68 m (with a standard deviation of 15.87 m)—see Table 4.4 for the different vessel types. A quick glance at the results reveals that an average reported draught can vary significantly between different vessel types. The minimum and maximum values were omitted from the table, since for virtually all types they ranged between 0 and 25.5 meters.

Table 4.4 also highlights some basic statistics about the total number of updates of the draught value, by vessel type. There was a significant difference across vessel types. Tankers reported around 13 draught values in 2015 on average, cargo vessels reported 6. This means that for these two types the draught value is updated more often, whereas for fishing vessels updates are much rarer. Thus, a typical cargo vessel or tanker updated this value accordingly once every 13 or 14 days, which is the highest figure among the types, while high speed craft (HSC) and pleasure crafts updated it only once every 69 and 64 days, respectively. However, for the latter vessel types this might result from the fact that their draught actually does not change.

Navigational status. The next attribute under analysis was the navigational status of the ship. As can be seen in Figure 4.5, its distribution seems reliable and there is

Table 4.4. Draught statistics for different vessel types recorded between January and December 2015

Type	Draught in metres (mean)	Draught in metres (std. dev.)	Total distinct values of draught (mean)	Days to draught change (mean)
Anti-pollution equipment	3.46	3.99	1.62	45.91
Cargo	7.09	3.98	6.38	13.29
Diving ops	4.28	3.70	2.33	56.53
Dredging or underwater ops	4.21	3.91	3.12	47.28
Fishing	0.90	2.43	1.13	57.88
High speed craft (HSC)	3.49	4.61	1.93	69.56
Law enforcement	2.56	3.62	1.81	56.69
Medical transport	8.05	7.32	1.33	36.56
Military ops	4.09	4.38	1.84	52.11
Non-combatant ship	6.70	7.04	1.77	37.03
Not available (default)	1.25	3.07	1.28	24.58
Other type	4.92	4.08	3.48	34.45
Passenger	4.52	4.13	2.82	33.24
Pilot Vessel	3.34	4.42	2.10	30.34
Pleasure Craft	2.41	3.13	1.83	64.03
Port Tender	4.03	4.60	1.80	29.55
Reserved	1.93	3.03	1.49	47.30
Sailing	2.44	2.98	1.47	58.15
Search and Rescue Vessel	3.06	3.49	1.71	56.46
Spare—Local Vessel	6.12	5.99	1.85	45.08
Tanker	7.84	3.84	13.05	14.06
Towing	3.94	3.85	1.83	53.70
Tug	5.01	3.50	3.40	32.93
Undefined or empty	5.76	6.98	1.79	28.66
Wing in ground (WIG)	4.22	5.08	1.87	32.71

Source: (Stróżyna et al., 2016a).

no issue with this attribute in the analyzed data. The vast majority of ships traveled with the status ‘under way using engine’ (status 0). Other popular statuses were: ‘moored’ (5), ‘at anchor’ (1), and ‘engaged in fishing’ (7). However, there are some vessels that sent the ‘default’ status (15).

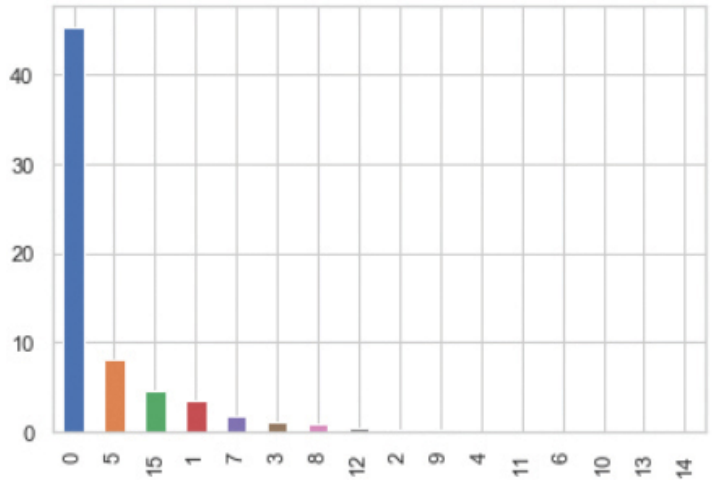


Figure 4.5. Navigational status

Source: (Stróżyna et al., 2016a).

SOG and COG. Another dynamic attribute in the analysis was speed over ground (SOG). Its logarithmic distribution is presented in Figure 4.6 (a). We can observe several values that are likely used as a default for missing values (peaks in the chart).

Several outliers were identified for course over ground (COG), as can be seen in Figure 4.6 (b), which is also presented on a logarithmic scale. Moreover, 0 and 360 were the most common values, 10 times more frequent than any other value (please note the log scale). This may mean that the default values were not replaced. There were also some values greater than 360.

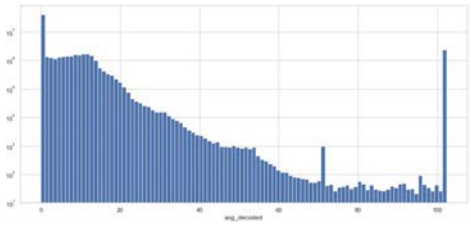


Figure 4.6 (a) Speed over ground (logarithmic scale)

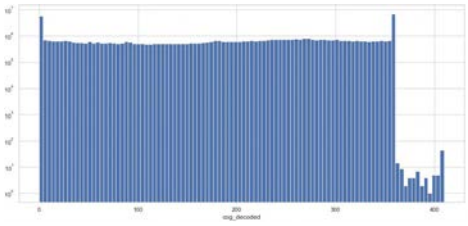


Figure 4.6 (b) Course over ground (log scale)

Source: (Stróżyna et al., 2016a).

Destination. As with the draught value, the destination should be set up manually and updated regularly by the captain. An initial statistical analysis was conducted to see whether the values of destinations they provided were valid. To explore

the declared destinations, firstly the data was cleaned by removing the special character "" and trimming it (removing leading and trailing spaces). The most popular destinations are presented in Table 4.5. Notice that a more sophisticated method is needed to obtain more robust results. For instance, "ANTWERP" and "ANTWERPEN" refer to the same port. The analysis concerned values of destinations and how often this parameter was updated. Unfortunately, the completeness, and hence the quality, of this variable is not satisfactory. Even such an initial analysis found an empty destination in 275,338,472 messages, which is over 20% of the messages, whereas 0 was set in 1,796,596 messages. Interestingly, "HOME" was declared in 1,815,822 messages.

Table 4.5. Most popular destinations in AIS data between January and December 2015

Destination	Number of AIS msg	Number of AIS msg (%)
(empty)	275, 338, 472	20.32
ROTTERDAM	20, 000, 984	1.48
AMSTERDAM	10, 167, 975	0.75
ANTWERPEN	6, 694, 372	0.49
SINGAPORE	6, 578, 468	0.49
HAMBURG	6, 182, 447	0.46
ANTWERP	5, 335, 668	0.39
SHANGHAI	4, 510, 860	0.33
NOVOROSSIYSK	4, 046, 389	0.30
TIAN JIN	3, 921, 142	0.29
TIANJIN	3, 610, 640	0.27
CONSTANTA	3, 410, 220	0.25
SHANG HAI	3, 374, 164	0.25
BREMERHAVEN	3, 286, 316	0.24
HARLINGEN	3, 222, 434	0.24

Source: (Stróżyna et al., 2016a).

The results show that for each analyzed AIS attribute, some problems can be noted, which in turn negatively influence the quality of the AIS data. These problems are observed for both the static AIS parameters (vessel identification, vessel type, and dimensions) and dynamic attributes (location, draught, and destination).

The most common quality issues with AIS data revealed in this study were:

- Duplicated identification numbers (MMSI). Usually, one can find a number of vessels with the same MMSI number and different types declared. For instance, for MMSI 123456789 there were 22 types assigned, while a vessel with MMSI 2443870000 declared 5 different types during 2015.

- No change to the default values in AIS transponder. This concerns, for example, the value of draught, dimensions, and navigational status.
- No update or a relatively infrequent update of dynamic values during ship operation (e.g., draught, or destination).
- Empty field or incorrect values (not meeting the standard requirements) in the destination or vessel identification attributes.
- Location spoofing or incorrect configuration of a GPS device, resulting in incorrect positions of ships.

In summary, there are two main reasons for the problems with AIS quality: wrong (purposeful or not) configuration of an AIS transponder (not changing the default values while installing the transponder on board) or the ship captain not updating the dynamic AIS attributes (intentional or not). Such a situation may also result from the cooperative nature of the AIS—although ships are required to use the AIS, there is no means to actually control whether ships provide correct values. This limits various maritime actors' awareness of the situation and impacts decision-makers' analysis of the situation.

4.3. Data enhancement

As presented in Section 4.1, there are different kinds of data sources in the maritime domain that provide heterogeneous maritime-related data. This data can be used to provide additional information about various maritime entities, which may be important from the point of view of maritime surveillance, threats detection, maritime transport monitoring, or risk assessment. Data from the Internet can be used to enhance proprietary data (e.g., sensors) and is fused with data from other sources, such as legacy systems or internal databases, thus facilitating the entity in conducting its operations based on the broader understanding of its environment.

However, in the existing maritime systems, usually only data received from sensors are used (Rhodes, Bomberger, Seibert, & Waxman, 2005; Vespe et al., 2008). Non-sensor data includes expert knowledge, for example, which might be further integrated with sensor data (Helldin & Riveiro, 2009). Mano, Georgé, and Gleizes (2010) proposed a system that collects data from radars and databases such as an environmental database, Lloyd's Insurance, and TF2000 Vessel DB. Ding, Kannappan, Benameur, Kirubarajan, and Farooq (2003) in turn proposed an architecture of a centralized integrated maritime surveillance system for Canadian coasts, fusing HFSWR, Automatic Dependant Surveillance reports, visual reports, and radar. The only research, which focuses on using open data available on the Internet for the purpose of maritime surveillance, is presented in (Kazemi et al., 2013).

In order to utilize potential stemming from availability of data on the Internet, potential open data sources first need to be identified and selected. To this end, an

appropriate framework/procedure for identifying, assessing and selecting online data sources must be applied. Such a framework should focus above all on the sources' quality and the data quality.

The standard approach to data quality defines quality as "the totality of features and characteristics of a product or service that bears its ability to satisfy stated or implied needs" (International Organization for Standardization, 1986). In the case of a business, these needs are expressed in the form of users' requirements. Therefore, it might be assumed that each potential data source should be analyzed and assessed taking into account two aspects: users' requirements and a set of selection criteria.

In this context, we made use of the quality attributes defined by the European Union (presented in Section 4.2): relevance, accuracy, timeliness, punctuality, accessibility and clarity, and comparability and coherence. They were used to assess the quality of data published by various Internet sources and select data sources for our case study—the SIMMO system presented in Section 4.6. The definition and measure of these attributes for SIMMO purposes are described in Sections 4.3.3 and 4.6.2.

4.3.1. Source selection method

This section presents a framework for identifying, assessing and selecting online data sources to be used by maritime entities, along with data from existing sources (e.g., sensor data) in a decision-making process. The proposed framework consists of the following steps:

(1) the identification of potential data sources;
(2) the assessment of data sources, including the definition and selection of quality criteria and the final selection of sources for a system;
(3) the design and development of the data retrieval procedure, including the definition of the cooperation model, the development of the data acquisition methods and the fusion of data.

The steps of the framework are presented in Figure 4.7.

In the following paragraphs, the steps of the framework are briefly described. In Section 4.6, the framework is described in more detail based on a case study from the SIMMO project.

4.3.2. Identification

The first step focuses on identifying potential data sources available on the Internet and related to a given domain or a given issue. Both the shallow and deep web should be considered as potential data sources.

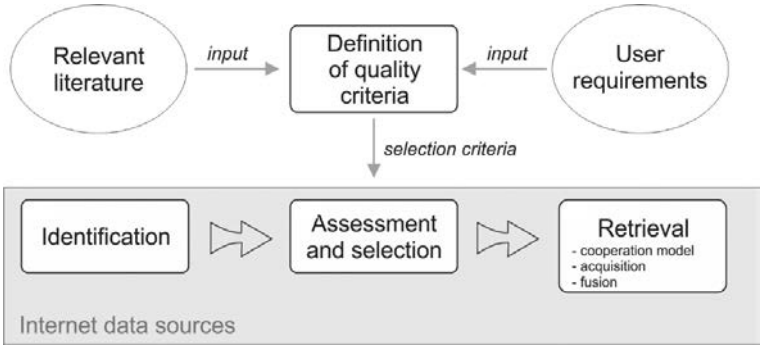


Figure 4.7. Source selection framework

Source: (Stróżyna et al., 2018).

Various search engines may be used to identify sources, including traditional ones like Google, or Bing, as well as meta-search engines or domain-specific ones (if they exist). Apart from search engines, a review of relevant literature should also be conducted, since there might be information or suggestions about which data sources are used in a given domain. Finally, (if possible) domain experts or future users should also be consulted, as they may also suggest potential sources of data (He et al., 2007).

4.3.3. Quality measures

The identified data sources should be assessed from the point of view of both the quality of the source and its compatibility with the users' requirements. At first, the assessment assumes a definition and selects quality criteria. This selection may result from uniform definitions or standards which are used in a given domain, if such a standard exists.

In our approach we propose adopting the data quality measures of the European Statistical System (2014). This selection was driven by the fact that in the case of maritime systems, there are no standards or procedures which would suggest or dictate the set of quality criteria to be used. The previous research on data quality in the maritime domain mainly concerned the quality of AIS data (such as the completeness, accuracy, integrity, etc. of AIS messages) (ABP Marine Environmental Research Ltd, 2013; Harati-Mokhtari et al., 2007; Iphar, Napoli, & Ray, 2015b). However, the methodologies and quality attributes proposed there do not fit well to the assessment of other maritime-related sources of data, especially data published on the Internet. As a result, we decided to look for commonly used approaches in other domains, like the one used in statistics.

We are aware that this framework is used mainly by statistical systems and originally may not fit well to quality assurance in all domains. Nevertheless, we believe that these basic quality criteria are rather universal and may be used after minor modifications in various areas of research. In our case, we made some modifications to adopt them to the characteristics of the online data sources.

In the practice of Eurostat, some of the quality criteria are combined (e.g., accuracy and reliability, timeliness and punctuality) (European Statistical System, 2014). We did the same in our research, which ultimately resulted in six quality measures. Each measure used in the proposed framework is briefly described below:

- *Accessibility*: the possibility to retrieve data from a source; it includes such aspects as the structure of a source, the technologies used in its development, the form in which the data are published, and the source's stability (changes in structure, errors, or unavailability of a service); it also takes into account the terms of use, privacy policy, requirements for login or registration, access to data (fees or subscriptions), etc.;
- *Relevance*: what kind of information is provided by a source and whether this information matches the users' or system's requirements;
- *Accuracy & Reliability*: the reliability of the information provided from the point of view of the users' requirements; it also evaluates data scope and coverage (how much information is available) and data accuracy (missing information);
- *Clarity*: the availability of an appropriate description or explanation of the data model and information about the source of information (data provider);
- *Timeliness & Punctuality*: data update (the time interval between an event and the data which describe it becoming available) and the time delay in publishing updated information;
- *Coherence & Comparability*: whether data provided in a source describes the same phenomenon or has the same unit of measure as data from other sources.

4.3.4. Assessment and selection

Once the quality measures are defined, the next step is to assess othe potential online data sources according to these measures. A systematic approach should be followed. In the framework, we propose using expert knowledge (domain experts assess and select data sources) by assessing sources based on the Delphi method with elements of the Analytical Hierarchy Process (AHP) proposed by Saaty (1990).

The Delphi method (Brown, 1968) relies on a group of experts; its aim is to achieve the most reliable consensus on a given issue. The method is a systematic approach—it consists of rounds in which the experts answer questionnaires and provide their judgements on a given topic. After each round, a facilitator provides

an anonymous summary of experts' opinions. In the next round the experts are encouraged to revise their earlier answers in the light of the replies of other experts. The process is continued until a consensus or a predefined stop criterion is reached (e.g., a number of rounds).

In the approach presented herein, the standard Delphi is enhanced with some characteristics of the AHP, which adds priorities (weights) to the decision-making factors (in this case quality measures). Moreover, following the AHP, the experts in the Delphi method are asked to evaluate a data source under a particular criterion using a four-level rating scale (high, medium, low and N/A), which are then converted into numerical values (accordingly, high $= 3$, medium $= 2$, low $= 1$, and N/A $= 0$) and normalized. Based on these evaluations, a final quality grade is calculated for each source.

Having the quality grade calculated, a final selection of data sources for a given system may take place. Here, we propose to define a threshold value for the quality grade, above which a source is selected. In order to define the threshold, again the Delphi method may be used.

4.4. Data extraction

In the final step of the framework, the design and development of the data retrieval procedure is foreseen. It includes the definition of the cooperation model and the development of the data acquisition methods. The method depends mainly on the type of source and format of publishing data. In Section 4.6.3, examples of retrieval methods for four different categories of data sources are presented.

Cooperation with data owners

The model of cooperation with deep or shallow web sources (data owners) requires the definition of aspects connected with web crawling, i.e., the crawling policy. Web crawling is a process performed by an intelligent agent (computer program), which visits and automatically retrieves the content of web-pages (Kobayashi & Takeda, 2000). The crawling policy consists of the following policies, which define the behavior of a web crawler (Castillo, 2004):

- a selection policy, which defines the pages to be downloaded;
- a re-visit policy, which defines when to check for changes on the pages;
- a politeness policy, which defines how to avoid overloading of websites.

Selection policy. A selection policy defines which pages should be downloaded by a crawler. As a crawler always downloads just a fraction of a web-site, it is highly desirable that the downloaded fraction contains the most relevant pages, and not just a random sample. This requires web-pages to be prioritized, which is a difficult task because the complete set of web-pages is unknown during crawling (Castillo, 2004). There are various strategies for crawling scheduling, based on website popularity in terms of links or visits (Cho & Garcia-Molina, 2003) or a similarity of a page to a given query (Diligenti, Coetzee, Lawrence, Giles, & Gori, 2000).

Re-visit policy. The Internet is a very dynamic in nature, and crawling even a fraction of it can take weeks or months. By the time a web crawler has finished its crawl, many events could have happened, including creations, updates and deletions. Not retrieving or retrieving such events late, and thus having an outdated copy of a resource, can result in late detection of or failing to detect important information about ships. Therefore, a cost function might be specified for each sourc, for example freshness, saying whether the local copy of data is accurate or not. Therefore, the objective of a web crawler is to minimize the fraction of time pages remain outdated by keeping the average freshness of pages in a collection as high as possible, or to keep the average age of pages as low as possible (Coffman, Liu, & Weber, 1998).

Different re-visit policies might adapted (Cho & Garcia-Molina, 2003):

- a uniform policy: involves re-visiting all pages in the collection with the same frequency, regardless of their rates of change;
- a proportional policy: involves re-visiting the pages that change more frequently more often.

In terms of average freshness, the uniform policy outperforms the proportional policy (Cho & Garcia-Molina, 2003).

Politeness policy. Crawlers can have a negative impact on the performance of a website, due to the fact that they retrieve data more quickly and in greater depth than human searchers. In cases where a single crawler is performing multiple requests per second and/or downloading large files, it might lead to server overload, especially if the frequency of access to a given server is too high.

One of solutions, utilized by websites owners is the robot exclusion protocol, also known as the robots.txt protocol. It is a standard used by website administrators to indicate which parts of their webservers should not be accessed by crawlers (Koster, 1996). However, this standard does not include a suggestion for the interval of visits to the same server, even though such an interval is the most effective way of avoiding server overload. In case no extra “Crawl-delay”: parameter is defined in the robots.txt file (which indicates the number of seconds of delay between requests), different practices are used by crawlers’ developers—from a 10-second

interval for access (Cho & Garcia-Molina, 2003; Heydon & Najork, 1999) to even a 1-second (Dill et al., 2002).

Taking into account the politeness policies, a separate cost-benefit analysis is needed for each web crawler for a data source, and ethical considerations should be taken into account when deciding where to crawl and how fast to crawl.

Crawling the deep web

The deep web encompasses pages, which are typically accessible only by submitting queries to a database. Regular crawlers are unable to find these pages if there are no links that point to them. Deep web crawling also multiplies the number of web links to be crawled. A popular approach to target deep web content is to use a technique called "screen scraping". This software automatically and repeatedly queries a given web form with the intention of aggregating the resulting data. Data extracted from the results of one web form submission can be taken and applied as an input to another web form, thus establishing continuity across the deep web in a way that is not possible with traditional web crawlers (Shestakov, Bhowmick, & Lim, 2005).

4.4.1. Data fusion and disambiguation

In cases where data are obtained from many heterogeneous sources, they must be fused, that is, a common data model that meets the initial system requirements has to be developed and used to organize new data in a homogeneous and integrated form. The concept of using additional data (e.g., from the Internet) as a complementary resource for existing data (e.g., sensors) is not an entirely new idea. This concept is called data fusion and is often described along with the JDL Model described by M. J. Hall, Hall, and Tate (2000) and D. L. Hall and McMullen (2004).

Data fusion is a complex process and there are a few challenges which need to be faced in order to successfully conduct it. Firstly, there are semantic interoperability problems related to the interpretation of data coming from different sources. Moreover, data imperfection, correlation, inconsistency, disparateness, and ambiguity are other problems. Finally, in each data source the same entity may be referenced in different ways and different categories may be used to describe the same issues. For example, different words may be used to name the same entity (e.g., a port or a ship) or categories used in two sources may be developed on different levels of granularity. Therefore, before the data are added to the database, such differences must be recognized and the data need to be aligned. For example, the system should recognize which entities are being referenced in the data and, based on that, should assign to this data a unique identifier, which

can be easily used for subsequent analysis. This process is called disambiguation (Małyszko, Abramowicz, & Stróżyna, 2016). In this section, we present a set of methods which might be used to solve this issue. In Section 4.6.3, examples of the methods developed for disambiguation of popular maritime entities, used in our case study—the SIMMO system—are presented.

Data fusion and disambiguation can be related to extract, transform, load (ETL)) tasks. ETL refers to a process in a database usage, especially in a data warehouse, that does the following:

- Extracts data from homogeneous or heterogeneous data sources; in ETL, these are usually databases which may be accessed directly or using dedicated API. In traditional ETL research, an important issue is reducing the overload of the data source resulting from data extraction (to ensure, that the performance of the original data source will not suffer) and, at the same time, keeping the data as up-to-date as possible (Vassiliadis, 2009).
- Transforms the data in order to store it in the proper format or structure, for the purposes of querying and analysis; transformation steps used here are often ad-hoc, developed to fit a given situation, and straightforward if studied individually. Still, as the number of such transformation steps may grow, a proper approach should be utilized to ensure efficiency and elegance in terms of semantics (Vassiliadis, 2009).
- Loads the data into the final target (database or data warehouse) for possible exploitation.

In the case of fusing data from different sources, the traditional ETL process needs to be extended to include entity disambiguation. In the literature, disambiguation is also referred to as duplicate detection, record linkage, reference matching or entity-name clustering and matching problem (Bilenko & Mooney, 2003). It is a well-known problem in the area of data integration that results from the fact that references to a single entity may vary in different sources for various reasons, such as typographical errors, abbreviations, etc. Moreover, in information systems the same entities may be referenced in completely different ways due to the fact that they are developed and maintained by various parties (Rahm & Do, 2000).

According to Rahm and Do (2000), the following steps should be followed to conduct disambiguation:

(1) **Data analysis**, whose goal is to identify errors and inconsistencies in data, which then need to be removed.
(2) **Definition of transformation workflow and mapping rules**, which then are used in the methods for data disambiguation.
(3) **Development of disambiguation methods**, which assumes implementation of rules from the previous step.

(4) **Verification**, whose goal is to evaluate whether the methods developed in the previous step give the expected results; this step, together with the previous ones, may be performed iteratively multiple times.
(5) **Transformation**, which processes the data using the selected methods and stores the results in a database.
(6) **Back-flow of cleaned data**, which assumes an update of a source database with new, cleaned data (if possible).

The basic tools used for the entity disambiguation problem are string similarity measures. These measures return a numerical value that represents a distance (or a similarity) between two strings. Based on them, two very similar strings may be recognized as referring to the same entity (the difference between them may result from a simple misspelling, for example (Alberga, 1967)). Well-known string similarity measures are the Levenshtein distance (Bilenko & Mooney, 2003) and the Jaro distance (Jaro, 1989).

Still, there is another challenge if there is no single, uniquely identifying attributes for a certain entity. In such situations, multiple attributes must be compared to establish some measure of similarity between the two records (Elmagarmid, Ipeirotis, & Verykios, 2007).

To successfully perform entity disambiguation, lexical resources may also be used to identify different ways a certain entity may be referenced. In Wentland, Knopp, Silberer, and Hartung (2008), one of the resources that was used for entity disambiguation was a *Disambiguation Dictionary*, which maps all ambiguous proper names to a set of the unique entities they refer to. If we assume a situation where the abbreviation ACC is used to refer to an entity known as the American College of Cardiology (the main name), such a mapping cannot be easily identified using the Jaro distance, for example. This problem may be solved if there is a proper dictionary in which alternative names for known entities are defined. Additionally, in many situations, such mappings may be ambiguous, (e.g., ACC may also mean the Asian Cricket Council). To resolve such difficulties, additional data must usually be taken into account (e.g., the context).

Data fusion is also an important process in the development of maritime surveillance systems. The existing solutions in this area focus mainly on data fusion techniques for sensor data, such as AIS, Vessel Traffic Services, radars, or video cameras (Kazemi et al., 2013). A more sophisticated approach—which assumes the enrichment of sensor data with open data, data available from various databases, or data stored in structured or unstructured documents (e.g., webpages, historical reports, and comments on ships' behaviors) has not yet been found (Brax, 2011). Therefore, in Section 4.6.3 a method for fusing and disambiguating maritime sensor data with data retrieved from open Internet sources is presented.

4.4.2. Data processing and analysis

Accessing and gathering the required maritime data is one key task. However, analyzing large sets of collected data, deriving the relevant information, and making timely interpretations, assessments, and reasoning on the situation are of the utmost importance.

According to maritime experts (Riveiro, Falkman, & Ziemke, 2008), the existing maritime surveillance systems lack some important features which should be improved upon:

- extension of the maritime area covered, since most of the existing systems include only coastline areas or Exclusive Economic Zones of a state;
- addition of ancillary information about ships, their current and historical routes, and the marine environment;
- detection of standard ship routes and the distribution of traffic at different times of the day, month, and year;
- detection of ships that require special attention due to high risk.

Therefore, systems that would allow retrieval, fusion, storage and analysis of sensor and non-sensor data from various maritime sources seem to be particularly needed. Such systems already exist on the market, but they are based on traditional architectures and approaches for data processing—centralized, relational database systems, SQL-based applications for managing and accessing data, clearly defined structured formats, static schemas, applications that require data to be loaded from disk into memory in order to process data, etc. These approaches and architectures are costly and known for their inefficient and poor scalability when large volumes of data need to be processed (Trujillo, Kim, Jones, Garcia, & Murray, 2015).

Considering sensor data (such as millions of AIS messages), additionally enriched with ancillary information derived from other sources, the amount of data to be analyzed becomes huge. For example, in the maritime domain a key role is played by vessel tracking data, namely, AIS messages. AIS data forms a continuous data stream—therefore, traditional methods relying on one physical machine might be computationally inefficient. Some studies have reported that processing of one month's worth of AIS data takes one day (Marine Management Organisation, 2014; Shelmerdine, 2015). Additionally, the individual responsible for security and safety management or risk assessment who must interpret all the available data, even in a restricted area of interest, would certainly be overwhelmed.

Therefore, firstly, information systems and tools with advanced analysis and reasoning methods which would provide real-time assessment of the situation and support users in risk assessment are required (Pallotta et al., 2013). Secondly, although the enhanced data create a great potential for analytics and reasoning stemming from (near) real-time information, the analysis of such large amounts

of data is a complex task that requires the development and application of appropriate technologies and tools for processing them—so-called big data technologies.

The big data technologies assume that a vast amount of data is acquired from various sources and in different formats, which is further processed, fused, and analyzed in (near) real-time. Moreover, in the case of anomaly detection, a relatively long period needs to be analyzed in order to identify the standard behavior of vessels and to find patterns such as the main routes that are followed by most ships or by ships of a given type. Also, when applied for security and safety purposes, anomaly detection needs to be performed online; it is crucial to reduce delays between the anomalous event and its detection. In Chapter 9, two examples of the application of big data technologies for the purpose of analyzing maritime data are presented.

4.5. Maritime data sources—a summary

As presented in this chapter, there are different kinds of data sources in the maritime domain that provide heterogeneous data regarding maritime activities. The main drawback is that many of these datasets are not publicly and freely available (they are accessible only to the authorized entities or on a commercial basis). Some of them also require special skills, algorithms, and applications to analyze them (e.g., SAR images). Despite this, there are still a number of open datasets that have a great potential and may be used by various maritime entities and in research.

For the purpose of the methods presented further in this book, the authors have analyzed various maritime data sources and selected those which were available for the study:

- AIS data on a global scale, covering different time periods, depending on the method;
- data about ships and their characteristics, acquired from the following services:
  - MarineTraffic,[27]
  - ITU MARS database,[28]
  - Q88,[29]
  - American Bureau of Shipping;[30]

27. https://www.marinetraffic.com
28. http://www.itu.int/en/ITU-R/terrestrial/mars/Pages/MARS.aspx
29. https://www.q88.com/Q88Search.aspx?c=1
30. http://www.eagle.org/

- data about detentions and inspections of ships published by the following groups:
  - Tokyo MoU,[31]
  - Indian Ocean MoU,[32]
  - Mediterranean MoU,[33]
  - Black Sea MoU,[34]
  - Paris MoU,[35]
  - US Coast Guard,[36]
  - Canada Government;[37]
- data about the classification of ships and their membership in classification societies, published by IACS;[38]
- data about risk indexes:
  - inform Index,[39]
  - basel AML Index,[40]
  - world Risk Index;[41]
- data about ship's accidents and reported piracy and terrorist attacks from the GISIS database;[42]
- selected services of the Copernicus Marine Environment Monitoring Serivce (CMEMS);[43]
- GIS data—political national borders and Exclusive Economic Zones.

Information from maritime experts. An additional source of information for the research was the results of a survey conducted among subject matter experts. This survey was conducted within the SIMMO project (Abramowicz, Filipiak, Małyszko, Stróżyna, & Węcel, 2016), which is elaborated in more details in the next section. The survey took a form of a mail survey with a questionnaire consisting of open-ended questions aimed at identifying the actual system and data sources used in the maritime domain, explaining shortages of the existing solutions, and

31. http://www.tokyo-mou.org/
32. http://www.iomou.org/
33. http://www.medmou.org/
34. http://www.bsmou.org/
35. https://www.parismou.org/
36. http://cgmix.uscg.mil/PSIX/PSIXSearch.aspx
37. http://wwwapps.tc.gc.ca/Saf-Sec-Sur/4/ PSCQ-SRPSC/eng/detentions
38. http://www.iacs.org.uk/shipdata
39. http://www.inform-index.org/
40. http://index.baselgovernance.org
41. http://collections.unu.edu/view/UNU:5763
42. https://gisis.imo.org
43. http://www.copernicus.eu

finding out how the process of detecting anomalies is performed. The questionnaire was sent to five marine experts from the Polish Naval Academy, out of which three responded. Based on answers the provided, it was possible to elaborate and consolidate various sets of information regarding maritime traffic monitoring activities and anomaly detection practices.

The information provided by the maritime experts, along with the data presented in the previous paragraph, was used in the process of designing developing, and evaluating the methods presented further in this book.

4.6. System for maritime monitoring—a case study

4.6.1. Outline of the system

The challenges indicated in the previous sections, such as the enhancement of AIS data, and the retrieval and fusion of maritime data from various, heterogeneous sources, were of interest to a research project called System for Intelligent Maritime MOnitoring (SIMMO). In this section, the set of approaches, methods, and tools developed in the SIMMO project are presented, serving as a case study for the topics described so far.

The general aim of the SIMMO project was to develop a system that allows the situation at sea to be monitored and analyzed in (near) real-time. The system was designed to support a variety of entities working in the maritime domain in analyzing large amount of data about ships in order to ensure the security and safety of maritime traffic. The research carried out in the project was designed to develop methods to enrich data from the AIS with information from heterogeneous internet sources and to automatically detect anomalies related to ships based on this data.

The SIMMO system uses two kinds of sources:

- satellite and terrestrial AIS, which provides information about the location of ships and generic static information about them;
- multiple open internet sources, which provide ancillary information about ships as well as other maritime-related data (ship owners, classification societies, ports, etc.).

Data acquired from both data sources is automatically integrated and fused by the system. Then, the intelligence analysis process on top of the fused data takes place in order to detect suspicious ships with regard to specified threat types. Finally, the enhanced information about ships and any anomalies detected are visualized in order to allow a user to analyze the current situation in the area

under observation, track selected ships, and take additional action if needed. The concept of the SIMMO system is diagrammed in Figure 4.8.

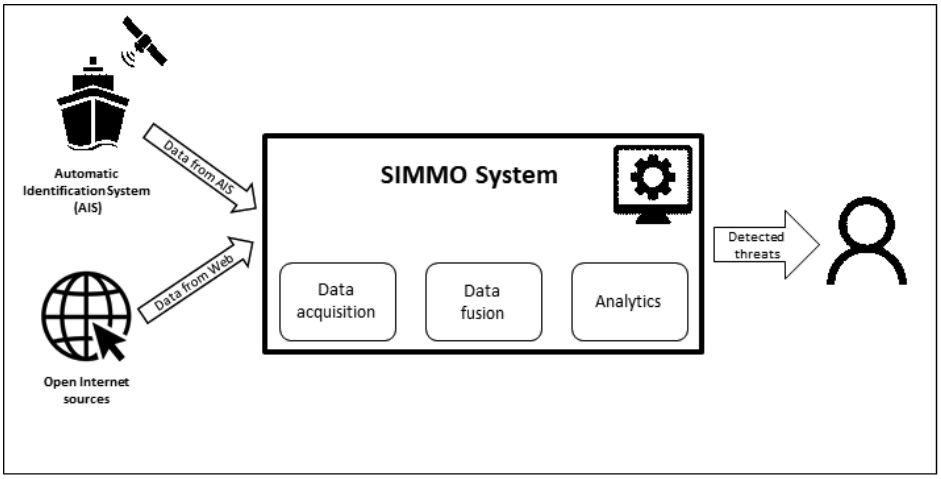


Figure 4.8. The SIMMO concept

Source: Own work.

The SIMMO system is meant to improve the process of analyzing the maritime situation by providing high-quality information about vessels and to automatically detect potential threats (suspicious vessels) with regard to defined criteria.

The SIMMO system offers the following advantages in comparison to the existing solutions:

- improved AIS data quality and content by supplementing the missing part of static AIS messages with information acquired from Internet sources,
- a wider scope of information about vessels, acquired from Internet sources,
- data from heterogeneous sources automatically fused into a consistent dataset,
- up-to-date information about maritime traffic as well as historical routes of ships,
- the possibility to track and analyze ships' movements worldwide, even on the high seas,
- user support in analyzing the current situation by automatically detecting and indicating potentially suspicious ships,
- history of ship anomalies detected by the system.

In the course of the project, a number of methods were developed concerning the retrieval and integration of data from heterogeneous sources (AIS, the shallow and deep web, data encoded in binary file formats, etc.). With regard to data acquisition and fusion, state-of-the-art approaches and techniques were designed

and implemented, encompassing methods for monitoring data streams, extracting information from both structured and unstructured information sources, and methods for merging and integrating data from multiple sources, among other things. As a result, the SIMMO system handles high-volume streams of heterogeneous data and improves the quality (when compared to data coming from only a single data source). Such enriched data collection is subject to further processing, the goal of which is to provide users with novel tools for situation analysis and decision-making.

The methods and technologies related to the data selection and retrieval in the SIMMO system are presented in the following sections, starting from the definition of user requirements, through the methodology for selecting sources, to methods for data retrieval, fusion, and disambiguation.

The results of the processing performed in the SIMMO system are presented in a ready-to-use web application—SimmoViewer—where users can obtain a near real-time picture of the maritime situation in a given area and can quickly identify dangerous ships.

The SimmoViewer can present the current positions of all monitored vessels based on AIS messages and the data fusion process (Figure 4.9). When a particular

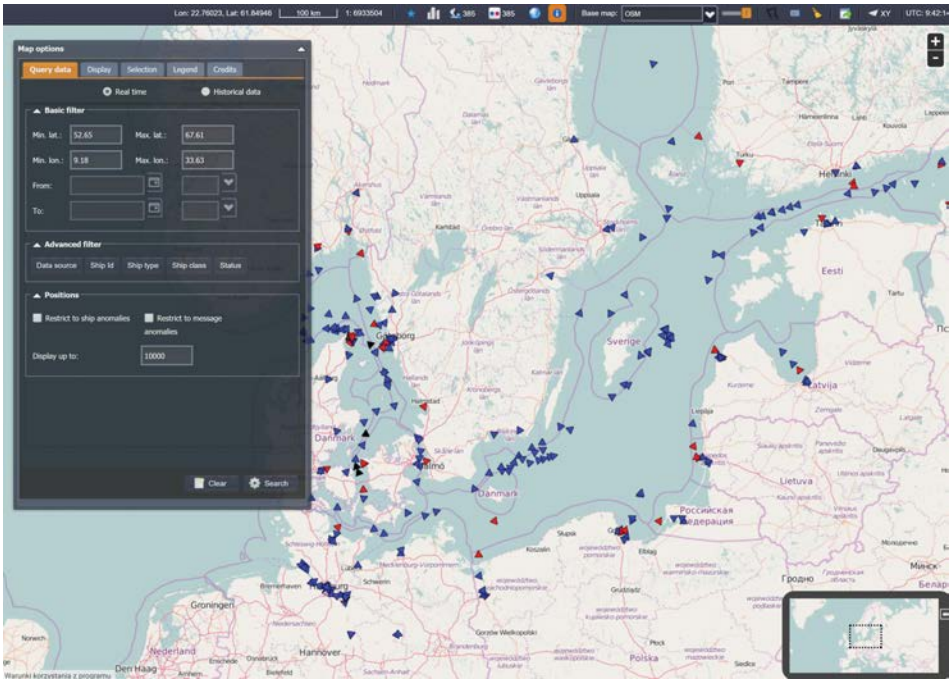


Figure 4.9. Presentation of a current situation at sea in SimmoViewer

Source: The SimmoViewer application developed within the SIMMO project.

position of a ship is clicked on, the system displays basic information about the ship, such as its coordinates, course and speed, destination port, estimated time of arrival, etc. (Figure 4.10)

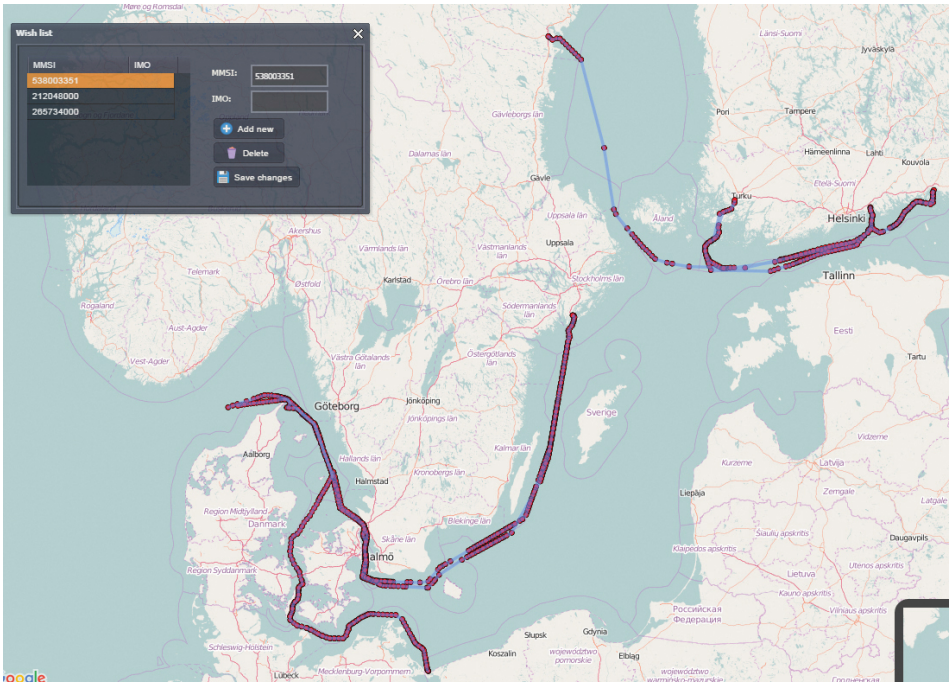


Figure 4.10. Tracking selected ships with SimmoViewer

Source: The SimmoViewer application developed within the SIMMO project.

Apart from the AIS data, the system also presents additional information about a given ship. It contains information gained from ancillary data sources as well as the results of anomaly detection. Basic vessel information, such as the flag, call sign, dimensions, classification society, vessel type, home port, and other information are provided (Figure 4.11).

Moreover, the system contains historical information about each ship. For example, there is a full record of flags, call signs, names and IMO and MMSI numbers used by a selected vessel. Classification surveys, delivery statuses, and inspections and detentions are listed as well. Finally, historical port calls are also accessible.

SimmoViewer is also an interface for analysis and reasoning methods. The detected anomalies related to a given vessel are displayed (Figure 4.12). It is also possible to narrow down the search results to display only vessels with such warning marks (detected anomalies) (Figure 4.13). Moreover, SimmoViewer offers

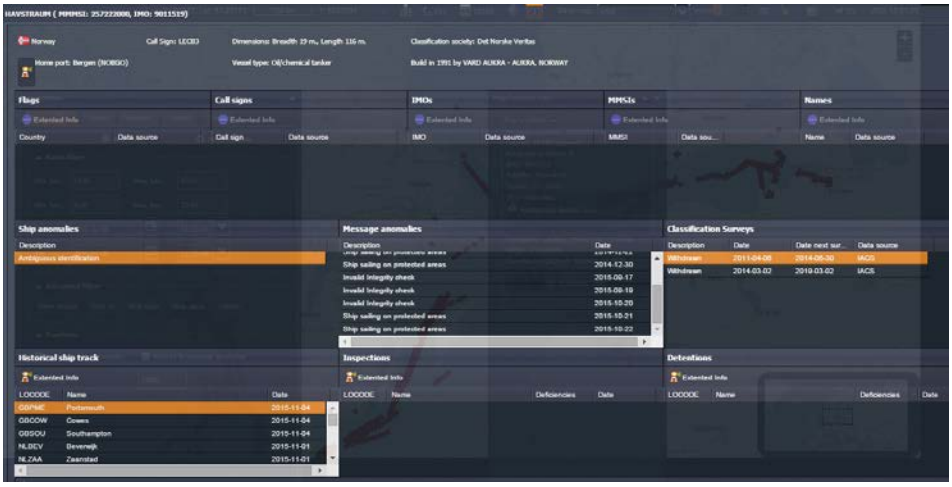


Figure 4.11. SimmoViewer: Detailed vessel information in extended information view

Source: The SimmoViewer application developed within the SIMMO project.

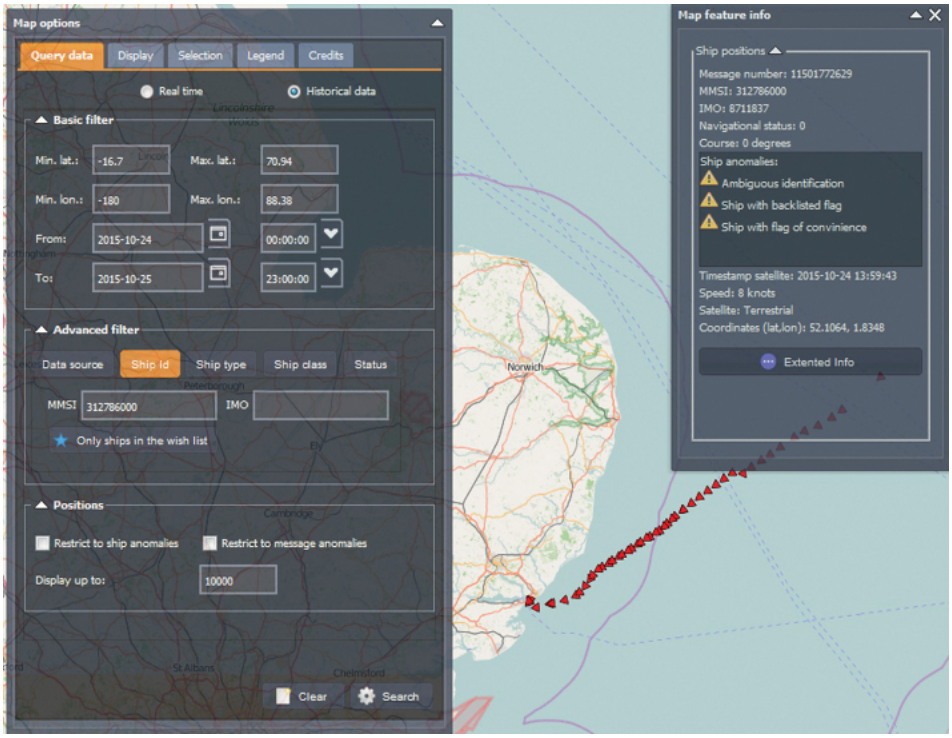


Figure 4.12. SimmoViewer: Detected anomalies warnings

Source: The SimmoViewer application developed within the SIMMO project.

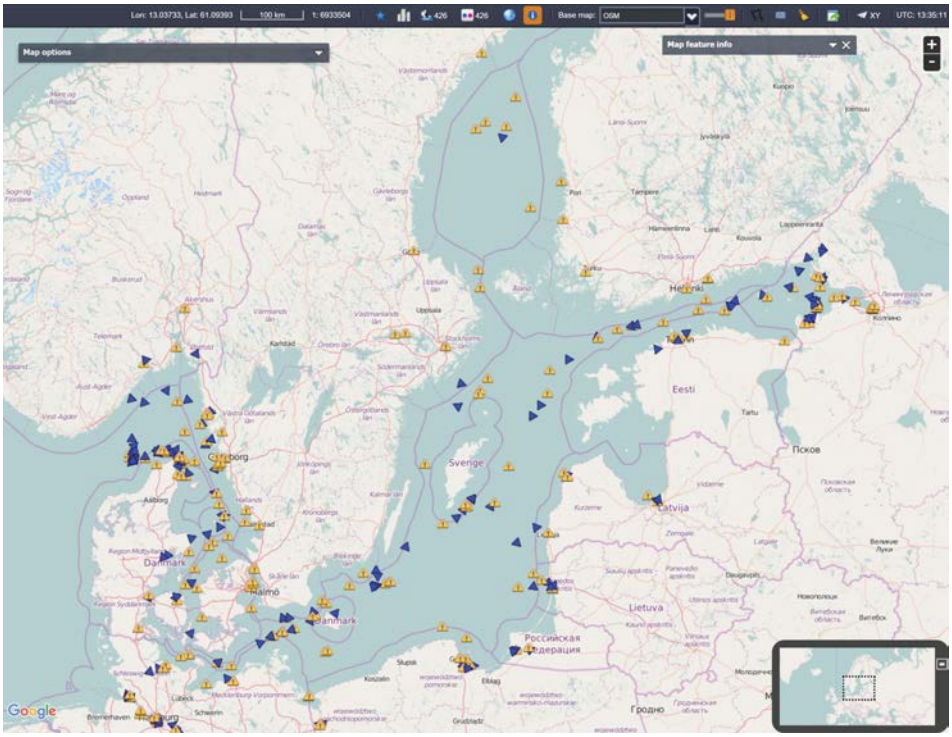


Figure 4.13. History of anomalies in SimmoViewer

Source: The SimmoViewer application developed within the SIMMO project.

a unique anomaly ranking method. It allows the user to set weights to a particular type of an anomaly. The customized weights are then used to calculate a threat score for each vessel.

4.6.2. Maritime data selection

In order to select appropriate online sources for the SIMMO system, the framework for the quality assessment of potential Internet sources, presented in Section 4.3.1 was used. The result of this process is described in the following paragraphs.

Identification of Internet data sources

As indicated in Section 4.3, the first step of the framework is to identify potential sources. In the SIMMO case, potential data sources related to maritime surveillance were identified using search engines, a literature review and consultations with

experts on the subject. The search engines encompassed conventional search engines (like Google) as well as meta search engines like Dogpile,[44] and Webcrawler.[45] Apart from the search engines, other data sources were also analyzed, including sources indicated in (Kazemi et al., 2013) and those suggested by maritime practitioners.

As a result, over 60 different data sources available on the Internet were found. These sources were part of both the shallow web and the deep web. They provided information in a structured, semi-structured, or unstructured manner. The list of identified Internet data sources is presented in Table 4.6. From the point of view of data access, we divided them into four categories:

(1) open data sources (O): websites that are freely available to Internet users,
(2) open data sources with registration (OR): websites that provide information only to authorized users,
(3) data sources with partially paid access (PPA): websites that provide a wider scope of information after a fee is paid,
(4) commercial (paid) data sources (PA): websites with only paid access to the data (a fee or subscription is required).

From all the sources identified, we selected for further analysis only the open data sources (categories O and OR). At this stage, we eliminated the commercial data sources and websites with paid access (category PPA and PA). The elimination of these sources resulted from the fact that they provide only very general, marketing information about the data they have and that access to the data is available only after a fee is paid or a contract is signed. Moreover, our attempt to conclude a contact with these data providers in order to access a sample data failed (requests for data access were sent, but no response was received). Therefore, we were unable to verify the data model or scope of data provided by these sources. Furthermore, in the project we did not foresee paying for an access to maritime data. Eventually, only sources with public content were selected for the project. Nevertheless, we believe it is sufficient to meet the users' requirements and provides advantages of open data.

Similarly, two other data sources (IALA and SafeSeaNet) were rejected due to the fact that access to the data required a long procedure with no guarantee that access would be granted. Due to the project's limited duration, there was not enough time to apply for the data. However, in case these sources are accessed, they can still be assessed according to the framework and added to the system in the future.

As a result of initial selection, 43 sources were taken into account as potential sources for the SIMMO system; these sources were assessed by the experts.

44. http://www.dogpile.com/
45. http://www.webcrawler.com/

Table 4.6. List of assessed Internet data sources

Type of information	Source Name	Type	A	C	R	TP	CC	AR	Quality Grade	Selected
General vessel data	**Marine Traffic**	PPA	H	M	H	H	H	H	98%	Yes
	US Coast Guard	O	H	H	H	H	M	H	98%	Yes
	Maritime mobile Access and Retrieval System (ITU MARS)	O	H	H	H	H	M	M	92%	Yes
	Maritime-connector	O	H	H	H	N/A	H	H	90%	Yes
	ShipFinder	O	H	M	M	H	M	L	73%	No
	AIS HUB	O	H	M	M	M	M	L	70%	No
	Equasis	OR	N/A	H	H	H	M	H	68%	No
	IMO GISIS	OR	L	H	H	N/A	M	H	68%	No
	Vessel finder	O	N/A	L	H	H	H	H	67%	No
	FleetMon	OR	N/A	M	H	H	M	H	67%	No
	Lloyd's Register Ship in Class	OR	M	H	M	L	M	M	65%	No
	ShipSpotting	OR	N/A	H	H	H	H	L	57%	No
	World Shipping Register	PPA	N/A	M	M	M	H	H	55%	No
	IHS	PA	—	—	—	—	—	—	—	No
	Clarkson	PA	—	—	—	—	—	—	—	No
	Internet Ships Register	PA	—	—	—	—	—	—	—	No
	Grosstonage	PA	—	—	—	—	—	—	—	No
	Lloyd's List Intelligence	PA	—	—	—	—	—	—	—	No
	Vessel Tracker	PA	—	—	—	—	—	—	—	No
	International Association of Lighthouse Authorities (IALA)	PPA	—	—	—	—	—	—	—	No
	SafeSeaNet Vessel Traffic Monitoring and Information System	OR	—	—	—	—	—	—	—	No

Type of information	Source Name	Type	A	C	R	TP	CC	AR	Quality Grade	Selected
Ship owners	InfoMare	O	H	L	M	N/A	M	N/A	55%	No
	Seaagent	PPA	N/A	L	H	N/A	L	N/A	33%	No
Weather	ICM Meteo	O	L	H	L	M	L	L	40%	No
	Meteooffice	O	L	H	L	M	L	L	40%	No
	Sailwx	O	N/A	L	L	M	L	M	33%	No
Classification of ships	**International Association of Classification Societies (IACS) Vessel in class**	O	H	H	H	H	H	H	100%	Yes
	American Bureau of Shipping (ABS)	O	H	H	H	M	M	M	88%	Yes
	International Association of Classification Societies (IACS) Transfer of Class	O	H	H	M	H	M	M	82%	No
	ClassNK	O	N/A	H	H	M	M	M	58%	No
	Leonardo Info	OR	N/A	H	H	N/A	M	H	58%	No
	Bureau Veritas Group	PA	—	—	—	—	—	—	—	No
	China Classification Societies	PA	—	—	—	—	—	—	—	No
	International Register of Shipping	PA	—	—	—	—	—	—	—	No

Type of information	Source Name	Type	A	C	R	TP	CC	AR	Quality Grade	Selected
PSC / Banning / Detentions	**Thetis Company Performance**	O	H	H	H	H	L	H	97%	Yes
	Tokyo Mou	O	H	H	H	H	L	H	97%	Yes
	Mediterranean MoU	O	H	H	H	H	L	H	97%	Yes
	Black Sea MoU	O	H	H	H	H	L	H	97%	Yes
	Government of Canada—Port State Control	O	H	H	M	H	H	H	90%	Yes
	Indian Ocean MoU	O	M	H	H	H	L	H	87%	Yes
	Riyadh MoU	O	M	H	H	H	L	M	80%	No
	Latin America Mou	O	M	M	H	H	L	M	78%	No
	Paris MoU	O	N/A	H	H	H	L	H	67%	No
	Abuja MoU	O	N/A	H	H	M	L	H	63%	No
	Caribbean MoU	O	N/A	H	H	N/A	L	H	57%	No
Maritime crimes	ICC Commercial Crime Services	PPA	M	H	L	H	L	M	60%	No
	Maritime Safety Information	O	L	H	L	M	L	H	53%	No
Tankers	**Q88.com**	PPA	H	H	H	M	M	M	88%	Yes
	Auke Visser's International Supertankers	O	L	M	H	L	M	M	63%	No
	International Association of Independent Tanker Owners	PA	—	—	—	—	—	—	—	No
Container ships	Containership-info	O	H	M	L	M	M	H	73%	No

Type of information	Source Name	Type	A	C	R	TP	CC	AR	Quality Grade	Selected
Fishing vessels	Commission for the Conservation of Antarctic Marine Leaving Resources (CCAMLR)	O	H	H	L	H	H	M	73%	No
	International Convention for the Conservation of Atlantic Tunas (ICCAT)	O	H	H	L	N/A	H	H	70%	No
	Indian Ocean Tuna Commission (IOTC)	O	H	H	L	N/A	H	H	70%	No
	Western & Central Facific Fisheries Commission (WCPFC)	O	H	H	L	N/A	H	M	63%	No
	FAO Vessel Record Management Framework (FVRMF)	O	M	H	L	L	M	H	62%	No
LNG vessels	Zeus Intelligent	PA	—	—	—	—	—	—	—	No
	LNG World	PA	—	—	—	—	—	—	—	No
Oil platforms	Oil and gas: offshore maps in UK	O	H	H	L	H	H	H	80%	No

Legend: A: Accessibility; C: Clarity; R: Relevance; TP: Timeliness & Punctuality; CC: Coherence & Comparability; AR: Accuracy & Reliability; H: High; M: Medium; L: Low; N/A: Not available; O: Open; OR: Open with registration; PPA: Partially paid access; P: Paid access

Source: (Stróżyna et al., 2018)

Assessment of Internet data sources

In order to select sources of the highest quality which are best suited to the users' requirements, the data sources were assessed using the six quality criteria presented in the previous section. Definitions of these criteria were adjusted to the needs of the SIMMO project (see Table 4.7).

Table 4.7. Quality measures used to assess Internet data sources

Name	Description	Weight
Accessibility	A possibility to retrieve data from a source; website structure and stability	0.30
Relevance	How well the data are fitted to the SIMMO system's requirements	0.30
Accuracy & Reliability	Data scope, Missing elements, Ship coverage	0.20
Clarity	Explanation of source's metadata model, Data provider	0.10
Timeliness & Punctuality	Data update, Time delay in publishing the data	0.05
Coherence & Comparability	Definition of a described phenomenon and units of measure	0.05

Source: (Stróżyna et al., 2018).

The process of assessing the data sources was conducted using the Delphi method. In fact, Delphi was utilized three times: for assigning weights, assessing sources, and specifying thresholds.

Each potential data source was assessed by assigning a mark to each quality criterion using a four-level rating scale: high, medium, low, and N/A. The rating N/A (not available) means that information required for a particular criterion (e.g., update interval or data coverage) was not specified by a source and, consequently, it was not possible to assess the source in this matter. In case of the measure *Accessibility*, a rating of N/A means that due to the terms of use or privacy policy, it is prohibited to automatically retrieve or use data published by a given source.[46]

The results were then summarized and the final marks in each criterion were determined (by majority rule). The results of quality assessment for each source are presented in Table 4.6.

46. In such sources the following provisions are included: "no part of the information contained in the website may be stored in a retrieval system" or "the use of web-robot or similar techniques to download data in an automated or regular manner is strictly prohibited."

Final selection of sources

After assessment, the final selection of sources took place. Firstly, all sources with an *Accessibility* measure of *N/A* were sorted out. This elimination resulted from the reasons indicated before, regarding access to the data and the data provider's prohibition of using information from these sources. Also, the sources with *Accessibility* assessed as *Low* were eliminated. This encompasses the sources with unstructured information (e.g., text written in a natural language). We excluded them because it was decided while defining the requirements for the system to include only sources with structured or semi-structured information. This was due to the limited time frame of the project and the fact that an automatic retrieval of unstructured information would require a significant amount of work on developing methods for natural language processing.

The sources with the measure *Relevance* graded as *Low* were eliminated as well. It is pointless to retrieve data that are not well-suited to the requirements of the SIMMO system. For example, the SIMMO system focuses only on collecting and analyzing data about merchant vessels; therefore, some categories of sources may have been excluded (e.g., fishing vessels, or oil platforms).

In the next step, each quality measure was converted into numerical value: High = grade 3, Medium = grade 2, Low = grade 1, N/A = grade 0. Then, a final quality grade was calculated according to the formula:

$$X_s = \sum_{i=1}^{n} \frac{\frac{x_i}{3} w_i}{\sum_{j=1}^{n} w_j} \times 100\%,$$

where s is the number of the analyzed sources, $n = 6$, x_i is the grade assigned by the experts to a given quality measure i, and w_i is the measure's weight. The grade was also normalized to the range of 0%–100% (therefore, each assigned measure is divided by 3).

Based on the quality grades that were calculated, a ranking of sources was created. Then, the experts were asked to decide on the threshold for final selection of sources. After two rounds of the Delphi method, the threshold was set to 85%. From the ranking list, only sources with the final grade above this threshold were selected for usage in the SIMMO system (listed in bold in Table 4.6).

To sum up, the application of the proposed framework for selecting data sources in the SIMMO use case allowed us to identify, assess, and finally choose open Internet data sources of the highest quality, which were then used by the SIMMO system.

Model of cooperation with data owners

In the next step, a model of cooperation with external data providers was defined. By external data providers, we mean the sources selected for the SIMMO system. For each selected source, a separate cooperation model was designed and described in the documentation. In defining the model, the following aspects were taken into account:

- the scope of available information—what kind of information is available in a source;
- the scope of information retrieved—what information will be retrieved from the source;
- the type of source—whether the retrieved content is published in the shallow web or deep web and in what form it is available (e.g., internal database, separate XLS, PDF, or CSV files);
- the update frequency—how often the information in a source is updated and whether the whole content is updated or only new information;
- the politeness policy—what kind of robot exclusion protocol was defined by the website administrators, for example, which parts of their servers cannot be accessed by crawlers, as well as requirements on a time delay between consecutive requests sent to the server;
- the re-visit approach—how often the SIMMO system will retrieve information from a given source, i.e., the intervals between consecutive downloads from the source, taking into account the politeness policy, if defined.

4.6.3. Data retrieval and disambiguation

One of the most important goals of the SIMMO project was to develop software modules which would automatically acquire data about vessels or other maritime objects from the selected Internet sources. Special attention was paid to the information about ports, flags, classification societies, and maritime-related companies which will enrich the data available in the AIS messages. The process of enriching the AIS data is essential for two reasons:

- AIS messages may not be complete (i.e., some attribute values may be missing for some reason); acquiring data from the online sources may fill these gaps.
- External internet sources may provide ancillary information, which is not included in AIS messages at all, for example, data about Port State Control, owners of vessels, etc.

Additional data retrieved from external sources is meant to provide end-users with a broader description of the vessels sailing in the area of interest to the user

and to facilitate a detailed analysis of the current maritime situation in that area. Such ancillary data, acquired from the Internet, is displayed in a display module (SimmoViewer) along with the standard data retrieved from the AIS. Moreover, this additional data is used in further analyses, designed to automatically detect vessels which pose a potential threat for some reason (e.g., ships which were recently under detention).

The SIMMO system retrieves data from many sources, and each of these sources may a have different structure and may publish data in a different way. Therefore, a separate Data Acquisition Module (DAM) was developed for each data source (Małyszko et al., 2016). DAMs connect to the data source in a defined manner, send appropriate requests, collect the documents returned, and extract required data. Each DAM is adjusted to the specific structure of the source.

The main steps of the data acquisition process are as follows:

- retrieval of a document that contains data available in a data source;
- extraction of the required data from the document;
- pre-processing of the data to make it more suitable for further analysis; and
- writing the data to the SIMMO database.

In the following subsections a brief description of these steps is provided.

Data retrieval

Data sources may publish data in many different ways. In the sources selected for the SIMMO system, the following basic formats are utilized to publish data:

- webpages—documents in HyperText Markup Language (HTML), which may be interpreted and displayed by web browsers,
- Comma Separated Values (CSV) files,
- Microsoft Excel Spreadsheet (XLS or XLSX) files,
- Portable Document Format (PDF) files,
- weather data in NetCDF files.

Below we describe these categories in detail and discuss how they are processed in the SIMMO system.

Shallow web sources publish data in the form of webpages (HTML documents) which can be directly fetched using GET queries defined according to the HTTP protocol. As a result, a source sends back an HTML document with data embedded in it. Such documents usually contain data concerning a single entity (e.g., a single ship) or a list of links to webpages that contain data about a single entity. The data itself may be extracted from the document using regular or XPath expressions.[47] In order to monitor whether a source has published new information or updated

47. http://www.rfc-base.org/txt/rfc-5261.txt

existing information, a list of URLs of known documents published by this source needs to be maintained. Moreover, a queue defining the order in which these URLs are to be visited needs to be created and managed.

For each shallow web source used in the SIMMO system, a separate DAM was prepared, responsible for the actual retrieval and processing of data. The DAMs share some common operations, such as queuing mechanisms, retrieval of HTML documents under a given URL, and writing data to the database. Still, operations such as data extraction from the HTML document are implemented separately for each source. This is a consequence of the different structures of the HTML documents.

Deep web and AJAX data sources also publish data in the form of HTML documents, but these documents are not directly accessible through a static URL link. Instead, they are dynamically generated in response to queries submitted through the query interface to an underlying database. In this case, in order to fetch the data, DAMs need to perform additional operations, such as posting a form or executing JavaScript code embedded in the HTML document.

This functionality is implemented with the Selenium WebDriver[48] toolkit and the web browser Mozilla Firefox. The toolkit allows for the automation of actions within a web browser. Then, it is possible to automatically submit instructions to one of the supported web browsers. In the case of SIMMO, the developed DAM opens a Mozilla Firefox browser window inside X virtual framebuffer (XVBF).[49] The process of data acquisition from these sources is presented in Figure 4.14.

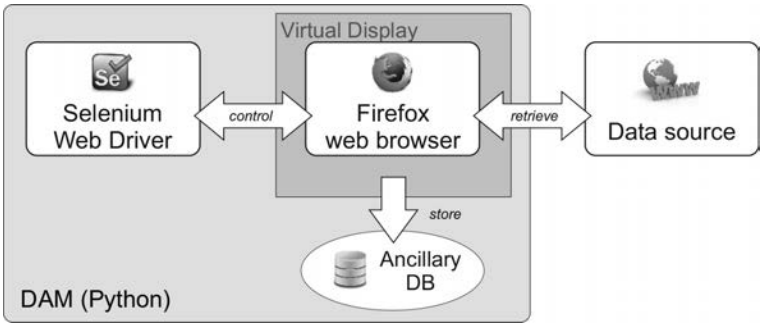


Figure 4.14. Pipeline of data acquisition from AJAX and Deep Web data sources

Source: (Stróżyna et al., 2018).

The third category is **data sources with CSV and XLS files**. In the sources used in the SIMMO project, CSV and XLS(X) files with the required data are published

48. http://www.seleniumhq.org/

49. TXVBF is software which allows applications with a graphical user interface to be run on computers without display hardware or physical input devices; see http://www.x.org/archive/X11R7.6/doc/man/man1/Xvfb.1.xhtml

on a regular basis under a certain URL (e.g., once a week). Therefore, it is relatively simple to retrieve the document, for example, by periodically running software (a crawler) that determines whether a new file has been published, and, if so, downloads it. Once the file is downloaded (and unpacked, if necessary), it can be read and its content can be processed sequentially, row by row, to get data about entities.

Data sources with PDF files are websites from which data can be accessed by downloading PDF files accessible under a certain URL. While the PDF format has some advantages,[50] it is difficult for automatic processing because PDFs are files designed to be read by humans. The processing of PDF documents becomes even more difficult if we want to automatically extract data from a table that is embedded in a PDF. Nevertheless, it is still possible. The processing pipeline for fetching and processing PDF files developed in the SIMMO system is presented in Figure 4.15.

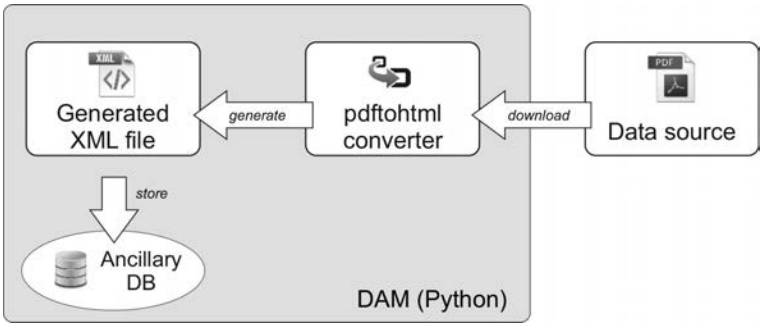


Figure 4.15. Retrieval of data from sources, which publish data in a form of PDF files

Source: (Stróżyna et al., 2018).

First, a PDF file is downloaded from a source to a local disk. Next, the file is converted to XML using the program pdftohtml,[51] which is included in the Ubuntu Linux operating system. This program produces an XML document that contains a text which is suitable for further processing.

The final category of data is **weather data** to be gathered from selected products of the Copernicus Marine Environment Monitoring Service (CMEMS) (Copernicus Marine Environment Monitoring Service (CMEMS), 2019). In this case, NetCDF files containing weather data for various maritime areas are regularly made available on the CMEMS FTP. In order to download weather data for further processing, it is necessary to write a piece of code for continuous monitoring of the Copernicus server folders and to download data as soon as new information is identified.

50. Files in this format are highly portable and easy to open for displaying in different applications.
51. http://linux.die.net/man/1/pdftohtml

Such software is usually called a *bot*. In our research, an appropriate bot was implemented to download weather data from the CMEMS FTP and to serialize desirable variables to the Apache Avro format.[52] It is basically a Python script that runs periodically to check whether new data are available and to download the most recent .nc files with weather data. Two steps are performed to avoid duplicate downloads:

(1) The bot traverses the FTP directory tree, stores FTP paths to .nc files in a local database (SQLlite), and flags them as `synced`.
(2) The bot downloads the next file from the `not_synced` queue and flags it as `synced` after a successful download.

After being successfully downloaded, the .nc files are serialized to Avro by special serializers—separate ones for each CMEMS product. Since only selected weather variables were relevant for our research, prior data filtering had to take place before data storage. The filtering is done by serializers, which take into account the data model foreseen for an Avro file where selected weather variables are defined.

The collected weather data is subject-oriented, that is, it groups data around the same phenomenon. For further calculations, we need spatio-temporal data all interesting variables need to be combined into a single row with fixed coordinates, collected for the same time. Thus, the downloaded data are available in two spatial resolutions: 0.5×0.5 degrees and 0.25×0.25 degrees.

Data disambiguation and fusion

The data retrieved from Internet sources concern different types of entities:

- vessels,
- ports that, depending on the context, may concern the current destination of a vessel, the home port, or a location where the vessel's inspections take place, etc.,
- flags that correspond to the country of registration of a vessel,
- classification societies,
- maritime companies (e.g., the vessel's owner or manager).

However, data about a single entity retrieved from different sources may vary from each other—in each data source the same entity may be referenced differently. For example, different words (names) may be used to call the same entity (e.g., a port or a ship). In other words, for each attribute of a certain entity there may be different values in data sources. Therefore, before the data are added to the database, such differences must be recognized and the data need to be aligned. The

52. https://avro.apache.org

pre-processing of the retrieved data includes data disambiguation—identification which entities the data concerns and data fusion, merging different data about the same entities into a single record.

Further on, we present how these steps are performed in the SIMMO system in the case of vessels.

We understand vessel data disambiguation as a process of assigning the same identifier to each data record concerning the same ship. The identifier (shipId) should be unique for a given vessel and all data concerning this vessel should have the same identifier assigned. Figure 4.16 presents a flowchart depicting how the disambiguation of vessel data may be performed.

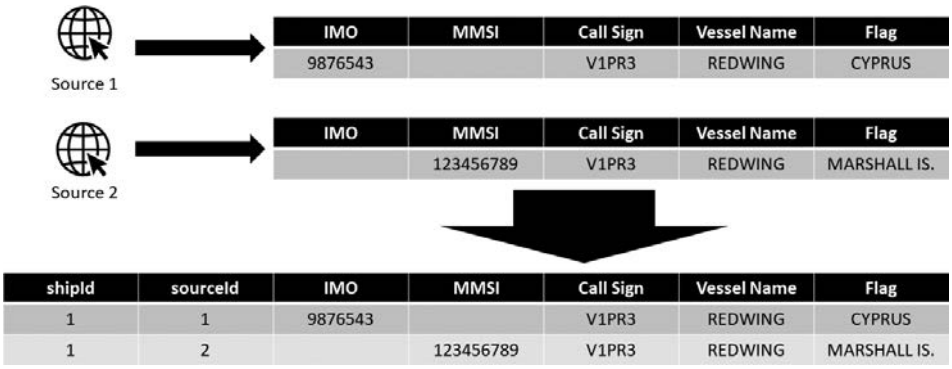


Figure 4.16. A simple schema presenting the goal of the vessel data fusion

Source: Own work.

Let's assume that we have two records with selected data about static vessel features from two different sources, for example, MarineTraffic and Maritime Connector. In the situation shown in the Figure 4.16, we may notice that the vessel's name and call sign in the two records are the same. This may indicate that both records concern the same vessel. In such case, both records should have the same identifier assigned (value in the shipId column). This identifier should also be assigned to any other data that concern this vessel.

The data disambiguation rules defining which attributes should be taken into account and in which manner were based on the statistical corpus analysis to ensure that the results of the disambiguation are correct.

The goal of the second step of data pre-processing—data fusion—is to create a single record consisting of data that relate to a particular vessel. Values for this record are to be selected from all data retrieved from all monitored sources. For example, in Figure 4.17 sample data are presented, where for a vessel with shipId $= 1$ there are three different records available from three different data sources. According to Source 1, the flag for this ship is the Marshall Islands, but

shipId	sourceId	IMO	MMSI	Call Sign	Vessel Name	Flag
1	1	9876543		V1PR3	REDWING	CYPRUS
1	2		123456789	V1PR3	REDWING	MARSHALL IS.
1	3	9876543			REDWING	MARSHALL IS.

shipId	IMO	MMSI	Call Sign	Vessel Name	Flag
1	9876543	123456789	V1PR3	REDWING	MARSHALL IS.

Figure 4.17. A simple schema of a vessel data fusion

Source: Own work.

according to Sources 2 and 3 it is Tuvalu. The goal of data fusion is to select one of these values (Marshall Islands or Tuvalu) to be the primary value for this attribute. The record with fused data consists of a set of such primary values for each vessel's attribute.

The data fusion may be performed based on different rules, for example:

- by selecting the most common value, or the value that occurs in the data sources most often. It may be assumed that the value is correct because many or most of the sources report exactly the same value (an *Argumentum ad populum* inference);
- by assigning different priorities to different data sources and selecting the value from the source with the highest priority. The priority should reflect how reliable the source is according to the quality assessment;
- by analyzing agreement between different attributes.

Similar pre-processing of data may also be performed for other entities, such as ports, flags, companies, classification societies, and vessel types. For example, in the case of ports, their names may differ between sources (due to abbreviations or different language versions, like *St. Petersburg* and *Saint Petersburgh*, or *Gdańsk* and *Danzig*). Matching various names of a port requires the use of different lexical resources and text processing methods. Such methods were developed within the SIMMO project in order to address this challenge. They are presented in detail in another paper (Małyszko et al., 2016).

Another important aspect is that data describing a certain entity may change over time. These modifications reflect changes which occur in relation to a particular entity. In the SIMMO system, it is important to store not only current data, but also the historical data, as they still may be useful in detecting anomalies. Thus, the system enables both current data and historical values of the selected attributes to be stored.

Data analysis

One of the goals of the SIMMO system is to detect anomalies in ships' behavior. Therefore, in the course of the research, a set of methods for detecting anomalies was developed. These methods are based on advanced reasoning and analysis methods for exploring standard patterns in spatio-temporal data and the detection of outliers.

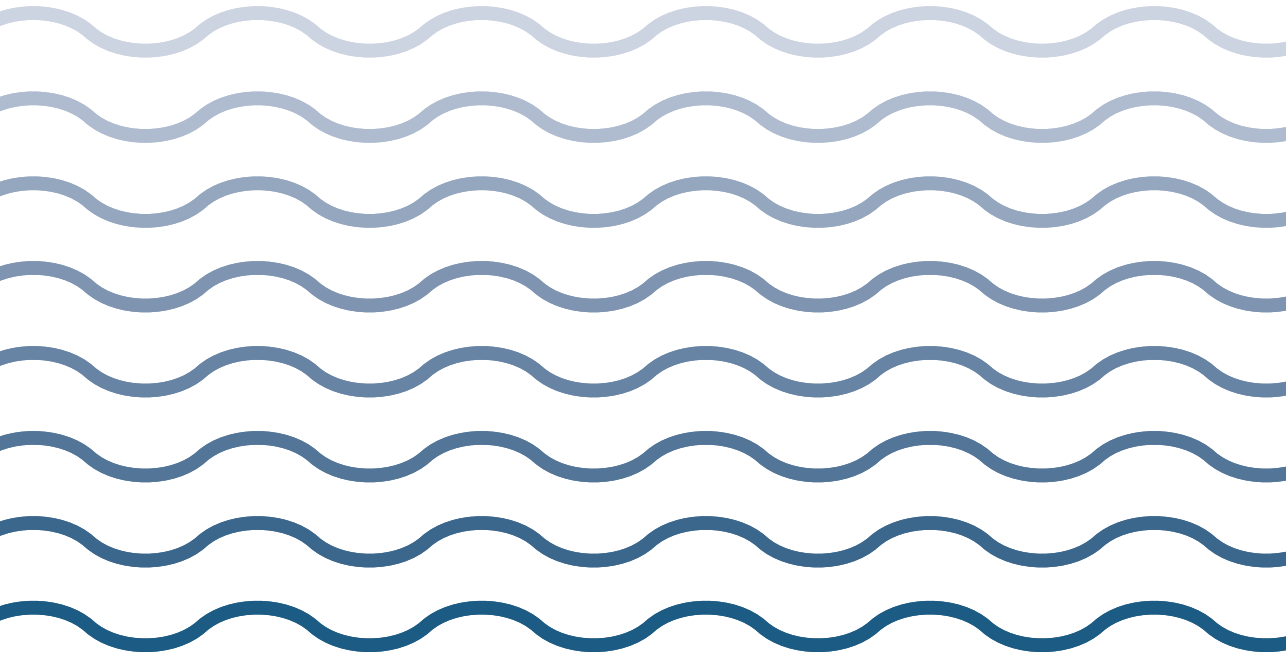
The analysis that the system conducts is both retrospective (based on historical data) and prospective, and relates to the behavior of ships at sea and to changes in their static characteristics. As a result, the system automatically identifies different types of maritime anomalies, thus aiding in the identification of suspicious ships which pose a potential threat.

The SIMMO system is able to detect the following anomalies:

- inconsistent or missing AIS data (i.e., incomplete static and dynamic information),
- ambiguous identification (i.e., duplicate or implausible MMSI or IMO number),
- sudden changes in a ship's identity (i.e., change of name, call sign, type),
- flying a black;osted- or a grey-listed flag,
- suspended/withdrawn classification status,
- registration of a ship's owner/manager as a poor-performing maritime company,
- calling at suspicious ports in the past,
- being listed as a banned or detained ship,
- occurrences of loitering at high sea (i.e., anomalies in a ship's behavior with regard to speed or course).

The selected methods for identifying the above-listed anomalies are described in Chapters 6 and 9.

Chapter 5



5. MARITIME ROUTING AND TRAFFIC NETWORKS

A standard procedure in today's shipping domain is a predictive voyage planning and monitoring. Voyages are planned berth to berth, resulting in a route that consists of waypoints a vessel should follow to assure security and safety. In voyage planning a number of aspects must be considered, such as minimum water depth in fairways and ports, the existing routes, Marine Protected Areas (MPA) and meteorological information like wind or tide. After the voyage started, the crew must monitor the execution of the route continuously in order to detect any deviations as soon as possible. There is also the possibility that the crew has to replan the route to adjust to changing meteorological conditions or traffic situations. This procedure of voyage planning and monitoring is regulated by the International Maritime Organization (IMO).

Nowadays such planning is performed manually by a ship captain with the use of a dedicated software. An alternative is an automatic route planning by specialized assistance systems that support the navigator both in planning before the voyage starts and in monitoring it when the ship is underway. Such systems analyse historical movements of ships and apply various methods to extract maritime traffic patterns from this data. Having the knowledge about typical behaviour patterns under different conditions, they can propose a safe route when planning a voyage. At the same time, other useful parameters, such as weather conditions, can be automatically taken into account.

This chapter presents a review of methods and algorithms to extract maritime traffic patterns and find an optimal route for a ship's voyage. Further on, the concept of an assistance system whose goals is to support a navigator both in voyage planning and monitoring is presented. The system was developed within the HANSA project.[1]

5.1. Ships routes prediction

A problem of finding an optimal route for a given ship has been addressed by scholars and practitioners for many years now. An optimal route can be defined as a blend of shortest time, minimal fuel consumption, and general safety of

1. The HANSA project was funded by the MarTERA partners German Federal Ministry of Economic Affairs and Energy (BMWi), Polish National Centre for Research and Development (NCBR) and Research Council of Norway (RCN) and was co-funded by European Union's Horizon 2020 research and innovation program under the framework of ERA-NET co-fund, https://www.emaritime.de/hansa/

navigation (H.-B. Wang, Li, Li, Veremey, & Sotnikova, 2018). *Routing* and *path planning* seem to be used interchangeably in the literature. Following Tu et al. (2017), there are some formal differences between them. Path planning in its simplest form can be defined as finding the shortest path between two points, using the great circle distance or rhumb line and considering the obstacles. Routing can be defined as a prediction of a vessel's next position based on its current position and a number of features, such as speed (Tu et al., 2017). Other scholars refer to it as *route design* (Cai, Wen, & Wu, 2014) or *navigation planning* (Tan, Weng, Zhou, Chua, & Chen, 2018). The term can be narrowed down to some specific meanings. For instance, weather routing adds additional layer of complexity by considering conditions such as wind or sea currents. Other research focuses on fuel efficiency in planning (Schøyen & Bråthen, 2015).

The problem of a ship's routing and prediction of a ship's future position is already a well addressed area in the literature. The existing methods take into account various aspects while planning the route, such as ship characteristics, weather, type of transported cargo, or economic factors. Tu et al. (2017) categorized the methods into three classes: physical, learning, and hybrid models.

The first class models a ship's motion by using mathematical equations and calculates motion characteristics using physical laws. In this group one can indicate curvlinear, lateral, and ship models. The second class models the route by learning motion characteristics from historical data, and thus implicitly integrating all possible influencing factors. These types of methods treat the ship manoeuvring system as a whole system. The hybrid methods build a model that either explicitly considers a certain part of influencing factors and is trained on historical motion data or it combines different learning methods. In our research (presented further in Chapter 9), we consider only the second and the third class since the physical models are mainly used in simulation systems.

The second and third class of methods include, inter alia, the well-established methods of graph or network theory, Neural Networks, or Gaussian Process methods. The graph-based methods assume discretization of location and navigational data, and construction of a graph which takes into account defined key maritime points or free areas of the environment. Then, an algorithm for finding the shortest path in the graph is used, like an A* or Dijkstra algorithm (Azariadis, 2017; Hornauer & Hahn, 2013). Also, a cell decomposition can be used, where a free space is divided into cells and the path is computed between these cells; but this is suited usually for short- or middistance route planning. The main drawback of these methods is that they may not be realistic since they do not take into account the behaviours and habits of ships.

A network design is a subject of research specifically in the case of linear shipping, where the problem of constructing routes and choosing which routes to serve is crucial at the beginning. Liner ships operate mostly along established routes and follow regular timetables that may not change for several years. Moreover, in the

planning phase information about the demand is additionally considered. Here various methods have been developed, like models with a single route, models with sets of routes without transhipment, hub and feeder route models, models that distinguish some ports as hub ports, or multi route models (Christiansen, Fagerholt, Nygreen, & Ronen, 2013).

For trajectory prediction also neural networks (Mazzarella, Arguedas, & Vespe, 2015; Nguyen, Vadaine, Hajduch, Garello, & Fablet, 2021; Singh & Heymann, 2020; Zissis, Xidias, & Lekkas, 2016) and Gaussian methods (Tu et al., 2017) are used. The first group includes methods that differ from each other with respect to the mapping function and network structure. The main advantage of neural network methods is their general good and stable performance and lack of need to provide assumptions or prior information on the ship or the weather. On the other hand, the training process of a neural network is usually very slow and in most cases the network architecture needs to be determined empirically. The Gaussian methods are mainly used for theoretical analyses, which due to their analytic properties, can be readily performed. However, these methods are characterized by high computational requirements and poor scalability, which is important in the case of big data or real time applications.

Trajectory prediction may also be conducted based on the clustering of historical vessel paths. To this end, the most popular clustering algorithm is DBSCAN (Density-Based Spatial Clustering of Applications with Noise) introduced by Ester, Kriegel, Sander, and Xu (1996). In DBSCAN, contrary to many unsupervised learning algorithms, the number of desired clusters is not its hyper-parameter (i.e., the number of clusters does not have to be known upfront). DBSCAN also detects and deals with outliers in an automatic way, which is desirable for (usually) noisy data such as AIS. Ester and Wittmann (1998) later extended this approach and prepared an incremental learning version of the DBSCAN algorithm. DBSCAN has been used in many contexts and variations with AIS data, such as for detecting fishing spots (Mazzarella, Vespe, Damalas, & Osio, 2014), or finding abnormal trajectories in a parallel manner (Z. Chen, Guo, & Liu, 2017). In another example, Pallotta et al. (2013) presented Traffic Route Extraction and Anomaly Detection (TREAD), which is a methodology for incremental and unsupervised machine learning approach for building a maritime traffic network through waypoints discovery, low-likelihood anomaly detection, and route prediction. The waypoints discovery component relies on the incremental version of DBSCAN. TREAD was later used and extended by Arguedas, Pallotta, and Vespe (2017) as Maritime Traffic Knowledge Discovery and Representation System, which aims at traffic network creation. Construction of the network—preceded by waypoints detection, route detection, and route decomposition—relies on the Douglas-Pecker algorithm for breakpoints detection (these will serve as nodes) along with a custom algorithm for creating traffic lanes (edges). Also Karataş, Karagoz, and Ayran (2021) extended

TREAD by adding direction checking through bearing, time interval and course clustering. The clustered AIS messages are then used as an input to a classification algorithm (Random Forest) that provides a different prediction model per cluster.

Another group of methods is evolutionary algorithms. H.-B. Wang et al. (2018) proposed to use a genetic algorithm in the weather routing research. Indeed, swarm and evolutionary algorithms can solve maritime routing and planning related problems. Different swarm intelligence methods have been used in various scenarios. For example, Kosmas and Vlachos (2012) used simulated annealing, whereas Tsou and Cheng (2013) proposed ant colony optimization for this task. A number of studies have demonstrated the usage of evolutionary algorithms in different configurations, such as multi-objective evolutionary algorithm (Marie & Courteille, 2009; Szłapczynska & Smierzchalski, 2009; Vettor & Soares, 2016), or real-coded genetic algorithm (Maki et al., 2011; H.-B. Wang et al., 2018). Dobrkovic, Iacob, and van Hillegersberg (2015, 2018) used genetic algorithm paired with spatial partitioning to enhance the process of clustering vessel positions and enable fast computation of increasing amounts of data. Their research is one of the first that focuses not only on proposing a robust and accurate algorithm but also on the speed of the algorithm, enabling it to be used in real-life applications where large data volumes have to be processed quickly.

There is a certain number of hybrid methods. The isochrone method was proposed by James (1957). Originally, it was not suitable for computers, but the method was extended by Hagiwara and Spaans (1987), as well as by Fang and Lin (2015). Calculus of variations, approaching the issue as a continuous minimum optimization problem, was used by Haltiner, Hamilton, and 'Arnason (1962). It was later extended by Bijlsma (2001). H.-B. Wang et al. (2018) point that this method is not very useful for practical applications. One can also use dynamic programming, which treats the issue as a discrete multi-stage decision problem (Bellman, 1952). It was later used by several scholars for the same problem (Calvert, Deakins, & Motte, 1991; De Wit, 1990; Shao, Zhou, & Thong, 2012). H.-B. Wang et al. (2018) argue that this method is highly complex and accurate.

Another popular approach to calculating a ship's route is the utilization of density maps. A density map depicts the most usual sea routes followed by vessels. Based on it a probable route can be defined. Generation of density maps requires a combination of a big dataset of historical locations with a real digital map and a route-generation algorithm. For example, Azariadis (2017) proposed a method for a calculation of short- to mid-range ship routes based on density maps derived from previous historical locations of liner and merchant ships. The information about density is represented by map pixels with colour values. Having obtained this information, they developed an evolutionary optimization algorithm (a modified A* algorithm) for finding the shortest path in graphs and grid-based algorithms. The resulting route can be then used to calculate travel time and predict an estimated time of arrival. A similar approach for

trajectory planning aimed at collision avoidance was proposed by Hornauer and Hahn (2013).

Constructing vessel motion models from historical data and predicting their future trajectories, like the density-based models, has been gaining in popularity along with the common usage of AIS (Blaich et al., 2015). Still, mainly only short term route estimation is achievable, while, as indicated by Tu et al. (2017), medium-term and longer-term estimation would be more useful (e.g., in the case of restricted manoeuvrability of some types of vessels).

There is also a group of methods that concern weather or economic factors. Bijlsma (2010) and Y.-C. Chang, Tseng, Chen, Chu, and Shen (2013) proposed methods for an optimal route planning that take into account ocean currents. Economic factors focus mainly on fuel consumption. Research says that a 10% decrease in a ship's speed may result in a 19% reduction in engine power and a 27% reduction in energy consumption and thereby a lower CO2 emission. Fuel consumption and a reduction of emissions are nowadays important factors which many ship owners take into account while planning a route. Therefore, methods have emerged supporting such planning. For example, Bijlsma (2008) developed a method for specifying the amount of fuel that can be consumed on a specific transoceanic route by computing a minimal-time route.

5.2. Maritime traffic networks

There exists a variety of approaches for extracting and representing maritime traffic patterns from historical AIS data. Following the classification of (Riveiro, Pallotta, & Vespe, 2018), the existing research on extracting and modelling traffic patterns can be divided into grid-based, vector-based and graph-based approaches.

Grid-based approaches are applying a grid in the considered sea area. To represent traffic patterns, the typical vessel behaviour is modelled in each cell by considering parameters such as speed, course or position. Such a procedure can be found in (Bomberger, Rhodes, Seibert, & Waxman, 2006; Ristic, 2014; Xiao, Ponnambalam, Fu, & Zhang, 2017). Bomberger et al. (2006) model the typical behaviour of ships with information about their positions and speeds within a cell. The authors show how their approach can be used for predicting the cell in which a vessel will be in 15 minutes. Xiao et al. (2017) propose a knowledge-based prediction approach. The traffic patterns needed for this approach are extracted by modifying the DBSCAN algorithm in such a way that the algorithm can be applied to a grid. Subsequently, these grid-based patterns are used to calculate the typical moving patterns of vessels by using Kernel Density Estimation (KDE). A similar approach was presented in (Ristic, 2014). A grid is applied to the considered sea

area and in an unsupervised learning phase maritime traffic patterns are described by extracting speed patterns inside each cell. For this purpose, KDE is applied.

Vector-based approaches assemble tracks or trajectories from associated AIS data points. From these tracks, significant points are extracted, such as points where ships manoeuvre, stop or enter the considered sea area (Riveiro et al., 2018). The approach proposed by Pallotta, Horn, Braca, and Bryan (2014) is based on (Pallotta et al., 2013), which is a method for identifying distinctive events in historical AIS trajectories. The main idea of (Pallotta et al., 2014) is to model the variation in the behaviour of vessels traveling on the same route. For this purpose, the authors approximate the movement of the vessels by modelling it as an Ornstein-Uhlenbeck process. Another vector approach can be found in (de Vries & van Someren, 2013). In their work, the authors propose an extension to the Piecewise Line Segmentation method. This extension enables the compression of historical AIS trajectories without losing relevant information about the trajectory such as stop points or distinctive turning points. Furthermore, the authors show how this representation method can be used for determining the similarity between different trajectories. Also, Rong, Teixeira, and Guedes Soares (2020) proposed a relatively straightforward and unsupervised method to characterize maritime traffic and determine maritime turning sections. In this approach a ship trajectory compression and turning point detection techniques were combined with the density-based clustering method. For the former Douglas and Peucker algorithm is used to compress ships trajectories and detect turning points—a point in a ship trajectory where a significant directional change is observed. Further on, turning points are clustered using the DBSCAN method into turning sections—areas where the turning behaviour is frequently observed.

Graph-based approaches model traffic patterns in a more abstract way than the two previously mentioned methods. All existing approaches share extracting significant points from historical AIS data and representing them in a graph (Riveiro et al., 2018). Such points can be, for example, points at which ships manoeuvre, reduce speed or anchor. Typical traffic patterns can then be described by a sequence of nodes in the graph. Oltmann (2015) introduces the Route Topology Model (RTM) for the North Sea region. The nodes in this approach are geographic points at which vessels usually perform a manoeuvre. The RTM is a topological representation of traffic, which means that only the nodes have a geographical reference. The edges only connect the nodes and thus model the reachability of the nodes by the ships. Hence, this approach does not enable modelling, e.g., the exact course of a waterway or a channel.

In contrast to (Oltmann, 2015), Varlamis, Tserpes, Etemad, Júnior, and Matwin (2019) propose a network representation, which is focused on the different navigational states vessels have during their journey. To create the graph, historical trajectories are analysed taking into account predetermined thresholds for course or speed changes. Points at which the respective threshold values are exceeded

are then clustered using the DBSCAN algorithm. A Markov Model is used to model typical behaviour. For this purpose, the transition probabilities from one node to another are stored.

Arguedas et al. (2017) propose a directed-graph representation for typical vessel traffic patterns. This graph is created by identifying waypoints in historical vessel trajectories. Waypoints are defined in this work as geographic points at which significant course changes can be detected. To extract such points from historical vessel trajectories, the authors apply the Douglas-Peucker algorithm, which is a method for smoothing curves (Douglas & Peucker, 1973). Those points are subsequently clustered with the DBSCAN algorithm. Afterwards, the order and the frequency in which vessels visit the determined waypoints is calculated, which yields a directed graph.

The idea of using waypoints extraction in order to identify traffic patterns is supported also by Dobrkovic et al. (2015, 2018). In the research they show that application of evolutionary algorithms, such as the genetic algorithm (GA) or ant-colony optimization, to discover sequential waypoints can be a viable alternative to other machine learning approaches. The proposed GA is able to provide accurate results once good criteria for GA fitness function have been found. Moreover, they address the problem of varying density traffic and high computational time of GA by using quad tree structures to pre-process the data and isolate the areas of high traffic density. Moreover, this approach handles routes with missing data.

In contrast to the works presented above, Lamm and Hahn (2019) propose to use a ship type specific graph for representing typical vessel movement patterns. The idea is to identify the points in historical vessel tracks at which vessels performed a manoeuvre. Hence, the resulting graph is called a manoeuvre net. A manoeuvre is defined in (Lamm & Hahn, 2019) according to the definition of Fossen (2011), who describes a maneuver as a change in course or speed. To identify a maneuver, Lamm and Hahn (2019) propose to use the cumulated sum (CUSUM) procedure. CUSUM is a method for change point detection in stochastic processes, whose usage for manoeuvre detection was demonstrated in (Lamm & Hahn, 2017). In order to distinguish between soft manoeuvres that are performed to maintain the course and strong maneuvers that are necessary in order to follow the vessel's route, a threshold for course and speed change is defined. The simple moving average (SMA) is then used to determine if the defined thresholds are exceeded. This procedure yields a set of manoeuvre points for each ship type, which are clustered by applying DBSCAN. To create a manoeuvre net the centre of each cluster is extracted by applying the medoid calculation. At the end, the consecutive manoeuvre points can be connected with the edges, which yields a manoeuvre net.

As described above, there exists a variety of approaches for extracting and representing maritime traffic patterns. All the approaches considered share a common basic procedure: at the beginning, historical AIS data is processed. The data

is either processed and analysed in a grid or the historical AIS data is merged with tracks and further processed accordingly. Another option is to extract distinctive points from historical tracks and model them in a graph. It becomes evident that all approaches only use historical AIS data as an information source. Context related information such as weather conditions are not considered in the analysis and representation. Furthermore, some of the chosen approaches are inefficient for large amounts of data, such as grid-based approaches or those using the DBSCAN algorithm. In addition, most approaches are evaluated with a relatively small amount of data covering only a small area.

In the next section, we present the approach for tackling these shortcomings that was applied in the HANSA project. We introduce the concept of Recommended Corridors (RCs) to represent context-sensitive maritime traffic patterns. For this purpose, the existing concepts are extended in such a way that context-sensitive patterns can be extracted and modelled. Furthermore, we describe a more efficient approach for extracting those patterns by using the Lambda-Architecture and in-memory computing technologies (see Chapter 9).

5.3. HANSA system—a case study

5.3.1. Outline of the system

As indicated in the introduction to this chapter, there is an emerging need to develop assistance systems that would support the navigator both in planning the voyage, monitoring the vessel at sea during the voyage, as well as in replanning the route if necessary. Such a system needs to have knowledge of the typical traffic patterns that most often ships follow. Moreover, it should also take into account other useful parameters, such as weather conditions, to propose an optimal route under different operational conditions.

The concept of such an assistance system was developed within the HANSA project (Retrospective Analysis of Historical AIS Data for Navigational Safety through Recommended Routes)[2]. In the project we developed a method for extracting vessel movement patterns from historical maritime data combined with historical weather information that was then applied in the HANSA system. In this section, a general overview of the HANSA system will be described along with a concept of a method for marine traffic network generation and utilization.

Recommended Corridors. In the project we proposed the concept of Recommended Corridors (RC) for modelling the most commonly used traffic patterns between a given origin to a destination. The idea of the RC is that it represents the

2. https://www.emaritime.de/hansa/

area in which vessels usually travel. It is intended to represent previous experience and best practices and is to be provided to onshore as well as offshore personnel for planning and monitoring a vessel's voyage. RCs must also consider the information typically used during voyage planning and monitoring, such as length, draught, and general manoeuvrability, as well as meteorological conditions, and context information, e.g., minimum draught along the potential route. Furthermore, RCs are ship type specific. This enables, among other things, the modelling of the differences in manoeuvrability that exist between the respective ship types. As a result, there might be different RCs for the same route depending on the considered ship type and meteorological conditions.

Since RCs are context-sensitive, they have the ability to dynamically adjust as soon as weather conditions change. This tackles the problem of increased workload for bridge crews as soon as replanning becomes necessary due to changing weather or traffic conditions.

In our approach, we propose to derive RCs from a graph-based traffic pattern representation, which will be referred to as 'mesh' from here on. In general, a mesh is a graph that represents all RCs in a given sea area and is context-sensitive and ship type specific. The proposed solution consists of a combination of a genetic algorithm and a method for generating a graph representing a maritime traffic network.

The visual representation of an RC, e.g., on an Electronic Chart Display and Information Systems (ECDIS), is suited for the visual support of mariners on ship bridges or Vessel Traffic Service (VTS) officers. Besides being a method for checking if a vessel is outside of an RC, it also enables a visual detection of deviations. This can reduce workload in dense sea areas and provides VTS officers another method for detecting anomalous and potentially dangerous vessel behaviours in the surveilled area.

5.3.2. Method for waypoints generation

In order to extract context-sensitive traffic patterns, we extend and combine the existing approaches (presented in Sections 5.1 and 5.2) in such a way that historical AIS data is augmented with historical weather information. This data is first used to generate ship type specific manoeuvre points by applying the CUSUM algorithm (Lamm & Hahn, 2017). Subsequently, these manoeuvre points are used as an input for a genetic algorithm in order to identify waypoints and then generate the mesh and RCs. The mesh consists of nodes and edges, whereas the nodes represent geographical points which are significant for vessel movement. The edges represent the connection between nodes—a vessel can travel from node A to node B if they are connected.

For generating the mesh, first we combine the CUSUM algorithm and a genetic algorithm (GA). The CUSUM algorithm detects the manoeuvre points based on

speed and course changes of vessels (see Section 5.2). Applying this algorithm to a set of historical vessel tracks yields a set of manpeuvre points (which was approximately 10 times smaller than the original dataset). These manoeuvre points are then used as an input to the main method.

In the following paragraphs the consecutive steps of the methods are presented in details.

CUSUM. CUSUM aims at processing the AIS data collected in the system in order to find preliminary waypoints for further analysis. CUSUM analyses trajectories of ships (collection of AIS messages sent by a ship in a given voyage) and detects these messages, which characterize a significant change of speed or course in a given trajectory (detection of significant manoeuvres). These messages are the preliminary waypoints.

By the abrupt change we understand a point on the timeline, at which properties of a current observation change, but before and after this moment, the properties are constant in some sense (Basseville & Nikiforov, 1993). Based on this definition, it is possible to associate AIS signals with a data stream by adopting certain assumptions. The planned or unplanned manoeuvre can be treated as a deviation in a single ship's voyage (trajectory). The main objective of the research was to detect significant manoeuvres of a ship (e.g., a sudden change of course) by a sequential analysis of a trajectory. CUSUM has a few implementations, such as a one-sided algorithm for observations with the expected direction of the changes (Lamm & Hahn, 2017), as well as a two-sided one, which handles increases and decreases of the observed variable. Since the manoeuvres in AIS data can be identified primarily by the increase or decrease of the speed or course, the two-sided algorithm has been taken into consideration. This form of CUSUM can be explained as using two CUSUM algorithms together (Basseville & Nikiforov, 1993).

At the beginning we can assume that AIS signals represent a certain stream of data (Faithfull, 2017):

$$Y = \left\{y_1 + y_2 + ... + y_n\right\}$$

The alarm time is defined as (Basseville & Nikiforov, 1993):

$$t_a = \min\left\{k : (g_k^+ \geq h) \cup (g_k^- \geq h)\right\}$$

t_a point means, that the decision function g_k^+ or g_k^- reached the previously defined threshold h.

We can distinguish the decision function gk (Basseville & Nikiforov, 1993):

$$g_k^+ = \left(g_{k-1}^+ + y_k - \mu_0 - \frac{\upsilon}{2}\right)^+$$

and the negative form:

$$g_k^+ = \left(g_{k-1}^+ - y_k + \mu_0 - \frac{\upsilon}{2}\right)^+$$

As input parameters, three parameters should be provided: μ_0, υ and threshold h. The first one is calculated dynamically and stabilizes the decision function with a moving average value from the last z observations (Lamm & Hahn, 2017). The second parameter, υ, requires the knowledge of the whole trajectory. Lamm and Hahn (2017) recommend using an upper quantile of all $|\Delta y|$, because this measure indicates the structure of a given voyage. In case of the threshold h, it can control the sensitivity of the algorithm. Depending on the requirements, we set this parameter between 1 (higher sensitivity) and 4 (lower sensitivity, with the risk of skipping some significant manoeuvres). The more sensitive the algorithm is, the more alarm points will be detected. So, the algorithm should not be too sensitive if we only look for significant manoeuvres.

It is worth mentioning that in the CUSUM algorithm only a single variable is considered (Y. G. Qi, Martinelli, Teng, & Jiang, 2010). If more than one parameter is required to monitor, it is recommended to integrate the variables.

Spatial partitioning. The genetic algorithm is made up of two main steps. Firstly, AIS points are partitioned using the spatial partitioning algorithm, then the genetic algorithm processes the AIS points, separately for each partition, and detects the final waypoints. The genetic algorithm is implemented in a parallel and distributed manner. To achieve that, we used geospatial partitioning—each partition is treated separately by the algorithm in parallel, and the merged sub-results are the final ones.

One of the main challenges in AIS data processing is their unequal spatial distribution, since densely populated areas *outshine* less popular sea areas. For instance, some areas aren't covered by satellites, which creates gaps in data and thus biases any analyses performed on such data. To mitigate this problem, Dobrkovic et al. (2018) propose to use a concept of QuadTrees. This issue is especially important when one wants to use a genetic algorithm for AIS data processing, since the densely populated areas would represent the fittest genes. The less dense areas would not be inherited, on the other hand. Following their suggestion, we have tested two tree-based data structures for spatial partitioning of AIS data: k-d B-trees and QuadTrees. In this approach, each resulting partition will be later treated separately. The k-d B-trees method is a specific juxtaposition of k-d trees and B-trees (Robinson, 1981). Similarly to k-d trees, a binary tree with nodes storing k-dimensional points is built—the *longest* axis is recursively divided using a hyperplane on a median point. However, the partitions are stored in leaf nodes, which is a feature borrowed from B-trees. In QuadTrees, as the name suggests, each node has *exactly* four children (Samet, 1984). This method recursively subdivides the most dense areas to four smaller ones.

There are other spatial partitioning methods, but not all of them preserve all the qualities relevant for us, such as representing disjoint areas. After performing

some experiments (see Chapter 9), we concluded that k-d B-trees are the right choice, since they minimize the unequal distribution better than QuadTrees.

Genetic algorithm. Genetic algorithms are a biologically inspired family of algorithms, in which the process of evolution is simulated (Sivanandam & Deepa, 2008). The algorithms constitute an important branch of the field of artificial intelligence, named evolutionary computing. Although this metaheuristics is rather a huge oversimplification of the real evolution, it catches the concept accurately. The overall *population* consists of entities called *chromosomes*. Each chromosome is built from *genes*. As in a real population, having good genes results in a higher chance of having an offspring. That goodness is measured by a *fitness function*. Two chromosomes with good genes produce a new one by combining their genes in a *crossover* process. This results in a new population, in which chromosomes with low fitness score are replaced by new ones. Moreover, some random changes are introduced to some of the genes—this process is called *mutation* and is used to maintain population diversity. Each such full cycle is called an *epoch*. There is an assumption that after a sufficient number of epochs, the resulting population will be much better (in terms of fitness) from the initial one. Instead of that, some convergence criterion may be used as a halting point. This approach can be applied to numerous problems. In our case, each gene will represent a waypoint candidate.

We use the genetic algorithm to discover waypoints from AIS data, as it was previously used in the literature (Dobrkovic et al., 2018). Having the AIS points partitioned, the genetic algorithm can be used for each partition to detect the waypoints.

The idea of a waypoint discovery is simple—a good waypoint candidate is a point that has many AIS points in its proximity. Naturally, this needs to be formalized. This can be done with a simple circle equation. First, we need to define a gene—in our case, it's a triple (x, y, r), where x represents longitude, y latitude, and r a radius (a constant for all genes). A single chromosome contains a fixed number of genes, which will be called a chromosome length. A set of chromosomes constitute a population. Contrary to the original paper (Dobrkovic et al., 2018), we initialise our population drawing random AIS points from the actual population (existing AIS points).

Fitness function. The fitness value of a chromosome is calculated using the following formula, which is our fitness function:

$$f = N^{-1} \sum_{i=1}^{N} \#\left\{(x, y) \in \mathbb{R}^2 : (x - x_{c_i})^2 + (y - y_{c_i})^2 \leq r^2\right\} \tag{5.1}$$

where N is the number of points (x, y) in a given partition. Every single gene in the chromosome carries a waypoint candidate (x_{c_i}, y_{c_i}), which actually denotes the

centre of the circle with a radius r (in degrees). The # operator marks the cardinality of a set. Since there are better distance measures than a simple Euclidean distance, this equation was later changed to:

$$f = N^{-1} \sum_{i=1}^{N} \#\left\{(x, y) \in \mathbb{R}^2 : \mathrm{hav}(x, y, x_{c_i}, y_{c_i}) \leq r\right\} \tag{5.2}$$

where hav is the haversine function, a formula for calculating the great circle distance between (x, y) and x_{c_i}, y_{c_i}. The standard haversine formula is presented as follows:

$$\mathrm{hav}(\lambda_1, \phi_1, \lambda_2, \phi_2) = 2r \arcsin\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)} \tag{5.3}$$

where λ_i and ϕ_i represent latitude and longitude (both in radians) of two points between which the distance is to be measured.

Contrary to the previous Euclidean formula, this distance is yielded in kilometres. If a chromosome is eligible for penalty (see the following subsections), its fitness is set to zero.

One-point crossover with a roulette wheel selection and mutation. The crossover operation in the genetic algorithm is an operation in which a new chromosome is generated from two existing ones. Our implementation generates a new population using a roulette wheel selection and one point crossover, as in (Dobrkovic et al., 2018). Conceptually, two parents for a new chromosome are selected using a roulette wheel, hence the name. The new parents are drawn from the whole population not in a uniform way—chances of being drawn are proportional to its fitness. Therefore, the process resembles a roulette with uneven sections.

Having the two parents selected, we use one-point crossover to generate a new chromosome. This procedure draws a random point at which the parents are combined, i.e., genes up to that point are taken from the first parent, and genes from that point out to the end are taken from the other parent. If the same parent gets drawn two times, the resulting offspring will be the same chromosome. Since it does not matter at which point we combine genes from the two parents, it will result in the same one, we can skip that part.

Finally, a mutation takes place. The genes are mutated by replacing one of them at that random position. The mutation is also called if the resulting chromosome has fitness value equals to zero.

Penalties for chromosomes. To prevent the situation in which all waypoints are created in a very dense area (leaving aside the less frequent areas), we introduced a mechanism for penalizing such configurations. A chromosome can be perceived

in terms of two values: fitness and diversity. The first term was already introduced. The second reflects how many different genes a given chromosome consists of. Diversity is just a proportion of the size of unique genes to the size of all genes.

This, however, would only enable one to detect exactly the same waypoint candidates. Since overlapping waypoints (i.e., genes close to each other) have to be penalized, we check if two circles are disjoint by checking the following condition. In the Euclidean distance, it can be presented as a situation in which two circles overlap:

$$(x_1 - x_2)^2 + (y_1 - y_2)^2 \leq (r_1 + r_2)^2 \tag{5.4}$$

Using the haversine function, a similar criterion can be formulated as follows:

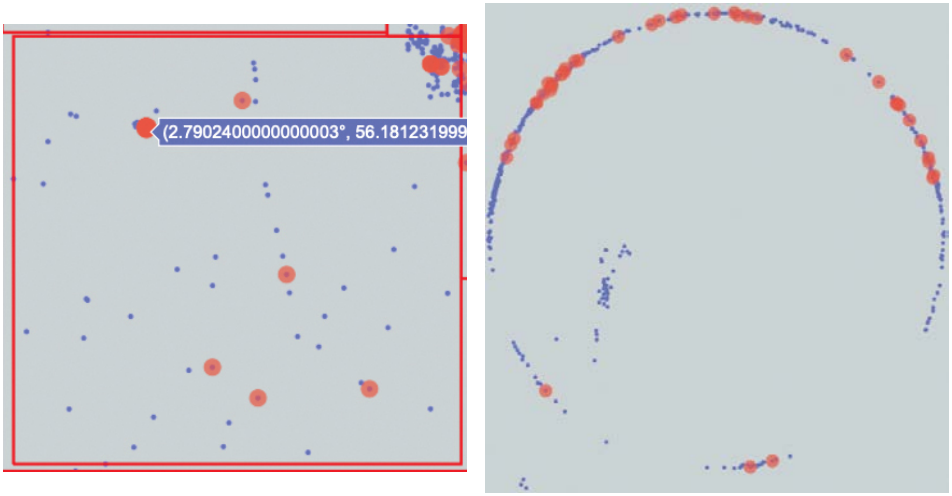
$$\mathrm{hav}(x_1, x_2, y_1, y_2) \leq 2r \tag{5.5}$$

If the condition is not met, the chromosome receives 0 for its fitness score.

To calculate this value, we need a cartesian product for a set of genes without duplicates. Notice that in this context a phrase without duplicates means the same *circles*, not the same coordinates (hence a standard cartesian is not enough). Finally, checking whether a chromosome is eligible for a penalty is done. The chromosome is eligible for penalty either if the minimal diversity score is not reached, or there exists at least one pair of genes which "overlap".

Merging close waypoints. Unfortunately, even penalties can't entirely prevent that a majority of waypoints are created in a single area. Sometimes, a small area is so dense that the algorithm creates numerous waypoints in this area. This behaviour is presented in Figures 5.0 (a) and 5.0 (b)—a suspiciously dense waypoint in open seas is presented. After zooming in, it turned out that this traffic was generated by a single ship traveling in circles for many days. By the definition of our genetic algorithm, this behaviour can't always be avoided. However, to minimize the negative impact of such behaviour in the further edge detection mechanism, we introduced a simple precision loss parameter using a rounding constant.

As a consequence, very close waypoints would be treated as the same and removed as duplicates later. However, setting the proper value is tricky, it's about reaching a compromise between accuracy and simplicity A set of conducted experiments revealed that in case of the HANSA system the value of 5,000 seems to be good enough.



(a) Suspicious "dense" waypoint (b) The same area zoomed in

Figure 5.1. An area with high concentration of AIS data

Source: Own work.

5.3.3. Method for traffic patterns and RC extraction

The genetic algorithm described in the previous section generates a set of waypoints. In this section we describe how mesh was generated. To this end several methods have been tested in an iterative approach. Waypoints are equivalent to nodes of the generated mesh. What is necessary, is to discover the edges, i.e., which waypoints should in fact be connected. It was conducted based on historical AIS data. By looking at every single trajectory of all vessels (that pass an area of interest) we can track which waypoints they 'visited'. It is therefore necessary for each AIS point to assign it a nearest waypoint. The approaches and methods to achieve this are described in the next paragraph.

Having all AIS points annotated with the nearest waypoints, it may seem straightforward to reconstruct the connections between waypoints. Unfortunately, due to incomplete distribution and often low quality of AIS data, several measures need to be undertaken to achieve good results. Details are described below in the paragraph Reconstruction of edges.

AIS enrichment. We refer to the process of adding information about the nearest waypoint to each AIS message as "AIS enrichment". We add both the identifier of a waypoint and the distance to it. To find the nearest waypoints we have used the kNN algorithm available in Sklearn[3]. It was necessary to adjust the metrics

3. https://scikit-learn.org/stable/modules/neighbors.html

as well as the partitioning function. We started with the Minkowski distance metric and k-d tree partitioning but ended up with a haversine distance and ball tree partitioning. Taking into account the number of AIS data to be enriched, the process of assigning waypoints to AIS data proved to be very time consuming. Therefore, we introduced many optimization techniques to make the task feasible. It was also necessary to optimize the code in PySpark i.e., Pandas UDF[4] was used instead of a plain iteration over rows in a CSV dataset. The results and details of the AIS enrichment process are presented in Chapter 9.

Reconstruction of edges. Having the waypoints assigned, the next step is the reconstruction of edges between these waypoints. The general approach to the reconstruction of edges is presented in algorithm 5.1. It lists the filtering functions that are applied either to obtain meshes for different conditions or just to improve the quality:

- FilterAis: the function selects a subset of data for a given vessel type (e.g., tanker) or weather conditions (e.g., heavy wind). It is also used to build a mesh on a subset of points, like only important manoeuvre points as identified by CUSUM.
- FilterTrajectory: the function is applied to trajectories of a ship. It is responsible for the selection of points out of which the edges will be constructed. For example, it can leave only the AIS points that reflect the changes in the waypoint (as in border points). In another variant it is used to consider only the points that are sent within a specific period of time (time-bound).
- FilterEdges: the function is applied to edges. For example, we can filter out the edges that are too long (e.g., distance > 250 km) or are very rare (e.g., followed by only a single ship). The method is mostly applied to visualizations.

The details and results of the method for edge reconstruction are presented in Chapter 9.

Having generated the mesh, the final step is getting an RC. By providing a starting location and a destination of a vessel, an RC might be generated based on the mesh.

Summing up, the added value of the proposed method lies in the speed of computation by applying relevant technologies (here Apache Spark) and optimizing the performance of the algorithms. With the speed of the calculation, it is possible to test many different scenarios and tune hyperparameters of the method to suit individual needs. Thus, the current approach can consider seasonality of movements (daily, weekly, yearly). We have generated many different meshes for different vessel types, different sizes of ships, different draught values, and also for

4. https://databricks.com/blog/2017/10/30/introducing-vectorized-udfs-for-pyspark.html

Algorithm 5.1. Edges discovery

```
1: function DiscoverEdges(AIS)
2:     AIS_f ← FilterAIS(AIS)
3:     AIS_w ← AIS_f.partitionBy('mmsi').orderBy('timestamp_ais')
4:     AIS_w ← AIS_w.withColumn('to_waypoint')          ▷ mark current waypoint
5:     AIS_w ← AIS_w.withColumn('from_waypoint')        ▷ mark previous waypoint
6:     AIS_w ← AIS_w.withColumn('changed') ▷ identify rows with changed waypoints
7:     AIS_c ← FilterTrajectory(AIS_w)
8:     edges ← AIS_c.groupBy('from_waypoint', 'to_waypoint')
9:     edges ← FilterEdges(edges)
10:    edges_d ← edges.withColumn('distance_km')        ▷ calculate distance between
   waypoints
11:    return edges_d
```

different weather conditions, measured on the Beaufort scale and considering the ice coverage where necessary.

5.3.4. System architecture

As described above, RCs are intended to support mariners both in voyage planning and monitoring. Hence, the underlying system for the calculation and provision of RCs must be able to regularly process large amounts of data and extract patterns from them. In the event that the route must be replanned, an alternative route has to be provided within a reasonable period of time. In addition to this, one has to consider that the provided traffic patterns may change in time and that RCs are not static and should be updated. In order to ensure the sustainability of the RC concept, the database has to be extended by new data regularly. Furthermore, the data must be analysed regularly in order to extract the most recent patterns. To meet these requirements, we proposed to use the Lambda architecture.

Figure 5.2 illustrates the architecture of the HANSA system that consists of three layers. The basic AIS and weather data are stored inside the Batch Layer, which will be called the master data set. As soon as new data, i.e., historical AIS and weather data, is available, it will be appended to the master data set. This layer also contains all expensive operations such as pattern extraction, mesh calculation and RC generation. It considers also some other static data sources, such as ports coordinates, or geographic information (continental contours, e.g.). The Lambda Architecture of the system is intended to separate batch data processing and calculations operations (such as pattern extraction) from requests, which tend to be resource consuming (Filipiak, Stróżyna, Węcel, Abramowicz, & Steidel, 2021).

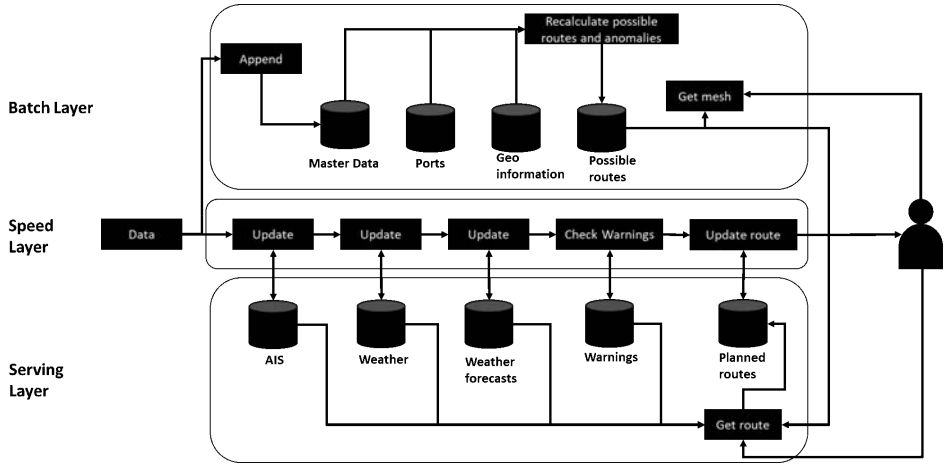


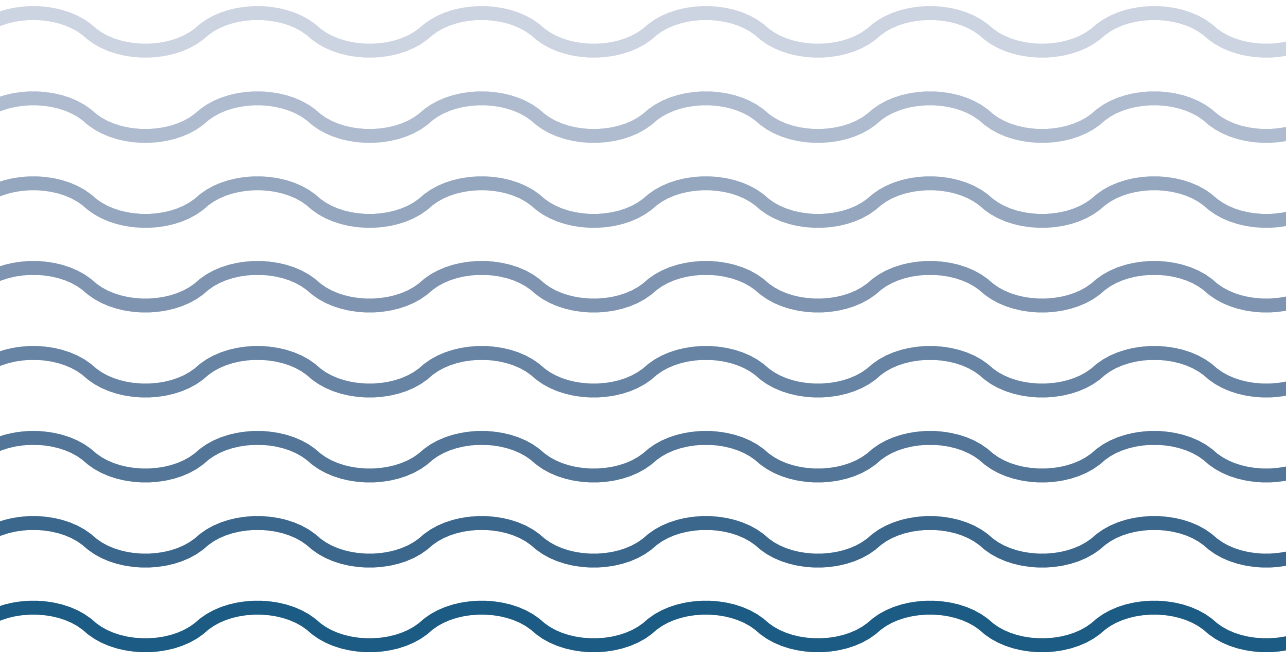
Figure 5.2. The Lambda architecture of the HANSA system

Source: Own work.

As in a standard Lambda architecture implementation, the end user, who e.g., wants to request an RC, accesses the Speed layer. The results and data provided by the Batch and Speed Layer are merged and are made available to the user.

The implementation details of the HANSA methods as well as the received results of methods evaluation are presented in Chapter 9.

Chapter 6



6. MARITIME ANOMALIES DETECTION

In the recent years, an increasing number of dangerous or strange behaviour at sea has been observed. In the EU context, illegal migration from African and Middle East countries is a particularly serious problem. In other regions of the world piracy and robbery as well as smuggling are real issues. Dangerous behaviour of ships may also lead to collisions. This, in turn, requires ensuring protection of sovereignty and infrastructure, counteracting terrorism and piracy from governments, or protection of the environment. As a result, the maritime domain faces a problem of detection and anticipation of such anomalous behaviours at sea. In order to achieve that, anomaly detection became one of the main issues of Maritime Surveillance. Usually, the surveillance is assured by operators that search and predict anomalous or conflict situations using surveillance systems. However, exploring and monitoring the anomalies may become a demanding task for operators due to two reasons: 1) the complexity, heterogeneity, dynamism and increasing number of data to be analysed; 2) information overload and a limited ability of humans to process huge volumes of data related to sea traffic. Therefore, surveillance operators need support in their daily activities by methods and systems with anomalies detection capabilities.

This chapter provides a review of the approaches and methods that have been used to detect maritime anomalies, preceded by an attempt to define what maritime threats and anomalies are as well as a categorization of maritime anomalies. Finally, the methods for detection of selected maritime anomalies, that have been developed in the SIMMO project, are presented.

6.1. Maritime threats and anomalies

Maritime threat is a possible danger, which happens at sea and may result in possible harm to states, organizations, people or objects. This harm can be of different nature such as economic, environmental, health related, or defensive. It is seen as something broader than anomaly.

An anomaly is in general “a deviation from the expected” (van Laere & Nilsson, 2009). An anomalous behaviour is a behaviour that is “inconsistent with or deviating from what is usual, normal, or expected, or that is not conforming to rules, laws, or customs” (Roy, 2008). Anomalies are defined as objects, observations

or patterns that do not conform to a well-defined notion of a normal behaviour (Chandola et al., 2009).

Thus, anomaly detection refers to the problem of finding vessels that behave differently from most other vessels (searching for unusual behaviour) and evaluating their threat potential (Martineau & Roy, 2011). From the point of view of data analysis, an anomaly is a rare item, event, observation in a data space, which represents a deviation from standard behaviour, or which appears to be inconsistent with the remainder of the dataset (Hodge & Austin, 2004).

Nowadays, a great variety of maritime threats and anomalies can be observed. Maritime threats may concern such activities as (Bakir, 2007; el Pozo et al., 2010; Lane, Nevell, Hayward, & Beaney, 2010; Riveiro, 2011): piracy and ship hijacks, trafficking, such as illegal migration, narcotics trafficking, smuggling of illegal goods, human and drug smuggling across maritime borders, stowaways and seaborne terrorism, illegal transhipment, grounding, collisions, oil spills, pollution, trash disposal. An anomaly may occur with a sudden change of speed, a deviation from the standard route, and a close proximity to other object on high sea.

According to the International Maritime Organization (2021) between 2005 and 2020 there were 4785 incidents of piracy and robbery worldwide, of which 22% occurred in East Africa and 26% on the South China Sea. In these incidents more than 300 ships were hijacked, 5310 crew members were held hostage, 490 people were assaulted, 320 people were wounded, and 77 people lost their lives (see Table 6.1). Political conflicts, including wars and terrorist attacks, are another source of risk. Major terrorist hubs are located in coastal regions in Sri Lanka, Yemen, Pakistan, the Philippines, and Indonesia (Lam, 2012). Accidents are yet another issue. Table 6.2 presents statistics on maritime accidents showing that each year hundreds of maritime accidents occur that not only decrease the security of ships, people, and the transported cargo, but also influence the risk and reliability of maritime transport services. However, a sharp decrease in 2020 did not result from a sudden increase of maritime security. Instead, it was due to a significant decrease of maritime traffic related to the COVID-19 pandemic.

6.2. Typology of maritime anomalies

Maritime reality shows that there is a large variety when it comes to anomalous behaviour of ships. There are also various classifications of maritime anomalies which can be found in the literature (Riveiro et al., 2018).

Maritime threats encompass mainly:

- Illegal immigration.

Table 6.1. Regional analysis of reports on acts of piracy and armed robbery in total in 2005–2020

Location of incidents	South China Sea	East Africa	West Africa	Indian Ocean	Worldwide
Total number of incidents reported	1249	1096	504	795	4785
Ship hijacked	43	195	40	17	336
Lives lost	20	10	31	6	77
Wounded crew	78	38	116	44	320
Crew hostage	459	3119	505	805	5310
Crew assaulted	142	35	125	122	490

Source: Based on data available in (International Maritime Organization, 2021).

Table 6.2. Reported maritime accidents per year in 2005–2020

Location of incidents	Total number of accidents reported
2005	299
2006	487
2007	400
2008	332
2009	246
2010	455
2011	350
2012	425
2013	355
2014	219
2015	317
2016	370
2017	347
2018	234
2019	219
2020	37

Source: Based on data available in
(International Maritime Organization, 2021).

- Smuggling and transnational crime at sea.
- Threats against freedom of the seas and maritime trade, including energy security.
- Potential expressions of terrorism and piracy at sea.
- Degradation of the marine environment.
- Conflicts and crises in the periphery of Europe (e.g., Russia and Ukraine, Turkey, Syria).

Andler et al. (2009) gathered different types of maritime anomalies in a taxonomy that classifies them into six groups, thus creating a good depiction of threat types to be dealt with in the maritime domain:

(1) Rendezvous (object or location):
   (a) Rendezvous with an air plane.
   (b) Rendezvous with a small boat.
   (c) Rendezvous between vessels.
   (d) Rendezvous between a mothership and a small boat.
   (e) Simultaneous arrival of ships with similar threat profile.
   (f) Illegal, unreported and unregistered fishing vessels entering ports of regional fisheries management organizations.
(2) Movement:
   (a) No match between the position and AIS.
   (b) Time between ports does not match expected route.
   (c) Stop and go in many harbours with pollutive cargo.
   (d) Vessel type does not match movement.
   (e) Ships crossing the same point within a limited time.
   (f) Unusual routing:
      (i) Presence in a non-typical area.
      (ii) Outside normal routing.
      (iii) Abnormal deviation from route compared to port of call.
      (iv) Presence in an area of piracy.
      (v) Fishing in a closed area.
(3) Cargo:
   (a) Cargo of high interest.
   (b) Cargo does not match port of call.
   (c) Cargo does not match crew.
   (d) Dangerous cargo.
(4) History:
   (a) Agreement of ship type, harbours visited, crew etc.
   (b) History of AIS spoofing.
   (c) Involved in smuggling.
   (d) Change of ownership flag.
   (e) Ship's whereabouts 3–6 months.

(f) Registered in a state with criminal ships.
(g) Registered in a land of high interest/suspicious flag state / suspicious port.
(h) Visited port of interest.
(i) Visited harbours known for criminal activities.
(j) Casualty history—worse than normal for type of vessel compared to industry.

(5) Owner / crew:
(a) Changing of crew members during route.
(b) Crew / owner with a criminal history.

(6) Tampering:
(a) Intelligence report on smuggling of drugs, weapons from previous harbour.
(b) AIS turn-offs.
(c) Non-cooperative behaviour.
(d) Oil spills.
(e) Possible changes in status during transit.
(f) Incorrect IMO number against reference dataset.
(g) Change of name in AIS under route, change of name with the same MMSI code.

Similar analysis was made by Roy and Davenport (2009), who additionally paid attention to the quality of the transmitted AIS data (e.g., missing or impossible data), criminal activities of ships, motion of ships (e.g., drifting, loitering, too high speed), and its position (e.g., proximity to other objects, presence in restricted zones, traveling outside normal or historical routes).

From the point of view of their location, maritime threats can be divided into three categories: 1) threats in harbours; 2) threats in coastal/territorial waters; 3) threats on high sea (Figure 6.1).

Maritime anomalies in ships behaviour can also be categorized as *static anomalies* and *dynamic anomalies*, which can be further divided into non-kinematic and kinematic. Examples of such anomalies are presented in Table 6.3.

Attention should be paid to the fact that the majority of merchant vessels are from open registries ("flag of convenience"), which often do not assure their compliance with international safety and security standards (el Pozo et al., 2010). This trend creates potential issues such as environmental threats and illegal or criminal activities because control by the flag states is often ineffective or none. It is worth emphasizing that illegal activities are not confined to territorial waters or Exclusive Economic Zones, but occur in distant waters. Therefore, these vessels present states with significant problems of protection and security.

Another threat is charter frauds and the risk of cargo loss. There are a number of maritime companies that, despite their poor financial condition and low reliability, still conduct charter services and transport cargo. Then, when a service is in progress, it may turn out that the cargo cannot be delivered on time (e.g., due to

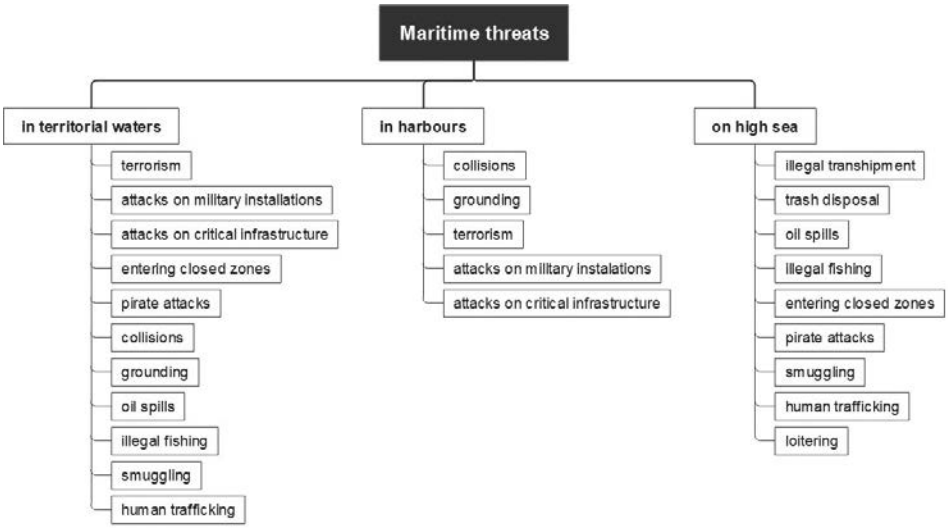


Figure 6.1. A typology of maritime threats

Source: Own work.

lack of money on the part of the ship owner to pay for the fuel, port charges, or the crew) and the owner of the cargo must incur additional costs to regain its property. Thefts of cargo or whole vessels by pirates or terrorists is another issue.

Table 6.3. Classification of categories of anomalies

Static anomaly	Dynamic kinematic anomaly	Dynamic non-kinematic anomaly
Vessel name	Course	Next or last port call
Flag	Speed	Cargo list
MMSI, IMO number	Manoeuvre	Draught
Owner	Reporting	Crew list
Port of registry	Location	Passengers

Source: Based on (Riveiro, 2011; Roy, 2008).

The studies by (Roy, 2008; van Laere & Nilsson, 2009) collected and prioritized user requirements regarding what types of anomalies should be detected by maritime systems. These requirements were obtained from different stakeholders such as experts working in armed forces, coast guards, ports, operation centres in the maritime domain, so they can be treated as rather exhaustive for maritime surveillance. For example, there are anomalies which Vessel Traffic Service (VTS) centres are particularly interested in. VTS operators in ports focus on the safety of port facilities. They indicate indicate:

- Control of dangerous goods.
- Identification of persons entering and leaving the port.
- Detection of explosives.
- Thefts in containers.
- Environmental issues.

On the other hand, operators in coastal VTSs pay special attention to:

- Grounding situations and collisions.
- Vessel entering restricted zones.
- Identification of unknown vessels.
- Vessels not following the standard route for the declared destination or sea lanes.
- Cargo of special interest.
- Vessels carrying dangerous goods sailing close to passenger ships or protected areas.
- Vessels with a history of being involved in illegal activities.
- Suspicious flag or port.
- Discovering oil spills and floating objects.
- Overturned boats.

Based on the maritime threats presented in the literature, we developed a typology of maritime anomalies (Figure 6.2) that summarizes the results of other studies and divide anomalies into the following categories: (1) Movement; (2) History; (3) Cargo; (4) Crew; (5) AIS reporting.

The AIS (*Automatic Identification System*) reporting anomalies exemplify issues which describe an intentional modification in transmitted AIS messages (more information on AIS and AIS data is presented in Section 4.1). Such modifications may concern a ship's position, static and dynamic data or switching off an AIS transponder, in order to hide the current activity of the ship. The category *crew* relates to characteristics of the people involved in the ship's activity. *Cargo* represents threats connected with characteristics of the goods transported by a ship. *History* relates to the historical behaviour of a ship, captured in different registers. And the largest category—*movement*—distinguishes possible abnormal behaviour of a ship at sea or in a harbour.

The presented list of maritime threats is of course not complete, but it presents the scale of the problem and indicates the need of dealing with these threats with appropriate risk management and by proposing methods for their automatic detection and the development of maritime surveillance systems.

To sum up, the conducted analysis shows, that there are a number of various maritime threats and anomalies, whose detection is of special interest for different entities working in the maritime domain. However, providing surveillance systems, dealing with all possible anomalies and being able to detect new abnormal

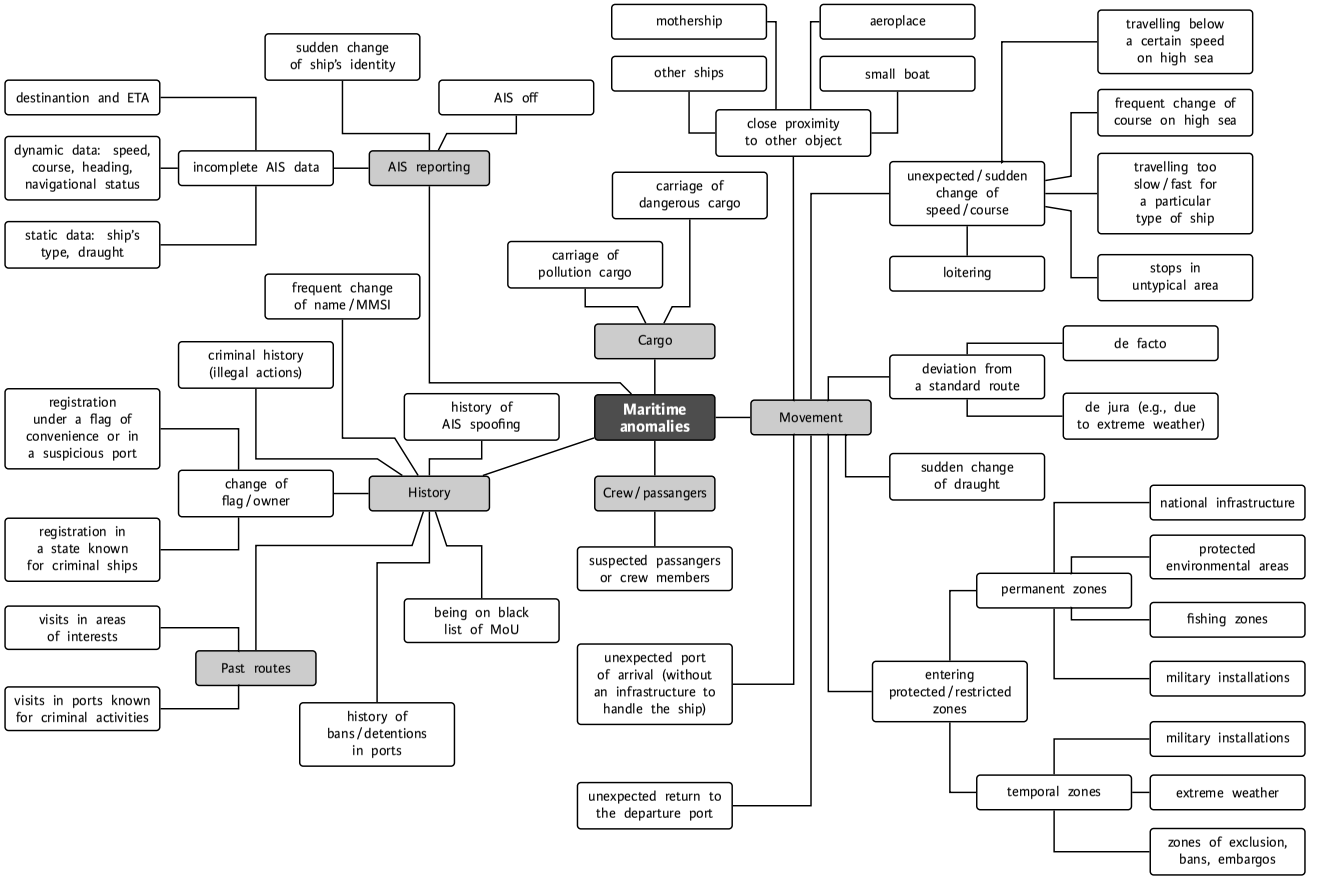


Figure 6.2. A typology of maritime anomalies

Source: Own work.

behaviours would require a huge amount of work and seems quite impossible. Therefore, we selected maritime threats and anomalies that are particularly interesting from the point of view of the research presented in the book, namely the process of risk and reliability assessment based on the data fused from various maritime-related data sources (chapters 2, 7, and 8), and the SIMMO project (Sections 4.6 and 6.4). These threats and anomalies encompass:

- Ships that do not report all the required data through AIS (inconsistent or missing data),[1] for example:
  - Incomplete vessel movement data, e.g., Speed over Ground (SoG), Course over Ground (CoG), wrong heading.
  - Incomplete static and voyage data, e.g., ship name, destination, time of arrival, type.
  - Ambiguous information: mismatched information or wrong entries, e.g., wrong ship type, MMSI, IMO etc.
  - Frequent change of vessel identity (MMSI), flag state, or ownership.
- Ships flying a black-listed or FOC flag, or registered in a black-listed port.
- Ships with a history of suspicious or dangerous behaviour:
  - Ships with a history of having a withdrawn or suspended classification status or belonging to a not well-respected classification society.
  - Ships with a history of calling suspicious ports.
  - Ships with a history of being on a detention or banned list.
- Suspicious or abnormal behaviour:
  - Deviations from a standard route.
  - Transhipment of cargo on sea.
  - Traveling through protected maritime areas.
  - Loitering at seas, e.g., unreasonable low / high SoG, sudden or unreasonable high frequency of course changes on high sea.
  - Liaison with other vessels on high sea.
  - Unreasonable switching off of AIS transponder or non-availability of LRIT data.
  - Location manipulation and AIS spoofing.

Each of these anomalies requires a specific method for its detection. For some of the anomalies such methods have already been developed and published in numerous papers. We briefly describe them in the following section. Still, the existing surveillance systems supports the users in detecting only a limited number of anomalies.

1. AIS system is described further in Section 4.1.

6.3. Anomalies detection: Approaches, methods

As presented in the previous section, the variety of maritime threats and anomalies is quite extensive. In general, the literature presents various methods for the detection of maritime anomalies. A few studies are actively being conducted in this area by several research communities—public and private organizations. Since there are already several papers available that review the recent achievements in this area (Riveiro et al., 2018; Sidibé & Shu, 2017; Tu et al., 2017), our goal is not to duplicate this work and rather focus on selected methods, those which are somehow related to the maritime threats and anomalies which we try to detect in our research (see the list in the previous section). Those readers who are particularly interested in a wider overview of the methods for maritime anomalies detection are encouraged to look at the works mentioned above.

Anomaly detection in the maritime domain is a complex process which requires acquisition of information from various sources, integration of this information and finally analysis of events with an operator's expertise and experience in order to detect abnormal behaviour (Matthews, Martin, Tario, & Brown, 2009). Moreover, a number of different aspects need to be taken into account, such as the type of input data, the type of anomaly, class labels and output data (Chandola et al., 2009). Possible patterns in data should also be considered. The specific problem associated with the detection and prediction of anomalous behaviour at sea is that normalcy is dependent on the context in question, therefore its detection requires application of different approaches, techniques, and data. As a result, the anomaly detection process must often be tailored to an application domain and properties of data (Brax, 2011). It is almost impossible to develop a system that recognizes and detects every type of abnormal behaviour. Still, over time, a variety of anomaly detection techniques have been developed.

The anomalies detection methods can be divided into three groups: data-driven, knowledge-driven and hybrid that combine data- and knowledge-driven methods (Kazemi et al., 2013; Lane et al., 2010). The main assumption of the knowledge--driven techniques is utilization of an external source of knowledge (i.e., expert knowledge). This group encompasses different representation techniques and reasoning paradigms such as rule-based, description logic, and case-based reasoning. In the maritime domain the knowledge-based systems use the expertise of an operator/analyst, which is then applied to different knowledge representation paradigms, such as if-then rules, situation cases, or description logic. Roy (2008) proposed an automated reasoning service that exploits ontologies expressed in description logic to support maritime staff in classifying vessels of interest and in identifying and categorizing maritime threats. However, the system proposed by Roy (2008) suffers from severe performance issues, e.g., reasoning is very slow and scales badly with the increasing number of ships. Moreover, most of the knowledge-driven solutions require constant updating because the classification

of events (e.g., events which so far have been known to be illegal) may change over time to be normal, and vice versa. The next challenge is the process of retrieving knowledge from subject matter experts in order to feed such systems (van Laere & Nilsson, 2009).

The vast majority of anomaly detection techniques used in the maritime domain is data driven. In the data-driven approach, anomaly detection is one of the six common classes of data mining tasks itself (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Data mining is a process of discovering trends and patterns in large datasets involving methods from such fields as statistics, machine learning and artificial intelligence. These tasks consist in the identification of unusual data records, which in turn can be either interesting for further analysis or data errors that require further attention. These methods look for, for example, abnormalities in the maritime traffic and estimate the degree of deviation from a learned normalcy (Riveiro et al., 2018)—here mainly unsupervised solutions are being developed. Data-driven methods may also look for predefined patterns in data such as a specific change in a ship trajectory (e.g., loitering vessels). Some of these methods are characterized below.

The data-driven approaches can be further divided into three classes (Chandola et al., 2009): classification-based methods, clustering-based methods, and statistical methods (parametric, non-parametric).

Among the classification-based methods the Bayesian network is commonly used to detect single-point vessel anomalies (such as location, course, speed) (Helldin & Riveiro, 2009; Johansson & Falkman, 2007; Mascaro et al., 2014), piracy on oil platforms (Bouejla et al., 2014), or to detect anomalies in the whole scenario of cooperation between ships (Fooladvandi et al., 2009). In 2005, a fuzzy ARTMAP classifier was proposed as a solution to maritime situation monitoring and increasing awareness (Rhodes et al., 2005). Detection of single point anomalies, i.e., those related to just one parameter, like speed, with the Bayesian network approach was first presented in 2007 by Johansson and Falkman (2007). They observed that Bayesian networks offer two interesting advantages over other approaches in anomaly detection: 1) Bayesian models are easily understood by non-specialists and 2) they allow a straightforward incorporation of expert knowledge. Then, Fooladvandi et al. (2009) proposed signature-based activity detection using Bayesian networks, based on knowledge acquired from experts.

Recently methods that apply various machine learning algorithms for anomalies detection have also emerged in the literature, e.g., neural networks (Nguyen et al., 2021; Singh & Heymann, 2020; Venskus, Treigys, Bernatavičienė, Tamulevičius, & Medvedev, 2019; Zhao & Shi, 2019) or deep learning (Hoque & Sharma, 2020; Karataş et al., 2021). They are used to predict a vessel trajectory, and based on the comparison to normal routes, provide information about anomalous behaviour. Neural networks were first used to detect anomalies in a ship's speed, position, and course (Bomberger et al., 2006; Rhodes et al., 2005). Newly, a method proposed by

Hoque and Sharma (2020) makes predictions based on a vessel's current course and detect anomalies taking into account the vessel's trajectory and engine behaviour. Nguyen et al. (2021) uses neural networks to learn a probabilistic representation of trajectories of ships and a contrario detector which detects location-dependent abnormal behaviours. A multi-class neural network was also used to classify intentional and non-intentional switching off of the AIS transponder based on a ship's position, speed, course, and timing (Singh & Heymann, 2020).

Among clustering-based methods the DBSCAN algorithm is applied to detect anomalies in a ship's speed (Kraiman, Arouh, & Webb, 2002; Pallotta et al., 2013) to identify some popular entry or exit points to a particular area (Pallotta et al., 2013) or to identify loitering (Patino & Ferryman, 2017). Recently, DBSCAN was further combined with Recurrent Neural Network (Zhao & Shi, 2019) or the Long-Short Term Memory (LSTM) architecture (Karataş et al., 2021) to predict vessel trajectories and then detect anomalies.

Another clustering method is the K-mean method used to detect various activities of ships (Tun, Chambers, Tan, & Ly, 2007). Hierarchical clustering methods are also used to learn the typical sailing patterns, which then can be combined with probabilistic methods (the Naïve Bayes classifier) (Zhen, Jin, Hu, Shao, & Nikitakos, 2017) or methods for measuring similarities between two trajectories (e.g., Longest Common Subsequence—LCS algorithm) (Karataş et al., 2021) to detect anomalous behaviour.

The third group of maritime anomaly detection methods are the statistical ones. Here both parametric, like regression and Gaussian Mixture Models (Kraiman et al., 2002; Lane et al., 2010; Laxhammar, 2008; Riveiro et al., 2008), and non-parametric, like kernels Gaussian Process methods (Brax, 2011; Laxhammar, Falkman, & Sviestins, 2009; Smith, Reece, Roberts, & Rezek, 2012), are used.

Laxhammar examined unsupervised methods for analysis of normal sea traffic patterns (Laxhammar, 2008; Laxhammar & Falkman, 2010; Laxhammar et al., 2009). It combines the Gaussian Mixture Model and the greedy version of Expectation-Maximization algorithm. This method was later compared with Adaptive Kernel Density Estimation (Laxhammar et al., 2009), which shows its superiority to the Gaussian Mixture Model (taking into account modelling normalcy and detecting anomalies). Later on, Laxhammar and Falkman (2010) introduced conformal prediction for distribution-independent anomaly detection in streaming vessel data in which no statistical assumptions nor usage of threshold are required. They used it for detection of discrepancies in ship type, between the type declared by a ship and the type estimated based on a ship's location and speed. The GMM was also used by Kraiman et al. (2002) and Riveiro et al. (2008) to detect anomalies in a ship's speed, sudden change of course and suspicious location, and by (Lane et al., 2010) to detect deviations from the standard route. Brax (2011), in turn, included the distance between two moving objects and context information in his model. His State-Based Anomaly Detection requires building normalcy models to

denote behaviours which are not suspicious and discover those which are. Smith et al. (2012) proposed usage of a Gaussian Process, based on GPS data, to detect anomalies in location, sudden change of route, anchoring and drifting.

Probabilistic graphical models for small vessel threats were examined by Auslander, Gupta, and Aha (2012) with Hidden Markov Models, Conditional Random Fields and Markov Logic Networks. The research question was whether all of these methods can outperform the rule-based perimeter maritime threat detection model. It turns out that only Markov Logic Network can provide better results, but it is much slower (including training time and inference time) than the rule-based model.

Recently also stochastic process modelling was considered to detect if a vessel deviates from a planned route by changing its normal velocity (d'Afflisio, Braca, Millefiori, & Willett, 2018). This model additionally detects if a vessel switched off its AIS transponder for a certain time and then tries to revert to the previous, normal velocity.

Riveiro et al. (2008) made an effort to improve maritime anomaly detection and situation awareness through interactive visualization using Gaussian Mixture Models. The process of abnormal behaviour detection is divided into acquisition, processing and analysis. Building a normal behaviour model relies on training data and then clustering it using Self Organizing Map and using Gaussian Mixture Models to process real world information. Similarly to Johansson and Falkman (2007), it tests whether a given observation exceeds a threshold value. If so, the end-user should analyse it and decide whether it was a false alarm. Fischer and Bauer (2010) in their paper suggest object-oriented world model (OOWM). Contrary to the previously presented systems, OOWM is focused on providing an interface for doing more high-level operations, like anomaly detection, rather than doing it itself. The system provides object-oriented representation of instances (vessels) and its attributes via access interface, which may be queried from external applications.

The Open Data Anomaly Detection System (ODADS) was presented by Kazemi et al. (2013). The most distinct feature of this solution is using open and closed data sources. In terms of architecture, all the information goes to a data storage through a data collector module. Then, an anomaly detector, which uses knowledge (expert rules) and data driven (statistical techniques) approaches, can distinguish abnormal behaviour. The obtained results are accessible via a display client.

GeMASS (GEnetic algorithm knowledge discovery for MAritime Security System), described by C.-H. Chen et al. (2014), supports a knowledge discovery process in maritime anomaly detection by using a genetic algorithm. Regarding the architecture of this system, data pre-processing (raw AIS data translation), real-time ship analysis (responsible for knowledge inference) and modules for decision/result update (for obtaining training datasets), knowledge discovery (contains the mentioned genetic algorithm) and data post-processing (for data accumulation) can be distinguished.

A review of the literature shows that for anomaly detection not only the methods which are applied are important but data representation as well. Data representation restricts what can be learnt and as a consequence what kind of anomalies can be discovered. It is also about the richness of attributes. Mascaro, Nicholson, and Korb (2011) studied possible advantages of including additional variables beside those from AIS, e.g., those related to the ship, the weather, and the time-related factors.

Finally, one of the key aspects in maritime anomaly detection approaches is "discretization"—we need to make a discrete decision, whether some behaviour is anomalous or not. But anomaly detection requires also discretization of other typical factors, like location, speed, course, etc. (Johansson & Falkman, 2007). A ship's behaviour in turn can be decomposed into events in order to learn whether an event reflects a routine activity or not.

This short overview of the methods used for the detection of maritime anomalies shows that there are a number of possible approaches. Still, there does not exist a single, general method for the detection of different types of anomalies. Moreover, in each of the proposed methods there is a huge potential for further improvement, for example by inclusion of additional data sources. Also, in the existing maritime systems, automatic and real-time detection of maritime threats is supported in a limited scope.

Another important aspect of maritime anomalies detection is the ability to process, store and analyse maritime-related data. This process requires utilization of multiple sources of data, which were described in Chapter 4. Among them, an important source of information is inter alia AIS, which generates a huge volume of data every day. Within a timespan of a few years, they stack up to terabytes of data, which then needs to be pre-processed (decoded), stored, and analysed. Therefore, there is a need for fast and efficient analytical methods that would have a potential to detect potential threats in real-time and support users in decision-making. Hence our motivation for the development of such methods. The results of our work in this area are presented in the next section as well as further on in Chapter 9. Data-driven methods developed for the SIMMO system (see section 4.6) and the results of conducted experiments are elaborated upon there.

6.4. Loitering-related anomalies detection

In this section one group of methods developed in the SIMMO project is presented, namely methods for detection of loitering behaviour at sea. Loitering can be related to:

- a ship's speed: a ships is traveling very slowly despite being on the high sea,

- a ship's course: a sharp change of course,
- a ship's location and route: unpredictable location or movement,
- non-typical travel time.

We distinguish three types of loitering detection methods—speed anomaly, route anomaly and travel-time anomaly. For all the three types, worldwide AIS data collected in 2015 are used as an input and the analytics is carried out in a batch. An in-depth description and applicability of the methods are presented in the next subsections.

6.4.1. Speed anomaly

A simple means to detect loitering is to look at the speed of a vessel. When a vessel moves on the high sea with a speed which is too slow for the class of the ship, it is an indicator of anomalous behaviour. At first, an assumption was adopted that a loitering occurs when "travelling at the speed less than 5 knots". While this could be true for the high sea, it is not correct for areas where sailing possibilities are restricted by external areas like straits. However, application of this rule would first require the annotation of the whole world with types of waterways. Likewise, if the ship is moving too fast, it can be dangerous, especially in an area with dense traffic or in ports. All in all, when the speed of a vessel deviates, then a warning should be issued. In order to be able to carry out such reasoning, we have to define the notion of *normal speed*, to which we can relate. Such a speed should be location-specific, i.e., defined for a certain geographical area.

In fact, such a "map" of normal speeds can be learnt from historical data. Therefore in our approach we tried to generate such a map. To this end first we divided the globe into sectors. In our experiments we used parallels and meridians to form quadrangles of 5 by 5 degrees. Then we calculated the average speed of vessels within sectors. We originally took the absolute speed, but the results were not satisfactory. The second approach was based on a relative speed, i.e., the current speed divided by a maximum speed of the analysed vessel. The assumption is that on the high sea vessels travel at full steam. Thus, the relative speed is comparable between vessels, whereas the absolute speed depends on a technical capacity. We additionally calculated variations of the relative speed to identify the areas where speed is more diverse, e.g., in ports.

Our hypotheses and results are presented in a series of figures. We start with an overview of the number of messages sent within a given sector (see Figure 6.3). As can be seen in that figure, there is a single region standing out with respect to the number of messages, namely the English Channel. There were over

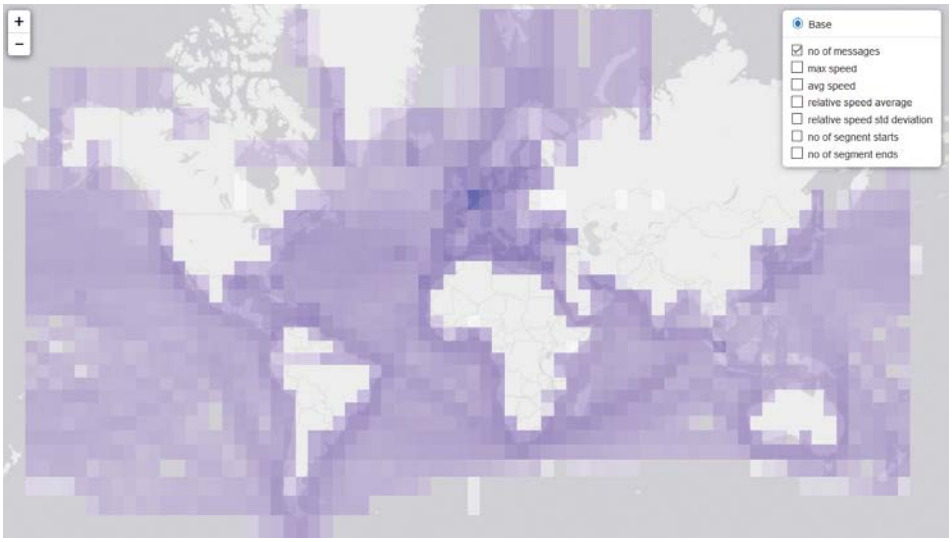


Figure 6.3. Number of messages sent from segments, worldwide

Source: Own work.

35 million messages sent in 2015.[2] The vast number of messages is probably caused by the popularity of this waterway, but it may also be related to the presence of terrestrial AIS receivers. This is important as we observed problems with synchronisation of clocks between various devices. Therefore, many anomalies detected in this crowded area are just false positives.

In the original approach we calculated the average speed of all vessels, based on AIS messages sent from a given region. After checking the quality of the data and devising some algorithms for quality improvements, we re-calculated the average speed based on the cleansed data (Figure 6.4). Again, the English Channel stands out but now it is a low average speed that is characteristic. Another specific region is the Canary Islands. The average speed on the high sea (e.g., the Atlantic Ocean) is relatively stable.

Measuring the average speed is not appropriate, because various ships have different capabilities with relation to the maximum speed. Therefore, a more sophisticated method for speed evaluation has been proposed. Instead of looking at the absolute speed, a relative speed, i.e., the current speed of a ship with relation to its maximum observed speed (not the one declared by a shipyard) was taken into account (see Figure 6.5). The relative speed at the English Channel was 0.13, which should be interpreted that, on the average, vessels travel at 13% of their maximum historical speed. The vessels rarely travel at full steam, even on the the high sea. Usually, it is 70% of their maximum speed.

2. It is important to see that the colours are based on a logarithmic scale, so the supremacy is even higher than visually interpreted from the figure.

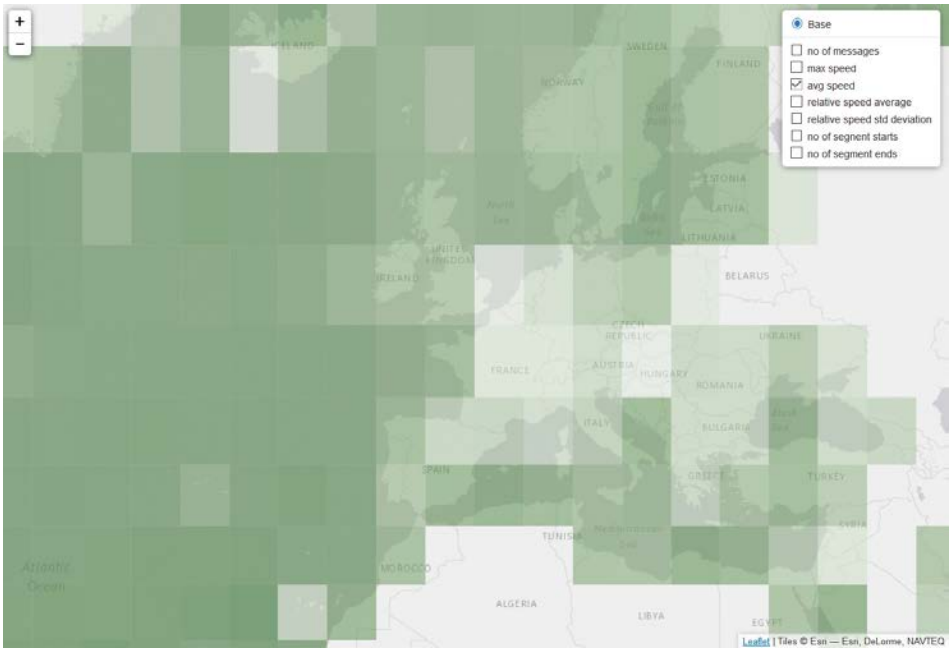


Figure 6.4. Average speed of vessels in a given segment, Europe

Source: Own work.

In order to form conclusions about anomalies in a given sector, we need to consider the variability of the relative speed. On the one hand, the speed of vessels can be relatively stable and then even a minor deviation could be a case of a loitering. On the other hand, in some regions, like near ports, there are ships traveling with high as well as close-to-zero speed. In statistics, there is a measure to characterize this variability—standard deviation. In Europe (Figure 6.6), the Mediterranean Sea is characterized with the highest variability in speed. Particularly, the region including the Suez Canal has a standard deviation of relative speed 0.33. Typically, on the high sea it is 0.10. Again, the Canary Islands provide an exception with the deviation of 0.33.

Having calculated the statistics for the whole globe (the relative speeds in particular), detection of anomalies related to speed can be conducted. The algorithm is as follows:

(1) Take a next point from the trajectory of a given vessel.
(2) Note the maximum historical speed of the vessel and calculate the relative speed.
(3) Calculate the region to which the current point belongs to.
(4) Get the average relative speed and the standard deviation for the region from the respective table (calculated earlier as devised in this section).

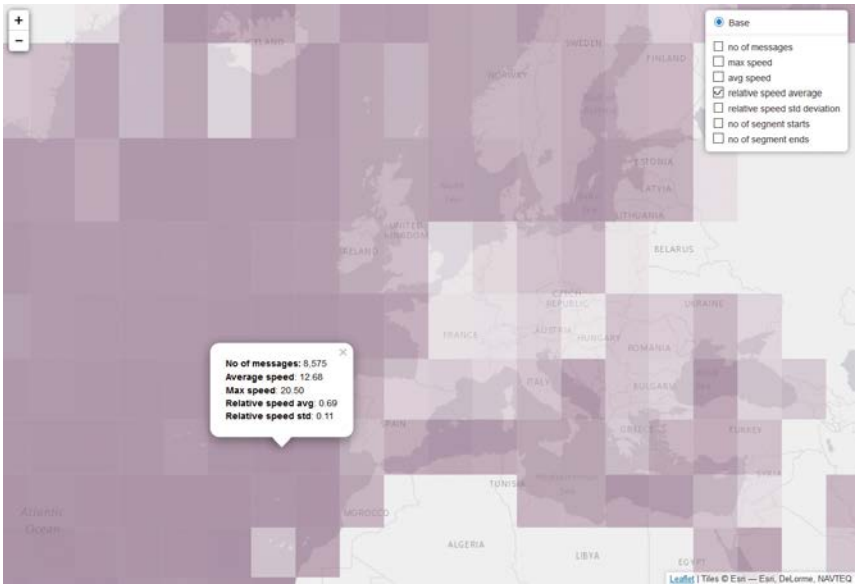


Figure 6.5. Average relative speed of vessels in a given segment, Europe

Source: Own work.

Figure 6.6. Standard deviation of the relative speed of vessels in a given segment, Europe

Source: Own work.

(5) Compare the relative speed with the average relative speed characteristic for a given region. The allowed deviation is determined by standard deviation.

The algorithm required some tuning, i.e., it had to be decided what the reasonable deviation was. Figure 6.7 compares two variants: with 1-σ and 2-σ, where σ is the standard deviation. The latter seems to return fewer false positives, therefore in further experiments this value was kept. However, this value can be further parameterized, if needed. The meaning of the colours is as follows:

- red: the ship is traveling at the relative speed lower than the average relative speed *minus* 2 times standard deviation (2σ);
- green: the ship is traveling at the relative speed higher than the average relative speed *plus* 2 times standard deviation.

The red-marked messages are considered as loitering.

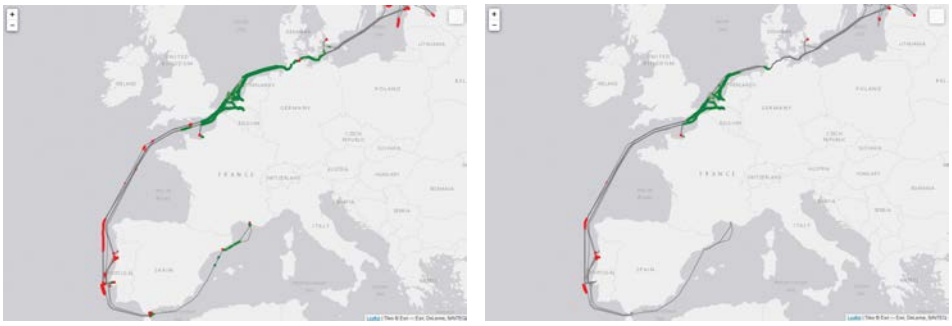


Figure 6.7. Relative speed anomalies with two deviation variants, MMSI 210688000

Source: Own work.

Route anomaly. The route anomaly is defined as an unpredictable movement, i.e., not following a trend or a pattern. The most typical examples are a sudden change of speed or course over ground. In this approach a prediction is made based on a current trajectory of an analysed vessel. We analyse the trajectory and based on the last three locations (from AIS) we extrapolate the next location.

During this analysis we came across several sub-types of how anomaly can be discovered:

- average speed anomaly: speed higher than possible for a ship; this way we also clean incorrect AIS data readings,
- location anomaly: a ship is found in another location than inferred from the previous course,
- triangle anomaly: a ship is traveling along the longer edges of a triangle instead of the shorter, e.g., making a zig-zag or traveling back and forth,

- angle anomaly: change of course over 90 degrees; we assume that a ship should not change course rapidly; if this is the case, it should be interpreted as loitering.

Unpredictable location anomaly. In this method the following algorithm is used:

(1) Take two preceding locations along with timestamps.
(2) Based on speed and timing predict the next location.
(3) If the real position is different from the one predicted, raise an issue.

In order not to raise too many warnings, we allow the deviation from the predicted position of 3 miles (the tolerance). We also do not try to predict if the time intervals between positions are longer than a specific amount of time (here 1 hour). Prediction is also not conducted at the beginning of the travel segment, when the necessary number of measurements is not yet available. Sample anomalies using the method are presented in Figure 6.8.

Figure 6.8. Trajectory of ship Amazonith (MMSI: 210688000) with unpredictable location anomalies

Source: Own work.

What can be concluded from the mentioned figure is that anomalies have the tendency to focus around certain areas. These are the regions where heavier

marine traffic can be expected, for example in the English Channel and around Portuguese ports. Another explanation is that more anomalies occur around destination ports. This can be connected with waiting for the permission to enter the port.

Sharp change of course. In this case the following heuristics is used: if the ship changes the course more than 90 degrees, as measured between three consecutive messages, then the issue is raised.

Thanks to the proposed method it was possible to discover quite interesting angle anomalies. For example, one vessel, while waiting for the entry to the port, was traveling in circles (see Figure 6.9). More rational behaviour would be rather to stop on the high sea, so it was another reason why such an example should be treated as a loitering anomaly.

Some of the discovered anomalies seemed to be false positives and required a more detailed analysis. For example, in some cases anomalies were discovered on seemingly straight course trajectories (see Figure 6.10 left). In these particular cases the turn was almost 180 degrees. However, what was peculiar, it always occurred two times in a row. Later on, we identified the source of the problem: messages

Figure 6.9. Angle anomaly—a vessel traveling in small circles

Source: Own work.

were received by various AIS devices, which had unsynchronized clocks. Thus, the problem resulted from the incorrect ordering of the points forming the trajectory. Fortunately, the device responsible for incorrect timestamps was responsible for less than 0.5% of messages. Another source of false positives can be manoeuvres close to ports (see Figure 6.10 right).

Figure 6.10. Trajectories with marked angle anomalies. Left: anomalies on straight trajectories. Right: false positives around ports

Source: Own work.

Travel-time anomaly. We also proposed a method to discover loitering by looking at the longer segments of ships trajectories, not only at single messages. We tried to estimate the typical travel time between certain areas. Loitering would be discovered when a non-typical travel time was detected. More specifically, it would happen when a vessel was not following a normal or historical route: different times of travel when compared to its own historical routes or routes of a similar ships (type, size, cargo).

For the detection of this type of loitering we needed typical travel times between trajectory segments. In our database, after execution of previous algorithms, we had already had trajectory segments, i.e., parts that have the same navigational status and contain consecutive locations of a ship. It was then possible to measure the travel time and distance between the starts and ends of many segments. If normally the travel takes 12 days and we observe 17 days, then the whole track can

be marked as an anomaly. Similarly, when the typical distance is 650 miles and the ships travels 850 miles then it is also a case for loitering.

Unfortunately, our hypothesis that it is possible to look for anomalies based on travel segments, i.e., longer tracks, could not be verified directly. It is noteworthy that the majority of segments started and ended in the same sector, which is a weak foundation for detecting cases of loitering. In fact, detection of such anomalies would require more a sophisticated approach for joining and combining shorter segments into meaningful multisegments that would allow one to compare travel times and distances. Given the expected computational complexity, this direction of research is foreseen for future work.

For this method we prepared a very interesting visualization, namely a count of starting and ending points of segments in each sector of the world (Figure 6.11). It perfectly reflects the shape of coastlines and the location of the most visited ports. In the visualization small dots are also visible—they represent the average location based on segment start and end coordinates respectively.

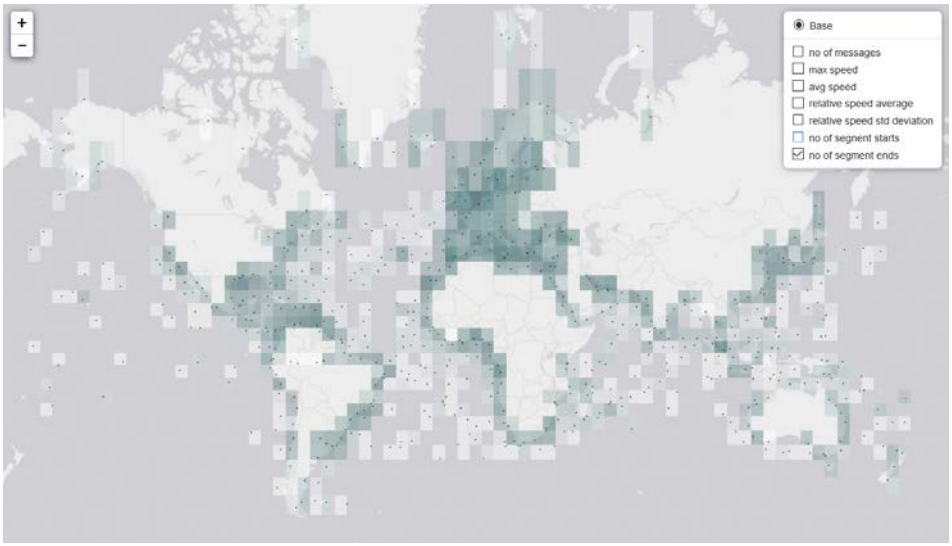
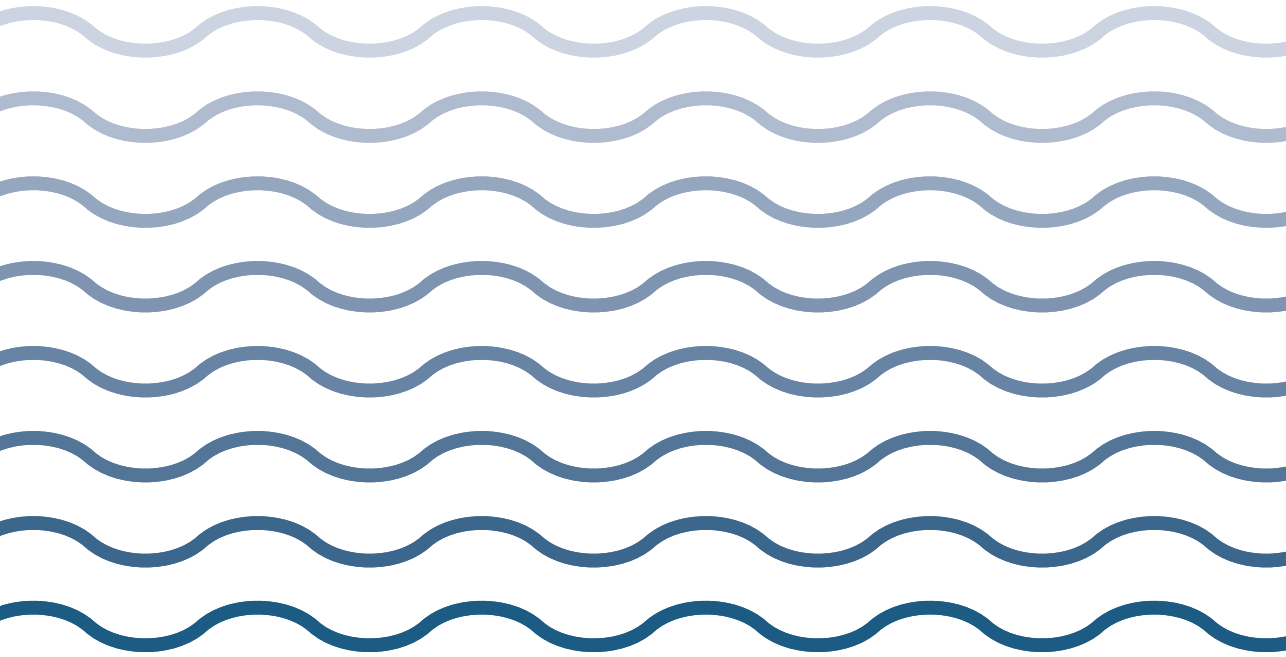


Figure 6.11. Number of segments ending in the given sector

Source: Own work.

Summing up, loitering detection is just the first example of anomalies detection methods developed in the SIMMO system. The rest of the methods, which focus on other types of anomalies, are presented in Chapter 9, since for their development we applied big data technologies. In Chapter 9 the updated statistics on the loitering methods, but this time calculated using the state-of-the-art analytics approach, are also presented.

Chapter 7



7. SHORT-TERM MARITIME RELIABILITY AND RISK ASSESSMENT

The chapter presents the assumptions and a concept of a method for a short--term assessment of maritime reliability and risk (MRRAM). The MRRAM method consists of three classifiers that include different variables that may influence the reliability of delivery being realized by a given ship. These classifiers are: ship-related, voyage-related, and history-related. For each classifier, its risk variables are discussed, with a justification why they are significant for reliability assessment. Finally, a proposed approach to the estimation of overall reliability and risk measure is described followed by presentation of results of analyses, which were conducted to evaluate the method.

7.1. Outline of the method

The short-term Maritime Risk and Reliability Assessment Method (MRRAM) calculates the reliability and risk of a given delivery (voyage) being realized by a given ship. As it was indicated in Section 2.3.1, reliability of a maritime transport service is one of the main factors that influence its quality. Thus, MRRAM contributes to the determination of this quality. Moreover, because MRRAM provides estimations related to a given voyage, it might be said that it focuses on a short-term horizon.

In the research we focus only on transport services being performed by merchant cargo vessels. This group includes ships that transport cargo for hire, such as general cargo, tankers, bulk carriers, and containers.

According to MRRAM, the reliability is determined at the beginning of a given ship's voyage and takes into account ship-related and historical risk variables (voyage-independent variables) as well as more dynamic characteristics of a ship and its operational environment that are known at the beginning of the voyage (voyage-dependent).

The result of MRRAM indicates both reliability of the transport service (reliability of a delivery) as well as the risk related to this service. This results from the inverse relationship between reliability and risk, which we define as:

> Maritime risk is the probability of occurrence of an undesirable event that in turn may negatively influence the reliability of a maritime transport service. The higher the risk, the lower the reliability of the service.

In MMRAM this undesirable event is a delay of a ship. As it was defined in Section 2.3.2, reliability is referred to the problem of providing delivery of ordered products in a timely and uninterrupted way. Thus, it highly depends on punctuality and travel time. The undesirable event that reflects these two elements is a delay of a ship.

Summarizing, we can say that by estimating the risk of delay of a given ship on a given route, we receive information about the ship's punctuality and travel time, and thus we are able to assess the reliability of the delivery being carried out by this ship.

In general, MRRAM is a novel approach for the short-term assessment of reliability of a maritime transport service, being realized by a given ship, taking into account both static (voyage-independent) and dynamic (voyage-dependent) characteristics of a ship and attributes of its operational environment that can be retrieved from the available maritime data.

The "individual" approach of MRRAM means that reliability is estimated separately for each ship based on their individual features. Some of these features, however, may change in time (voyage-dependent). Therefore, it should be updated (if needed) for each new voyage.

Moreover, MRRAM calculates the reliability and risk for a given voyage. In comparison to the existing methods (which calculate the risk for longer time periods) the risk horizon of MRRAM may be seen as short-term (a single voyage). However, MRRAM takes into account only these variables that are known at the beginning of the voyage. We assume here that they are stable during the whole voyage. Thus, the results provided by MRRAM concern a short-term horizon (a given voyage) and are not updated during the voyage. A more dynamic approach, which assumes that results might be dynamically updated during the ship's voyage, is provided by the method for a ship's punctuality prediction (presented in Chapter 8).

In the context of maritime risk management, MMRAM relates also to the Formal Safety Assessment (FSA) methodology (see Section 3.2.1). From the five steps of FSA, MRRAM addresses the first two:

(1) Hazard identification: the results of this step is the risk variables typology (see Section 3.3) which was used to develop risk classifiers presented further in this chapter.
(2) Risk assessment: the design of a risk model for determination of a risk value for a given ship and for a given voyage.

A result of MRRAM is a measure (Reliability) that, with a certain confidence level, indicates the level of reliability of the transport service being carried out by a particular ship. The measure provides a probability that the service will be completed successfully—punctually (without a delay) and within a standard travel time. Knowing Reliability, the level of risk that the transport service will not be realized on time can be calculated as $1 - \text{Reliability}$.

7.2. Risk classifiers and variables

As indicated in Chapter 3, the reliability of a transport service can be determined taking into account various variables (see the typology of risk variables presented in Section 3.3). These variables may be connected with the ship and its characteristics (e.g., size, age), with the voyage (e.g., destination port, planned route), or may relate to the operational environment of the ship (e.g., geopolitical variables, weather).

One of the difficulties in the design of a reliability and risk assessment method is a large number of input variables that might be considered. For this reason, a hierarchical architecture consisting of three reliability and risk classifiers, which further depend on other variables (*V*s), is proposed (see Figure 7.1). The three classifiers gather thematic-related variables:

- Ship-related variables: actual information that relates to the ship and its characteristics.
- History-related variables: information that relates to the past behavior and characteristics of the ship, but which may pose a potential risk at the moment of the analysis and may influence the level of the service reliability.
- Voyage-related variables: information that relates to the given voyage.

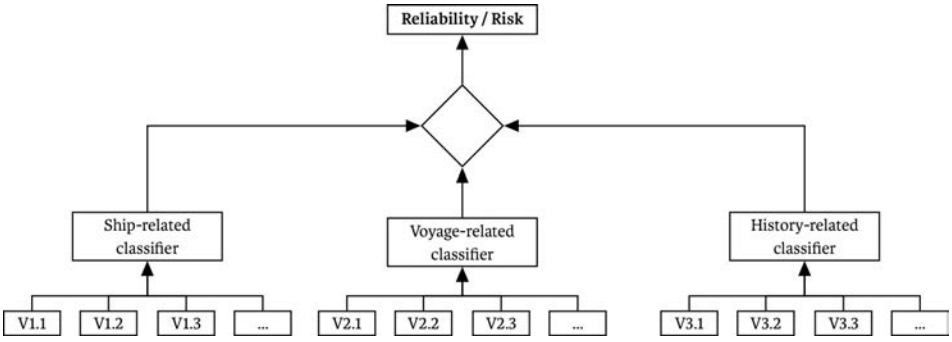


Figure 7.1. The MRRAM variables

Source: Own work.

Altogether, the presented classifiers determine the level of risk for a given voyage and indicate an overall reliability of the transport service.

In the proposed method, risk variables (depicted as *V*s), which are considered by respective classifiers, are taken from the typology of risk variables presented in Section 3.3. Depending on the classifier, different risk categories could be included. However, for MRRAM only three categories from the typology were included

and for each category a set of variables was selected. They are summarized in Table 7.1. This selection results from the availability of data which are required to determine the value of a given classifier. Here we selected only those variables that may be determined using data sources available for this research (see Section 4.5). Nevertheless, if other data sources are available, the list of variables for each classifier may be extended.

Table 7.1. Risk variables

Name	Description	Variables
Ship-related	static attributes of a ship, which do not change during a given voyage	size, flag, age, type, classification society, classification status
Voyage-related	variables specific for a given voyage, which may change from voyage to voyage	travel time, congestion, hazards on the route and geopolitical risk, type of cargo, weather, predicted delay
History-related	historical information about a ship and detected past anomalies	past delays, detentions and bans, accidents, cargo loss, pollution events, visited ports and other anomalies

Source: Own work.

It should be emphasized that the three reliability classifiers include both the variables that, in other research, have been indicated as important in determining the quality of a transport service (see Sections 2.3.1 and 2.3.2) and a set of new variables that have not yet been analyzed together, like the characteristics of a ship or voyage. The former variables include especially:

(1) Travel time: the amount of time needed to transport cargo from port A to port B; this variable may take into account, e.g., an average speed on the predicted route, weather forecasts or historical information, like past travel time of the ship on this route or with a given crew.
(2) Punctuality/Delays: information whether a ship is able to complete a delivery before or at a previously designated time, where designated time is the ETA provided by a captain at the start of the voyage; these variables depend mainly on the predicted travel time as well as historical statistics regarding the punctuality of the ship in the past.
(3) Completeness: information whether during a shipment damage to cargo may occur; this may take into account, e.g., past statistics on cargo damage / incompleteness for a given ship, or past accidents with registered loss of cargo.

It is assumed that there are certain relationships between various risk variables (dependability), as included in the typology. These relationships have to be taken into account when determining the value of a given classifier. Therefore, while

designing MRRAM it was decided that for the classifiers the concept of Bayesian Networks (BN)[1] should be incorporated to model the relationships between the variables and estimate the risk.

Moreover, because the actors operating in the maritime domain might have different missions, then different risk variables can have different values for them. Therefore, MRRAM assumes differentiation of the critical variables which may influence the overall level of risk/reliability more heavily. This is done by assigning weights to the classifiers in the calculation of the final reliability measure. The weights might be adjusted depending on who is the receiver of the MRRAM results. This approach makes it possible to take into account different context and business scenarios in which the method can be used (Stróżyna & Abramowicz, 2015).

The next characteristic of MRRAM is the confidence measure of the provided results, which depends on how much information is available at the start of the ship's voyage. In reality, some variables, and thus the reliability variable, may not be known or available. Therefore, MRRAM assumes that we do not need to know the values of all risk variables in order to calculate reliability, some information might be missing. However, the fewer input values are provided, the lower the confidence of the results.

We can present this characteristic based on a simple example. Let us assume that we know only the basic characteristics of a ship (e.g., type and age) and some voyage-related information, like the departure and destination ports, type of cargo to be transported and the ETA provided by the captain. Moreover, we have past and actual AIS data available that allow us to see the past routes of this ship as well as those of other ships. Despite the fact that other variables are not available, we can use MRRAM to determine, with a given level of confidence, the reliability of this transport service taking into account only the available data. But due to the fact that some information is missing (e.g., information about the weather or past detentions) we would say that the calculated reliability of the transport service has 50% confidence level. However, if we had a little bit more information, for example the weather conditions on the predicted route, it would be possible to provide more input to MRRAM and thus calculate the classifiers with a higher confidence. More examples of utilization of MRRAM, based on real data, are presented in Section 7.3.

As indicated above, MRRAM consists of three classifiers: ship-related, voyage--related, and history-related.

The ship-related classifier depends on the ship's features. It can include such variables as: size, age, owner (known/unknown, owner on a list of poor performing companies), flag, classification status, crew size and experience, etc. Some of these variables can change from time to time (e.g., owner, classification status), while others are rather the same for the ship's lifetime (e.g., size, type).

The voyage-related classifier is associated with a specified ship's voyage from

1. The concept and justification to use BN was presented in Section 3.2.3 and in (Stróżyna, 2017a).

port to port. It may take into account travel time, delays on the route, congestion on the route, characteristics and state of the departure and destination ports (congestion, history of accidents, and geopolitical hazards like political unrest, corruption, civil disorders, terrorism, crimes, etc.), type of cargo carried (dangerous or harmful substances), and weather (e.g., predicted extreme weather conditions). In general, it includes variables whose values are unique for a given voyage.

The history-related risk variables are connected with historical information about the ship. Here, such variables can be included as past anomalies in ships behavior (missing elements in AIS messages, sudden changes in a ship's name, type, or identification number, ambiguous or invalid identification of a ship, loitering at high sea, sailing through protected areas), history of accidents (including casualties, pollution, cargo loss), port state controls and detentions in ports, history of visited ports, past classification statuses.

The risk variables for each classifier are assumed to take a binary or a categorical value (with a finite list of possible inputs). Moreover, for each classifier, the key issue is to determine the parametric relationship between the different risk variables and thus connect together the knowledge about the ship and its voyage. In the proposed method, it is done by incorporation of Bayesian Networks (BN). This approach allows for an inclusion of both deterministic and probabilistic information. Moreover, it enables to model the structure of variables which influence a given classifier and based on that estimate the conditional probability. For each classifier a separate BN has been constructed.

Depending on how many input values for a given BN are available, the confidence measure of a given classifier is estimated as a ratio of the number of input variables provided to the total number of variables included in the BN. Thus, depending on how much information is provided as an input, the confidence of each classifier may vary.

Using the three classifiers, the overall reliability (risk) measure for a given ship can be calculated. Moreover, the method assumes that each classifier is assigned a different weight. The weight reflects the significance of a given classifier (a classifier's relative importance) in a given context or scenario. This results from the fact that for each context (actor), different variables can be critical. The weights may be adjusted depending on the scenario/context.

As a result, the overall reliability measure in MRRAM is calculated as follows:

$$Reliability = 1 - RiskOfDelay \tag{7.1}$$

$$RiskOfDelay = \sum_{i=1}^{3} w_i \times \underset{(C)}{f_i} \tag{7.2}$$

$$w_i \geq 0 \tag{7.3}$$

$$\sum_{i=1}^{3} w_i = 1 \tag{7.4}$$

$$C = \frac{n}{N} \tag{7.5}$$

where $\sum_{i=1}^{3} w_i \times \underset{(C)}{f_i}$ is the probability of a delay of a given ship in a given voyage; f_i is the value of a given risk variable; w_i is the weight assigned to the classifier i; C is the confidence measure that relates to the ratio of available information in the classifier; n number of inputs (risk variables) provided; N total number of risk variables included in the classifier.

Knowing the value of the overall risk index, it is possible to rank the ships according the risk they pose (the reliability of supply). The ranking can be then used to compare different ships. The ships with level of risk above a defined threshold may be treated as particularly dangerous and more attention may be paid to them. Similarly to the weighted calculation of the overall risk level, also here the value of the threshold may be adjusted depending on the context and entity.

Taking into account the availability of data in this research, we selected risk variables that are used in MRRAM for determining particular risk classifiers. These variables are presented in Figure 7.2 and are described in the next sections. However, it is possible to adjust the list of the used variables, depending on the availability of data as well as context in which the method is going to be used. Thus, new variables may be added, while others can be discarded. However, inclusion of a new variable would require re-training the classifier to which the variable would be added.

7.2.1. Ship-related classifier

In this section selected risk variables for the ship-related classifier are described. These variables are a ship's: Age, Size, Type, Flag, ClassificationSociety, and ClassificationStatus. In general, these variables do not change, or change relatively rarely (e.g., from time to time a ship can change the classification society or classification status and once a year its age increases). Each of these variables is characterized in the following paragraphs.

Age. It is fairly reasonable to presume that the age of a ship is a variable that should be considered for the assessment of the maritime risk and the reliability of a transport service. Age in MRRAM reflects the number of years since the ship was built.

According to Nivolianitou, Koromila, and Giannakopoulos (2016), based on the age ships can be classified as:

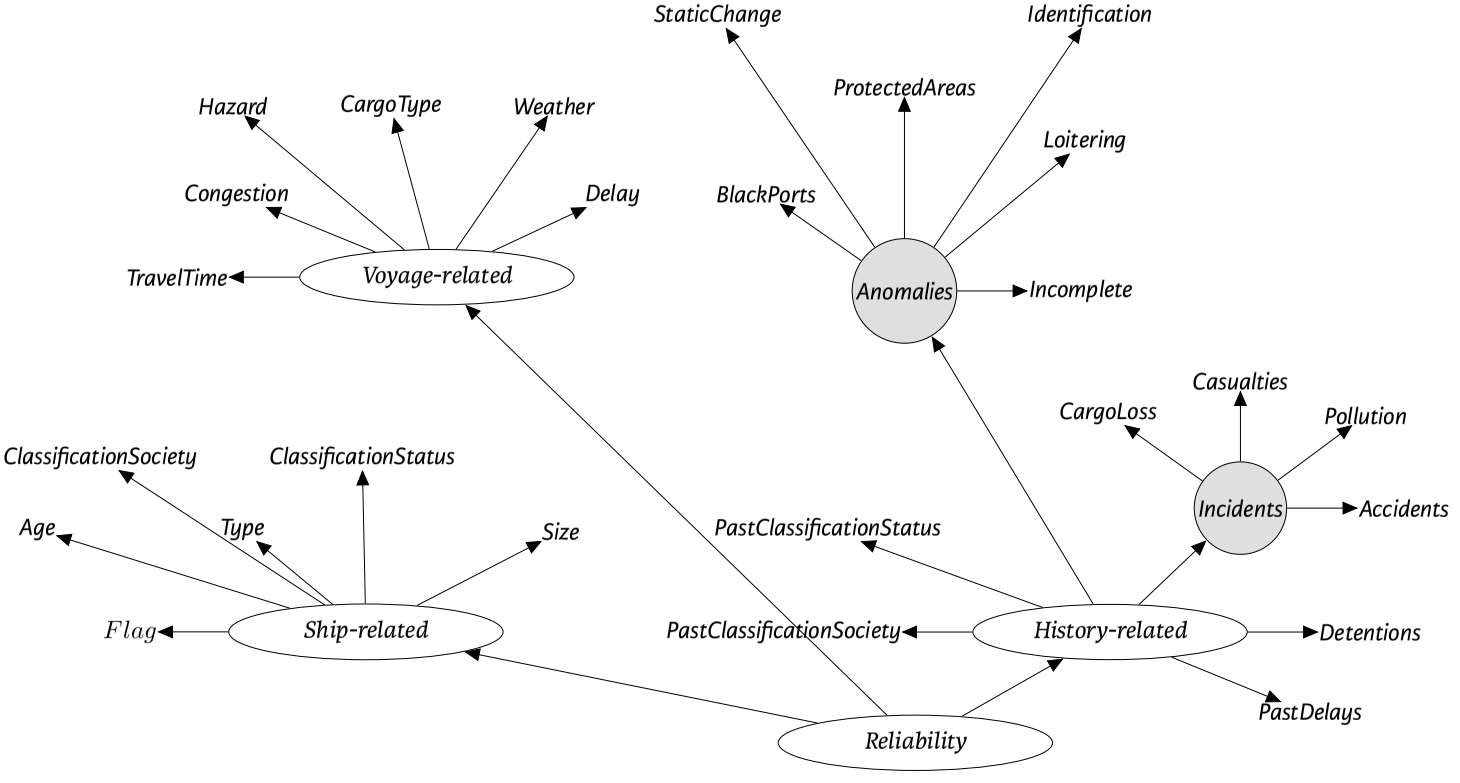


Figure 7.2. MRRAM graph

Source: Own work.

- New: 0 to 5 years.
- Middle age: 6 to 25 years.
- Old: 26+ years.

In general, new ships would perform better than the old ones. An aging ship is one of the risk variables since the duration of the ship's usage influences the overall operational reliability of the ship. Along with age, a hull's structural strength declines, due to corrosion and physical damage sustained during cargo operations, and the ability to resist waves and the intact rate drop. According to the accident statistic (Lam, 2012), vessels aged 15 years or above account for 86% of total loss in ship accidents, meaning that the age has a great impact on the stability and strength of the vessel.

However, the operational reliability of a ship may be improved due to technological upgrades made by the vessel owner. Such inclusion of technological innovations increases the performance and prolongs the lifetime of the vessel, and thus counteracts the ageing effect.

Size. The ship's size depends on its gross tonnage. Size may influence the risk of a maritime accident, i.e., it may be more frequent for big ships to have an accident due to their limited maneuverability and ability to speed up or slow down. Besides, not all ports or canals are adapted to deal with the biggest ships.

According to Equasis (2013), ships may be grouped into four size categories:

(1) Small ships: 100 to 499 GT.
(2) Medium ships: 500 GT to 24999 GT.
(3) Large ships: 25000 GT to 59999 GT.
(4) Very large ships: above 60000GT.

The small category starts with 100 GT because it reflects the main tonnage threshold for merchant ships to comply with the SOLAS Convention.

Type. The type of a ship is another variable that potentially may influence risk and reliability. For example, certain types might be riskier than others because they carry a specific type of cargo that is dangerous by nature (e.g., chemicals, oil, or gas).

The survey among the maritime experts gave us suggestions on, which types of merchant ships may be perceived as especially dangerous. According to the provided answers, these are container ships and tankers, including LNG/LPG tankers and chemical tankers.

Flag. A ship must be registered in the registry of the country whose flag it is flying. However, during the lifespan of a ship the flag may change. As it was already explained in Section 2.2, nowadays a real problem is 'Flag of convenience' (FOC),

which is a business practice of registering ships in a sovereign state, different from that of the ship's owners. In 2013, about 37% of ships were associated with a FOC state (Equasis, 2013).

Moreover, FOC allows shipowners to be legally anonymous, which hinders prosecution in civil and criminal actions. There are examples of FOC ships that have been found engaged in crime, offering substandard working conditions, and negatively impacting the environment. Therefore, the fact that a ship is from a FOC country may increase the risk it poses and may negatively influence the reliability of the supply. Besides, such ships are targeted for special enforcement by the countries they visit.

In general, flag states can be grouped into three categories: black (high risk), grey (middle risk), and white (low risk). The colors of flags are assigned and published by well-known maritime organizations, such as: the Paris MoU,[2] the Tokyo MoU,[3] and the US Coast Guard.[4] The colors of flags are determined based on the risk assessment that reflects the safety performance of ships registered to each flag state as measured by the number of port state inspections and detentions recorded over a three-year period.

ClassificationSociety. Classification society is a non-governmental organization that establishes and maintains technical standards for the construction and operation of ships; it also validates that a ship meets the security and safety standards and carries out regular surveys of ships to ensure compliance with the standard. If a ship is in compliance with the classification standards, the classification society issues a classification certificate.

SOLAS 74 convention (International Maritime Organisation, 1974) requires ships to be designed, constructed, and maintained in compliance with the requirements of a recognized classification society. A ship without 'class' can neither be insured nor mortgaged. It is also difficult to find a crew willing to sail on a ship that does not have a classification certificate. Moreover, nobody would risk giving cargo on such a ship and it would hardly have any value in charter or on the sale market. As a result, it might be said that a ship's classification is a must. However, there are still ships without a class since classification in most countries is still not a legal requirement. Such ships may be treated as potentially risky.

Today, there are more than 50 classification societies worldwide. Thirteen largest marine classification societies are members of the International Association of Classification Societies (IACS). But as it was stated before, each classification society may define their own standards. Thus, some societies may have stricter rules and standards than others. Also, the process of classifying a ship, the scope

2. https://www.parismou.org/inspections-risk/white-grey-and-black-list
3. http://www.tokyo-mou.org/inspections_detentions/NIR.php
4. https://www.uscg.mil/hq/cgcvc/cvc2/psc/security/flag_list.asp

of classification surveys, and assignment of a class may look differently for each society.

As a result, there may be classification societies with 'less strict requirements' and thus some ships may receive a classification certificate even if their 'quality' is not good enough in comparison to ships classified by other societies. In MRRAM, such classification societies are called as 'unreliable'.

ClassificationStatus. A classification status is designated by a classification society upon a classification survey. During the survey, a ship's structure, design, and safety standards are checked, including an inspection of engines, shipboard pumps, and other vital ship machines. Based on the survey, one of the following statuses is granted to a ship:

- *Delivered*—a vessel is in the class.
- *Suspended*—the class can be suspended when:
  - A ship does not operate in compliance with the rules of the classification society.
  - The owner of the ship fails to submit the vessel to a survey after defects or damages affecting the class have been found.
  - Repairs, alterations, or conversions affecting the class are carried out without requesting the attendance of a classification society.
  - A class renewal survey has not been completed before the deadline or within the time granted for the completion of the survey.
  - A ship is not entitled to retain its class due to reported defects.
- *Reinstated*—the class is reinstated upon satisfactory completion of an overdue survey or upon verification that due or overdue problems with a ship have been satisfactorily dealt with.
- *Withdrawn*—the class can be withdrawn:
  - At the request of the owner.
  - When the causes of class suspension have not been removed within a specified period of time.
  - When a ship is reported as a constructive total loss or scrapped.
- *Reassigned*—a ship can be reassigned to a class after suspension or withdrawal of class, if the required repairs have been carried out.

A class is designated only for a specified period of time (usually 5 years). Upon expiry of the class, a class renewal survey has to be performed in order for the ship to remain in the class.

From the point of view of risk analysis, special attention may be paid to the ships with statuses *withdrawn* and *suspended* as well as those which have not been classed by any classification society. Moreover, detection of past occurrences

of status withdrawals or suspensions may also suggest that a ship is suspicious because in the past there have been problems meeting the class standards, a similar situation may also be taking place right now or happen in the future.

ShipClassifier. Having the values of the above described variables, MRRAM calculates the value of the ship-related classifier. This value is calculated as the probability of delay in a given voyage (PossibleDelay), taking into account Flag, Age, Type, Size, ClassificationSociety, ClassificationStatus:

$$ShipClassifier = P(\text{PossibleDelay} \mid \text{Flag}, \text{Age}, \text{Type}, \text{Size}, \text{ClassificationStatus}, \text{ClassificationSociety}) \tag{7.6}$$

7.2.2. Voyage-related classifier

The voyage-related classifier includes variables that concern the current voyage of a ship and which may potentially influence the reliability of the transport service. The variables selected for this classifier include: TravelTime, Delay, Congestion, Hazard, CargoType, and Weather.

In the following paragraphs the variables are shortly characterized. However, TravelTime, Delay, Congestion, and Hazard are variables that are also part of the method for a ship's punctuality prediction (presented in detail in Chapter 8), and can be calculated using this method. Therefore, in this section only a general understanding of these variables is presented.

TravelTime. Travel time concerns how much time is needed to travel from the origin to the destination port. This time can be calculated based on information provided by the captain in AIS messages—Estimated Time of Arrival (ETA). In MRRAM, travel time is used to compare it with the average travel time on this route. The average time is calculated based on historical voyages of this ship on this route or voyages of other ships on this route. If the travel time provided by the captain is significantly higher or lower than the average, it means that the ship is planned to sail much faster or slower than in the past or than the other ships. Such a deviation of travel time is a risk—too fast or too slow sailing may be suspicious and potentially dangerous, and thus negatively influence the reliability of the service.

TravelTime is a binary variable that says whether, according to the ETA declared at the beginning of the voyage, the predicted travel time significantly deviates from the average (in this case TravelTime equals 1), or it falls into the tolerance limit (then TravelTime equals 0).

The tolerance limit is:

$$\text{Average(TravelTime)} \pm \text{StandardDeviation(TravelTime)}$$

Delay. Delay is a deviation between the ETA declared by the captain at the beginning of the voyage and the actual time of arrival to the destination port. In MRAM, Delay is understood as the average difference between the declared and actual time of arrival on this route (average delay). The average delay is calculated based on historical voyages of this ship on this route or voyages of other ships on this route.

In MRRAM, it is assumed that if the average delay for the route is higher than a defined threshold (tolerance limit), there is a risk of a delay also in this voyage. This, in turn, negatively influences the reliability of this transport service.

Delay is a binary variable that says whether the average time of the delay on a given route is higher than the defined tolerance limit (in this case, Delay equals 1), or it falls into the tolerance limit (then Delay equals 0). The tolerance limit is a numerical value that can be defined for each new calculation (e.g., for each voyage) and can be adjusted to suit a user's needs and the context.

Congestion. Congestion is a variable connected with the density of maritime traffic. It says weather the current density of ships in a given area is greater than the standard value. In case of MRRAM, the standard density is represented by the average density of ships. The average is calculated separately for each month (due to seasonality, explained further in Chapter 8) and for each maritime area.

Congestion occurs when the density in the last 24 hours is greater than the average. The greater the density of ships, the higher the potential risk of delay (e.g., due to queues in popular canals, or the risk of a collision in high density zones).

Congestion is a binary variable that says whether the average density on a given route is significantly higher than the monthly average (then *Congestion* amounts to 1), or not (then Congestion amounts to 0). The threshold in this case is calculated as:

$$\text{CongestionThreshold} = \text{Average(Density)} + \text{StandardDeviation(Density)}$$

Hazard. The security of a voyage also depends on other variables, like geopolitical risk or maritime accidents. In MRRAM, all these issues are grouped together in a single variable: Hazard. This variable says whether the route a ship is going to follow is potentially dangerous (e.g., due to the risk posed by countries visited on the way). It includes variables that reflect hazards that may happen on the route of a given ship and thus may potentially influence the reliability of delivery and the level of maritime risk for the ship in the voyage.

Hazard is a binary variable. Its value depends on the hazard index that is calculated for the predicted route of a ship. This index (described in detail in Chapter 8) includes three types of variables:

(1) Maritime Accidents: it takes into account the number of maritime accidents that have happened in a given area in the past.

(2) Piracy: it takes into account reported accidents of piracy and armed robberies that have happened in a given area in the past.
(3) Country Risk: it analyzes the risk of the departure and destination country of a given ship as well as the countries the ship is sailing through.[5]

If a ship is sailing through such dangerous areas, there is higher likelihood that an unforeseen and unfortunate event may happen. These variables are hard to quantify but need to be thought about.

Similarly to the previous variables, Hazard says whether the predicted route of a ship includes any area that is classified as dangerous (then Hazard is 1) or not (Hazard is 0). An area is perceived as dangerous if the calculated hazard index is higher than the defined tolerance limit. The tolerance limit is a numerical value that can be modified for each new calculation and adjusted to suit a user's needs and the context.

TransportedCargo. The cargo itself may also influence reliability. The first issue concerns the type of cargo—whether the transported goods are dangerous. In this case, special care should be taken to avoid an explosion and other serious problems. Moreover, shipping operations must obey the International Maritime Dangerous Goods Code (IMDG) published by IMO (International Maritime Ogranization, 2011).

The second issue is cargo stacking, which should obey the SOLAS Convention and have a cargo security certification approved by the authorities (International Maritime Organisation, 1974). An accident statistics show that, in recent years, a large number of maritime incidents were directly or indirectly caused due to stacking issues and movement of goods carried by a ship.

MRRAM takes into account only information whether the cargo transported by a ship is dangerous or not. This fact should be declared by the captain in AIS messages (an appropriate type of ship should be set).

TransportedCargo is a variable that says whether the type set in AIS messages at the beginning of the voyage is dangerous (TransportedCargo equals to 1) or not (TransportedCargo equals to 0).

The information only from AIS may be perceived as a simplification since just the type of ship may not always reflect that the cargo is dangerous. But in this research the author has no access to more detailed data about the transported cargo. However, if such data is available, this information might be easily included in MRRAM.

Weather. Shipping operations on the route and, thus, the supply reliability are also determined by environmental variables, here collectively referred to as Weather. This variable is an external risk from the operational environment of a ship and may concern current and forecast weather conditions (such as wind, fog, rain, snow,

5. A country means in this case Exclusive Economic Zone belonging to a given country.

clouds) and natural hazards (such as earthquakes, tsunami, and other natural disasters). Depending on the route and the season, this variable may influence the shipping risk more heavily.

Weather conditions are often a very unpredictable variable since they may change suddenly (for example, abrupt changes in the state of the sea may drastically influence visibility). It is observed that very often weather predictions, i.e., meteorological forecasts at sea, are wrong.All these aspects may affect travel time and thus influence supply reliability. However, according to Gaonkar et al. (2011) this influence is rather weak.

In case of MRRAM the Weather variable concerns only prediction of weather extremes, such as heavy rain/snow, wind or limited visibility due to dense fog on the planned route. This variable is a binary variable, where 1 means that some weather extremes are forecast on the planned route, while 0 means otherwise.

VoyageClassifier. Having the values of the above described variables, MRRAM calculates the value of the voyage-related classifier. This value is calculated as the probability of delay in a given voyage (PossibleDelay), taking into account TravelTime, Delay, Congestion, Hazard, CargoType, and Weather:

$$\begin{aligned}\text{VoyageClassifier} = P(&\text{PossibleDelay} \mid \text{TravelTime}, \text{Delay},\\ &\text{Congestion}, \text{Hazard}, \text{CargoType}, \text{Weather})\end{aligned} \tag{7.7}$$

7.2.3. History-related classifier

The history-related classifier concerns the overall past operational history of a ship and includes all incidents from the past which suggest that a‘ship is potentially dangerous and that the reliability of its service may suffer. The risk variables that were selected for this classifier are: PastDelays, Detentions, PastClassificationSociety, PastClassificationStatus, BlackPorts, Incomplete, Loitering, ProtectedAreas, StaticChanges, Identification, CargoLoss, Casualties, Pollution, Accidents. Some of them are grouped together:

- CargoLoss, Casualties, Pollution, Accidents are grouped in the variable Incidents.
- BlackPorts, Incomplete, Loitering, ProtectedAreas, StaticChanges, Identification are grouped in the variable Anomalies.

In the following paragraphs the above variables are characterized.

Detentions. A ship can be subject to Port State Control (PSC)—the inspection of foreign ships in other national ports for the purpose of verifying competency of the captain and officers on board, checking whether the condition of the ship and its equipment complies with the requirements of the international conventions,

and whether the vessel is manned and operated in compliance with applicable international laws. After a PSC, in the case of occurrence of any deficiencies that are clearly hazardous to safety or the environment, a ship can be detained. The ship is also called to take follow-up actions to rectify the deficiencies indicated during the control.

Ships that were detained can be also classified as 'under-performing' or 'banned.' These categories include ships that have been detained three or more times by maritime authorities during the last 12 or 24 months. Such lists are open and published on a regular basis (most often monthly) by the Port State Control Committee of MoUs. These ships are subjected to more frequent inspections at ports within the MoU region.

The detained or banned ships are perceived as very risky. Therefore, it is important, while conducting a risk assessment, to take into consideration the fact that a ship was detained or banned in the past. Moreover, it is also critical to include information about detentions and bans not only in a particular port or region, but also in other regions of the world, administered under other MoUs.

The Detentions variable includes information whether a given ship was ever detained in any port or classified as banned. If so, the variable takes the value of 1, and 0 otherwise.

PastClassificationSociety and PastClassificationStatus. The role of classification societies and classification statuses in assessment of a ship's risk has already been discussed in Section 7.2.1. Similarly to ClassificationSociety and ClassificationStatus, the variables PastClassificationSociety and PastClassificationStatus include, respectively, information whether in the past ship belonged to an 'unreliable' classification society (then PastClassificationSociety equals to 1) and if it ever had a potentially dangerous classification status, like 'suspended' or 'withdrawn' (then PastClassificationStatus is 1).

PastDelays. Similarly to the variable Delay described in Section 7.2.2, PastDelays includes information whether in the history of the ship there have been any delays noted for a given voyage. If so, PastDelays equals to 1, and 0 otherwise.

Incidents. Maritime incidents is a category that includes information about a ship's accidents in the past and further consequences of these accidents, such as casualties, pollution, or loss of cargo. The fact that a ship has a history of maritime accidents may significantly influence the level of risk and reliability of its service. Also, a casualty history or loss of cargo are other risk variables that might be taken into account.

Each variable of Incidents is a binary variable that takes the value of 1, if:

- Accidents: there were any accidents in the past reported for the ship.

- Casualties: there were casualties in the reported accident(s).
- Pollution: there was a pollution due to the reported accident(s).
- CargoLoss: there was a loss of cargo due to the reported accident(s).

Anomalies. Anomalies is a category that includes information about detected anomalies in ship's behavior which have happened in the past (not during the ship's current voyage). The anomalies may concern for example changes of the ship's identity or static characteristics, occurrences of loitering at high sea, events of ambiguous identification, visits in suspicious ports.

The information about past anomalies for a given ship should also be included in the risk assessment. We can assume that if in the past the ship has behaved in a risky manner or suspiciously, it may happen again in the future. Therefore this group of variables is included in the history-related classifier.

Anomalies that are considered in MRRAM encompasses six types of behavior: BlackPorts, Incomplete, Loitering, ProtectedAreas, StaticChanges, and Identification. The information about occurrences of these anomalies is calculated using the analytical methods developed within the SIMMO project, in which the author participated. Detailed information about these anomalies can be found in (Stróżyna, Małyszko, Węcel, Filipiak, & Abramowicz, 2016b) and (Węcel et al., 2016).

BlackPorts. The history of visited ports is another variable which may influence ta ship's risk. If a ship frequently visited ports known for criminal activities or potentially dangerous due to various reasons (collectively called as black-listed ports), then this fact may suggest potential involvement of the ship in such illegal actions and thus influence its reliability.

Similarly, if a ship during its voyage called a port that was not declared as its destination port (e.g., the ship declared sailing to a European port but on its way it stopped at Roatan port (Honduras), known for a high crime rate, and then continued its travel to Europe) such an event may also be considered as suspicious and should be included in risk assessment.

If in the history of a ship there are any visits in such suspicious ports, BlackPorts equals 1, and 0 otherwise.

Incomplete. As it was mentioned in Section 4.1, AIS is a cooperative system and as a result sometimes (on purpose or not) vessels transmit incomplete AIS messages which do not include all required information. According to the SOLAS convention (International Maritime Organisation, 1974), ships shall maintain the AIS transponder turned on and should provide all required information, such as the ship's identity, type, position, course, speed, navigational status, and other safety-related information. In a situation when all or some parts of this required information is not provided, it should be detected. Such events mean that the vessel does not obey the IMO regulation, which may negatively influence the safety

and security of other ships at sea. This may also indicate lack of collaboration of the ship captain.

Information about instances of sending incomplete AIS messages is reflected in the Incomplete variable, which takes the value of 1 if such events were detected in several subsequent messages in the history of a given ship.

StaticChanges. AIS data include both static and dynamic information about a ship. While the dynamic data (position, speed, course) should be updated on a regular basis, the static part contains some information that in general should not be changed at all, such as IMO number, MMSI number, ship's name, call sign, type, dimensions. A ship should provide the same value for these static attributes all the time. In rare situations this information can in fact change, e.g., a tanker changed the owner who changed its name, and as a result the transmitted name of a ship changed.

However, there are ships that during their voyage suddenly change one of the above-listed elements and start transmitting a new value. For example, a ship may suddenly change its type from 'Fishing vessel' to 'Tanker.' Such behavior may indicate that the ship has some unclear intentions and poses a threat to other nearby ships.

Occurrences of such anomalies should also be included in calculations of a ship's reliability. In MRRAM, it is included in the StaticChanges variable that takes the value of 1 if such events were detected in the history of the ship (and 0 otherwise).

Identification. Ships, in general, can be identified using two numbers:

- IMO (International Maritime Organization): a unique reference for ships that was introduced under the SOLAS convention and remains linked to the hull for a ship's lifetime, regardless of a change in name, flag, or owner. It is assigned to all merchant ships above 100 Gross Tonnage. The IMO number consists of three letters 'IMO' and a seven-digit number, including a six-digit sequential unique number followed by a check digit. The check digit is used to verify the integrity of the IMO number.
- MMSI (Maritime Mobile Service Identity): a unique identifier of ships (or other entities, such as coast stations, ship stations, group calls, etc.), used to identify ships in AIS. The MMSI number consists of nine digits, where the first three are Maritime Identification Digits (MID) ranging from 201 to 775, denoting the administration (country) or geographical area of the administration responsible for a ship. The list of MIDs assigned to each country is published by ITU.[6]

Both identifiers are included in AIS messages. However, they may be manipulated by a ship master and there are ships that transmit wrong IMO or MMSI. Such

6. http://www.itu.int/online/mms/glad/cga_mids.sh?lang=en

events are maritime anomalies and in general should not happen. Similarly to changes in elements of static AIS messages, events of ambiguous identification should also be included in the calculation of a ship's risk.

In MRRAM, such events are reflected in the Identification variable that include the following situations:

- Implausible MMSI: hip transmits too short MMSI or MMSI does not start with the digits 201–775.
- Implausible IMO: ship transmits too short IMO or verification of the IMO number fails.
- Duplicated MMSI / IMO: there is another ship that transmits the same MMSI or IMO.

If the history of a ship includes events of ambiguous identification of the ship, Identifiaction equals to 1 (and equals to 0 otherwise).

Loitering. Loitering occurs when a ship being at high sea starts sailing with an unreasonably low speed, e.g., the speed over ground (SOG) reported by the AIS message is below 5 knots. Such events are rather unusual and may mean that this is due to some technical problems. This might be proved by checking the ship's navigation status (e.g., the ship transmits the "not under command" status). However, an unreasonably low speed may also suggest suspicious behavior, especially when at the same time liaison with another vessel occurs.

Another indication of loitering is when the average speed during the past 12 hours is below a certain threshold, even though a ship was sailing at high sea, where normally ships sail with a maximum or high speed.

The third indication of loitering may result from a comparison of a 'standard' behavior. The 'standard' behavior may be calculated as average speed profiles in specific geographical regions. Such a profile (average) may be determined after analysis of speed of all ships which have passed through a given region. Then, any SOG that significantly deviates from the average may be assessed as loitering.

Another type of a ship's anomaly is a route anomaly—an unpredictable movement, i.e., not following the trend or pattern. The most typical examples are:

- Average speed anomaly: speed higher than possible for a ship.
- Location anomaly: a ship is found in another location than inferred from the previous course.
- Triangle anomaly: a ship is traveling along the longer edges of a triangle instead of the shorter, i.e., making a zig-zag or traveling back and forth.
- Angle anomaly: change of course over 90 degrees (under assumption that a ship should not change course rapidly).

All the above mentioned speed- and course-related anomalies, collectively called Loitering, may be treated as a potential threat. Therefore, they are included

as another risk variable in MRRAM. Similarly to other anomalies, Loitering equals to 1 if any occurrence of speed- or course-related anomalies were detected for a given ship.

ProtectedAreas. Maritime Protected Areas (MPA) are areas that restrict human activity for a conservation purpose, typically to protect natural or cultural resources. MPAs can be established by local, regional, national, or international authorities. Depending on the authority, different limitations on development, fishing activities, moorings, and bans on disrupting maritime life may be in force. Among MPAs there are marine reserves where human impact is kept to a minimum.

Having in mind the goal of establishing MPAs, maritime traffic through these areas should be limited. It especially concerns ships that transport dangerous substances, which in the case of an accident might cause irretrievable damage the maritime ecosystem. Therefore, it is important to report events when merchant ships are traveling through MPAs.

In MRRAM, such events are reflected in the ProtectedAreas variable, which says whether a ship has traveled via an MPA (then the variable is equal to 1).

HistoryClassifier. Having the values of the above described variables, we can determine the value of the history-related classifier. Its value is calculated as the probability of delay in a given voyage (PossibleDelay), taking into account PastDelays, Detentions, PastClassificationSociety, PastClassificationStatus, BlackPorts, Incomplete, Loitering, ProtectedAreas, StaticChanges, Identification, CargoLoss, Casualties, Pollution, Accidents:

$$\begin{aligned}\text{HistoryClassifier} = P(&\text{PossibleDelay} \mid \text{PastDelays}, \text{Detentions}, \text{BlackPorts},\\ &\text{PastClassificationSociety}, \text{PastClassificationStatus},\\ &\text{Incomplete}, \text{Loitering}, \text{ProtectedAreas}, \text{StaticChanges},\\ &\text{Identification}, \text{CargoLoss}, \text{Casualties}, \text{Pollution}, \text{Accidents})\end{aligned}$$

The next section presents examples how the values of the MRRAM classifiers are determined (training and testing Bayesian Networks for each classifiers) as well as calculation of the final Reliability measure provided by MRRAM based on real maritime data and real examples of ships voyages.

7.3. Application of the MMRAM method—an example

In order to perform the evaluation of the MRRAM method and show its applicability, it was implemented and tested using real maritime data and for selected examples of real ships' voyages. The process started with selection of real world examples of past voyages for the 25 different European ports. In total, a set of 255 voyages was collected. It was further divided into a training set (consisting of 229 voyages), and a testing set (26 voyages). Then, for each voyage the values of the variables of the three MRRAM classifiers were collected and cleansed. Here, the real AIS data, the results of the SPP method (see Chapter 8) as well as data collected from different Internet sources were used. Altogether, for each voyage 27 different risk variables were defined.

In the next step the MRRAM method was implemented. For the purpose of this research, the method was implemented using R, which offers a number of helpful packages that allows to conduct, inter alia, a development of Bayesian Networks, statistical analysis and validation of results and models.

Having collected the data for the variables and implemented the method, in the next step the static, voyage and history classifiers were calculated. To this end, a model for BN for each classifier was trained and validated using the voyages from the training set. Then, the received models were used to estimate the reliability for the testing set voyages.

In the final step, the results provided by each of the classifiers were joined to provide an estimation of the overall reliability measure, according to formula (7.1). Here, exemplary weights for each classifier were defined. As a result, predictions of the reliability for the testing set voyages were calculated. Below the results of this process are described.

7.3.1. Data sources and infrastructure

In the presented study of the MRRAM method, selected data sources from Section 4.5 were used, namely AIS data, covering a one year period (January–December 2015) and a global scale, data about ships and their characteristics, data about detentions, inspections, classification of ships, data about risk indexes, ship's accidents, piracy and terrorist attacks, GIS data, and selected Copernicus data about weather extremes. Taking into account the amount of data available, the conducted analysis was very data-intensive. To deal with this Big Data challenge appropriate infrastructure in a cloud had to be set up and used. Cloud services make available the necessary infrastructure, which is scalable (can be expanded dynamically when required) and ensures that the desired analysis can be carried out more efficiently, with multiple, simultaneous access (Sauer & Norkus, 2015).

To this end, services and resources offered by the Microsoft Azure platform were used.

First of all, the Azure Storage services were used to store the big amount of AIS data required for the analysis in a form which enables their fast and efficient analysis. Both AIS as well as additional data from the Internet were loaded to Azure and converted to the Apache Parquet format. The format provides efficient data compression and encoding schemes with an enhanced performance to handle complex data. This process required developing an appropriate data loading and preprocessing pipeline using Azure Data Factory. In total about 120 GB of data were loaded, processed and stored.

Then, to test and evaluate the method also analytics solutions, which provide predictive analytics, machine learning, and statistical modeling for Big Data, had to be used. MRRAM was implemented using the open source R language. In this context, the Azure HDInsight service was used, which is a cloud Hadoop solution that provides open source analytic clusters for Spark and integration with R Server. Thus, it makes it possible to run the developed R scripts but with the advantage of a large parallel analytics. This significantly speeds up the computation time and the process of getting the results.

To do the analytics, the following HDInsight cluster was set up:

- 1 R server edge node (16 cores, 112 GB RAM, 32 disks, 800 GB SSD),
- 2 head nodes (8 cores, 56 GB RAM, 16 disks, 400 GB SSD each),
- 3 worker nodes (8 cores, 56 GB RAM, 16 disks, 400 GB SSD each).

7.3.2. Analysis results

In the first step of the MRRAM evaluation a set of 255 real world examples of ship voyages was selected.[7] The sample was then divided into a training and testing set. The training set consisted of 229 voyages for which the real (actual) reliability of the voyages could be determined using the results provided by the SPP method. To this end, the information about the actual delay of a ship on arrival at a destination port in a given voyage was taken into account. The real delay was provided by the SPP method (Chapter 8) and it was the difference between the ETA declared by the captain and the actual arrival time at a port. For the purpose of this research, we assumed that if the actual delay was greater that the absolute value of 3 hours (meaning that the ship arrived 3 hours before or after the declared ETA), then the actual reliability equaled 0 (meaning a delay occurred). Otherwise it equaled 1, meaning the ship arrived more or less on time. The values of the actual reliability were then used to train the MRRAM classifiers.

7. These 255 examples are the results of the route prediction method presented further in Chapter 8.

The testing set consisted of 26 voyages. These examples were used to test and validate the three risk classifiers that were trained using the training sample and the whole MRRAM. For these voyages, the actual reliability was also determined in the same way as for the examples from the training set. However, in this case this information was used to validate the results provided by MRRAM for the voyages from the testing set.

After preparing the training and testing sets, the values of variables for the three classifiers had to be collected or determined. Here, the variables defined in Sections 7.2.1, 7.2.2, and 7.2.3 were used (see Figure 7.2). In total, it was 27 variables.

Having collected the required data, each MRRAM classifier was trained. This means that for each classifier a Bayesian Network was created, and then, the conditional probabilities in the BN were determined using the parameter learning method with the Bayesian estimators, available in the R package "bnlearn". The model was fitted using the examples from the training set and the data collected for the variables. Then, a validation of the quality of each Bayesian classifier was conducted using the method for k-fold cross-validation for Bayesian Networks from the R "bnlearn" package.

In general, the BN developed for each classifier allows for determination the probability of a delay for a given ship and on a given voyage, taking into account the variables of a given classifier (either the ship-, the voyage-, or the history-related information).

We assume here that the probability of a delay corresponds with the reliability of the transport service (in other words, the risk that the transport service/the delivery by a given ship might not be realized on time). In general, the higher the probability (risk) of the delay provided by the classifier/MRRAM, the lower the reliability of the service.

Then, the three risk classifiers were tested using the testing set. The received results confirmed that the MRRAM classifiers are effective tools to estimate the probability of a delay and the short-term reliability of a transport service (a supply) (see Table 7.2). All classifiers are characterized by good accuracy (80.77%, 65.38%, and 84.62% accordingly, for the ship-related, the voyage-related, and the history-related classifier), and good precision (86.96%, 93.33%, and 87.50% accordingly). Thus, it might be concluded that they are able to provide accurate estimations and, as a result, might be used in the prediction of a ship's risk and reliability, taking into account the assumed risk variables (i.e., the variables that relate to the ship, its voyage, and its history).

In the case of the voyage-related classifier, the accuracy and precision of the results is the lowest. It might result from the fact, that the weather conditions and sea state might be an important variable here (see Section 7.2.2). Thus, the quality of the voyage-related classifier could be improved by taking into consideration more information or more precise information about the weather conditions and

Table 7.2. Results of the cross-validation for the risk classifiers

Classifier	Accuracy	Precision	Sensitivity	F1
Ship-related	0.8077	0.8696	0.9091	0.8889
Voyage-related	0.6538	0.9333	0.6364	0.7568
History-related	0.8462	0.8750	0.9545	0.9130

Source: Own work.

sea state on the predicted route. This aspect was beyond the scope of this study and is foreseen as future work.

In the final step of the analysis, the overall reliability measure of MRRAM was determined (it was conducted using the method described in Section 7.2). To this end, the weights reflecting the significance of a given classifier (its relative importance) were defined. The following weights were set up: 0.3 for the ship-related, 0.5 for the voyage-related, and 0.1 for the history-related classifier. Moreover, the confidence measure for each classifier and the whole MRRAM were determined, taking into account the number of input variables. The final results are summarized in Table 7.3.

The final reliability measure is the probability that the delivery/transport service will be realized on time, which is 1 – the risk of the ship being delayed. The higher the MRRAM overall reliability measure, the higher the reliability of the delivery being realized by a given ship. In other words, the higher the risk that the ship will be delayed, the lower the reliability of the supply.

To assess the estimations provided by MRRAM a simple thresholding rule to discriminate between 'high-risk' ships (with a high probability of a delay) and 'low-risk' ships (with a low probability of a delay) for different risk thresholds was used. It means that if the estimated risk of a delay is higher than the threshold, the MRRAM reliability equals 1, and 0 otherwise.

Here, different values of the threshold were analyzed, taking into account the False Positive and False Negative rates, i.e., the percentage of ships that were delayed and were wrongly misclassified as 'low-risk' and the percentage of ships that were classified as 'high-risk' and the actual delay did not occur. These results are presented in Table 7.4.

For example, for the risk threshold of 0.13 (reliability threshold 0.87) MRRAM correctly classified 80% of ships as either delayed ('high-risk') or not ('low-risk'); 15% of ships were classified as 'high-risk,' although they were punctual, and 3% of ships that were actually delayed were wrongly classified as 'low-risk.'

Regardless of the selected risk threshold, the results show that MRRAM is characterized by relatively high accuracy and precision. Nevertheless, in order to select the value of the risk threshold a measure that combines both precision and sensitivity might be used. An example of such a measure is F1, helps to find a balance between precision and sensitivity. Taking into account this measure,

Table 7.3. Summary of estimations for the MRRAM overall reliability and risk measure

Voyage	MMSI	Ship classifier Weight = 0.3	Voyage classifier Weight = 0.5	History classifier Weight = 0.2	Actual delay in hours	Actual Delay [TRUE / FALSE]	MRRAM Overall Delay Risk	MRRAM Overall Reliability	Confidence
Bilbao 1	309072000	0.3303	0.2000	0.4223	13.01	1	0.2835	0.7165	70%
Bilbao 2	477547300	0.3358	0.1111	0.3850	3.98	1	0.2333	0.7667	81%
Teesport	255804960	0.0833	0.1538	0.4070	−17.26	1	0.1833	0.8167	70%
Montoir	246823000	0.3273	0.0323	0.4268	1.72	0	0.1997	0.8003	74%
Kambo	311000237	0.0000	0.0000	0.3481	5.08	1	0.1446	0.8553	70%
Muuga	212118000	0.2692	0.2500	0.3594	3.61	1	0.2776	0.7224	63%
Livorno	258802000	0.4138	0.0741	0.4432	−79.62	1	0.2498	0.7502	70%
Valencia	636014997	0.6667	0.1818	0.3189	1.80	0	0.3547	0.6453	70%
Algeciras	257588000	0.2722	0.2000	0.2548	8.06	1	0.2326	0.7674	70%
Istanbul	353992000	0.2556	0.2500	0.4154	48.09	1	0.2848	0.7152	63%
La Spezia	548765000	0.1364	0.0000	0.2458	−3.92	1	0.0901	0.9099	70%
Le Havre	249701000	0.2852	0.2467	0.2548	261.91	1	0.2598	0.7402	70%
Mersin	538004227	0.2953	0.0833	0.6000	3.34	1	0.2503	0.7497	81%
Marseille	210318000	0.3025	0.2500	0.2985	5.93	1	0.2754	0.7246	70%
Gdansk	565978000	0.2775	0.1724	0.4571	313.66	1	0.2609	0.7391	74%
Gothenburg	212722000	0.3333	0.0294	0.4643	482.47	1	0.2076	0.7924	81%
Fos	372016000	0.2904	0.2576	0.6000	4.07	1	0.3359	0.6641	81%
Aliaga	538090282	0.2693	0.1200	0.5493	4.84	1	0.2507	0.7493	78%
Tenerife	538004027	0.4167	0.0000	0.3551	3.42	1	0.1960	0.8040	70%
Swinoujscie	538003236	0.3000	0.0000	0.4183	18.12	1	0.1737	0.8263	63%
Sines	247226900	0.3141	0.2500	0.3514	3.84	1	0.2895	0.7105	78%
Barcelona	371426000	0.2932	0.0345	0.2453	1.94	0	0.1543	0.8457	70%
Piraeus	636091659	0.2500	0.2692	0.2946	3.81	1	0.2685	0.7315	70%
Genoa	220620000	0.3349	0.0000	0.3286	7.79	1	0.1662	0.8338	74%
Felixstowe	354625000	0.2000	0.0714	0.4000	−1.12	0	0.1757	0.8243	74%
Lysekil	309553000	0.2735	0.1250	0.4234	17.37	1	0.2292	0.7708	63%

Source: Own work.

the risk threshold equaling 0.05 might be used. Moreover, the risk threshold might also be included as one of the parameters of MRRAM. Then, its value could be determined automatically by maximizing the F1 measure.

Another approach to combine the results of all the classifiers might be application of the ensemble methodology (Rokach, 2010). The main idea of ensemble is to weigh several individual classifiers and combine them in order to obtain a classifier that outperforms every one of them. This is one of the directions of future work on the method.

Summarizing, it might be concluded that MRRAM is an effective tool for determining the short-term reliability of a transport service and the risk of a delay of an individual ship on a given voyage, and it might be used to identify ships with a high or low reliability.

Table 7.4. Comparison of the MRRAM method depending on the risk threshold

Threshold	0.05	0.1	0.13	0.15	0.18	0.2
Accuracy	0.8462	0.8077	0.8077	0.7692	0.7692	0.7308
Precision	0.8462	0.8400	0.8400	0.8333	0.9000	0.9412
False Negatives	0.0000	0.0385	0.0385	0.0769	0.1538	0.2308
False Positive	0.1538	0.1538	0.1538	0.1538	0.0769	0.0385
True Negative	0.0000	0.0000	0.0000	0.0000	0.0769	0.1154
True Positive	0.8462	0.8077	0.8077	0.7692	0.6923	0.6154
F1	0.9167	0.8936	0.8936	0.8695	0.8571	0.8205

Source: Own work.

The estimated risk classifiers also provide interesting information regarding the relationships between input variables and the occurrence of the delay of ships (see the conditional probability tables in Appendix A). For example, if we know that a ship is classified by an unreliable classification society (the ClassificationSociety variable), the probability of being delayed or being punctual is very similar (it amounts to 44% and 45% accordingly). It is similar with the reliable classification societies—in this case the probability of being delayed or punctual is almost the same (55% and 54%), on condition that no other information is available. This might suggest that the classification society is not such a relevant variable in determining the punctuality of a ship. But then, for ships with the history of loitering (the Loitering variable) the probability of being delayed amounts to 24%, while being on time only 18%. Similar differences might be observed for Size, Congestion, Delay, Hazard, TravelTime, Detentions, PastDelays, ProtectedAreas, StaticChanges. Some of these dependencies are not obvious and without conducting a data analysis they might be difficult to observe. Therefore, it might be concluded that the results

of MRRAM may suggest which variables are especially significant for determining a ship's reliability, and thus they might be used to determine the importance of a given classifier in the estimation of the overall reliability (the weight) but also in the discovery of some general relationships between different ship's characteristics and/or its operational environment, and the ship's behavior.

7.3.3. Ranking of ships

As it was mentioned in the description of MRRAM, it might also be used to rank ships from the point of view of the predicted risk and reliability. To evaluate this functionality, the next experiment was conducted.

Let us assume that we have a list of ships that may be potentiality used to realize a given supply on a selected voyage. Having this list, we would like to estimate which ship is characterized by the lowest risk, and thus the highest reliability, and which ships are potentially unreliable and should not be taken into account in the first place. For the purpose of the experiment, the list of ships from the testing set was used and the results of MRRAM presented in the previous section.[8] Then, three different rankings were created, depending on the importance of a given risk classifier (different weights).

In the first ranking, the voyage-related variables are the most important (weight 0.5), then the ship-related (weight 0.3), and the history-related (weight 0.2). The ranking was built taking into account the overall reliability measure. Similarly, the next two rankings were created, but with a different importance of classifiers (0.3, 0.5, and 0.2 in ranking 2, and equals weights in ranking 3). The results are presented in Table 7.5.

The results show that in all three rankings there is a single ship (marked MMSI 548765000) that is characterized with the highest reliability. This result is, however, justified because of the characteristics of this ship—it is a middle age ship, classified by a reliable classification society and a delivered classification status, it is of a safe type, with a relatively good history (no history of accidents, pollution, casualties, or loitering behavior), as well as good predictions regarding the voyage (standard travel time, no dangerous cargo, or no predicted delay).

At the bottom of the rankings, there are: a very large ship of a dangerous type, with a history of being inspected/detained (MMSI 636014997) and a ship that belongs and belonged to an unreliable classification society, with a history

8. Let us assume here that the results concern the same voyage, although we know that they were calculated for different voyages. However, in reality a given voyage may be realized on different routes, so the estimations for voyage-related variables may differ for different ships even for the same voyage.

Table 7.5. Ranking of ships depending on the importance of the risk classifiers

Ranking 1 (WS = 0.3; WV = 0.5; VH = 0.2)				Ranking 2 (WS = 0.5; WV = 0; WH = 0.5)				Ranking 3 (WS = 0.33; WV = 0.33; WH = 0.33)			
Position	MRRAM Overall Reliability	Confidence	MMSI	Position	MRRAM Overall Reliability	Confidence	MMSI	Position	MRRAM Overall Reliability	Confidence	MMSI
1	0.9099	70%	548765000	1	0.8089	70%	548765000	1	0.8726	70%	548765000
2	0.8554	70%	311000237	2	0.7548	70%	255804960	2	0.8090	70%	371426000
3	0.8457	70%	371426000	3	0.7365	70%	257588000	3	0.8006	70%	311000237
4	0.8338	74%	220620000	4	0.7308	70%	371426000	4	0.7853	70%	255804960
5	0.8263	63%	538003236	5	0.7300	70%	249701000	5	0.7788	74%	220620000
6	0.8243	74%	354625000	6	0.7277	70%	636091659	6	0.7762	74%	354625000
7	0.8167	70%	255804960	7	0.7009	70%	311000237	7	0.7606	63%	538003236
8	0.8040	70%	538004027	8	0.7000	74%	354625000	8	0.7577	70%	257588000
9	0.8003	74%	246823000	9	0.6995	70%	210318000	9	0.7427	70%	538004027
10	0.7924	81%	212722000	10	0.6857	63%	212118000	10	0.7379	74%	246823000
11	0.7708	63%	309553000	11	0.6682	74%	220620000	11	0.7378	70%	249701000
12	0.7674	70%	257588000	12	0.6672	78%	247226900	12	0.7287	70%	636091659
13	0.7667	81%	477547300	13	0.6645	63%	353992000	13	0.7261	63%	309553000
14	0.7502	70%	258802000	14	0.6516	63%	309553000	14	0.7243	81%	212722000
15	0.7497	81%	538004227	15	0.6409	63%	538003236	15	0.7227	81%	477547300
16	0.7493	78%	538090282	16	0.6396	81%	477547300	16	0.7163	70%	210318000
17	0.7402	70%	249701000	17	0.6327	74%	565978000	17	0.7071	63%	212118000
18	0.7391	74%	565978000	18	0.6237	70%	309072000	18	0.6977	74%	565978000
19	0.7315	70%	636091659	19	0.6229	74%	246823000	19	0.6948	78%	247226900
20	0.7246	70%	210318000	20	0.6141	70%	538004027	20	0.6930	63%	353992000
21	0.7224	63%	212118000	21	0.6012	81%	212722000	21	0.6896	70%	258802000
22	0.7165	70%	309072000	22	0.5907	78%	538090282	22	0.6871	78%	538090282
23	0.7152	63%	353992000	23	0.5715	70%	258802000	23	0.6825	70%	309072000
24	0.7105	78%	247226900	24	0.5548	81%	372016000	24	0.6738	81%	538004227
25	0.6641	81%	372016000	25	0.5523	81%	538004227	25	0.6174	81%	372016000
26	0.6453	70%	636014997	26	0.5072	70%	636014997	26	0.6109	70%	636014997

Legend: WS: Weight ship; WV: Weight voyage; WH: Weight history.

Source: Own work

of accidents, detentions and casualties, and for which the predicted travel time deviates from normalcy and thus a delay is foreseen (MMSI 372016000).

Apart from the first and the last position, each ranking differs from each other. This may suggest that depending on the importance of a given classifier, different ships are indicated as those which may be used for the realization of a delivery. In the case of ranking 1, where the ship-related variables are the most important, at the top of the list dominate middle age ships of the safe type and delivered classification societies. In ranking 2, where the voyage-related variables are prime, at the top there are ships with a standard travel time and no delays predicted.

In conclusion, it might be said that the results provided by MRRAM might be used to rank ships from the point of view of reliability and to select the most and the least appropriate ships to realize the transport service.

7.3.4. Summary of the results

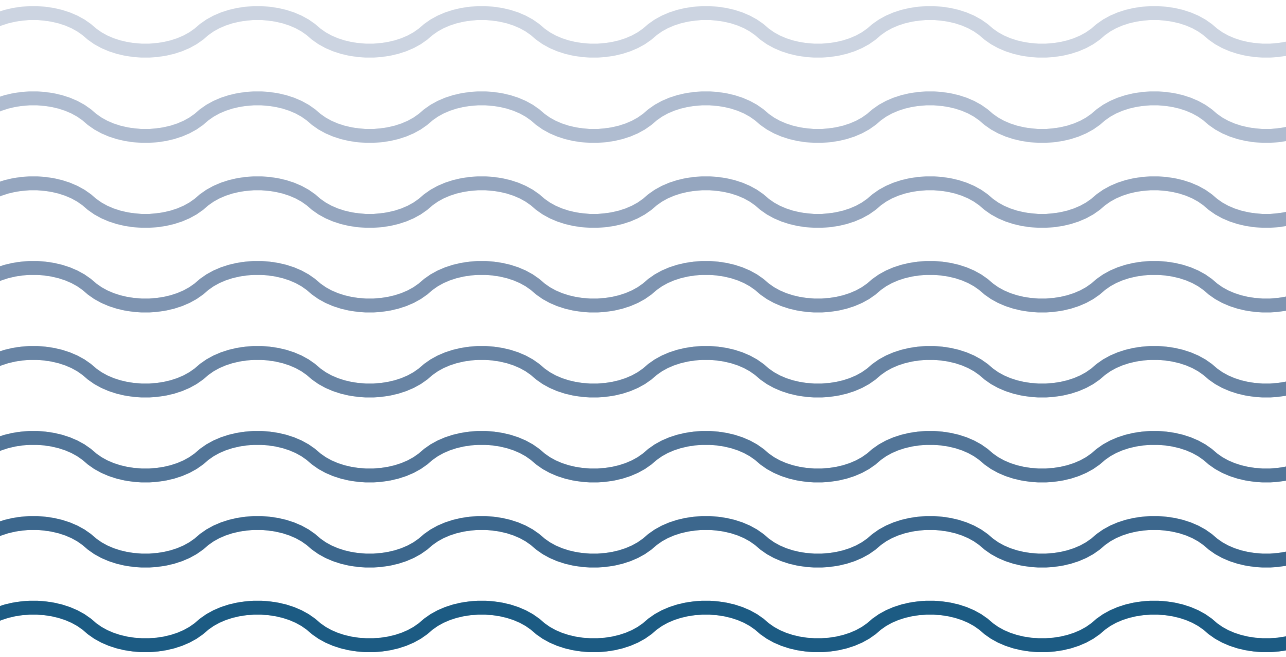
The aim of the presented analysis of MRRAM was to show its effectiveness in providing a short-term reliability assessment of ships, its ability to provide an accurate assessment for real world examples, and its usefulness in the decision-making process.

The presented results of the experiments proves the quality of the MRRAM method. Their results confirmed the ability of the method to provide accurate reliability assessment for real world examples of voyages and using real data. Depending on the propensity to take the risks (risk threshold), and the importance of different risk variables (weights of classifiers), the accuracy of the MRRAM results slightly varies and it amounts between 73.08% (for higher risk thresholds) and 84.62% (for lower risk-thresholds). Nevertheless, based on the research it was proved that for the selected set of sample voyages, the method is able to correctly predict the reliability of a ship in 73% to 84% of cases, depending on the defined risk threshold.

Moreover, the reliability estimation based on MRRAM, conducted before or at the beginning of a ship's voyage, seems to be a useful and helpful tool for potential users who, based on the received results, might check whether the information (estimation) of travel time provided by a captain is realistic, taking into account additional information about the ship, its history and current conditions of the ship's operational environment.

Finally, the last experiment confirmed that MRRAM is an effective and useful tool for creating a ranking of ships from the point of view of the reliability of delivery and, as a result, could be a support in decision-making when a user needs to decide which ship is the best/ is appropriate for the realization of a punctual and reliable delivery of goods on a given route.

Chapter 8



8. SHIP'S PUNCTUALITY PREDICTION

One of the key elements in assessment the reliability of the transport service is determination whether the service will be timely realized and if a ship will punctually arrive at a destination port. Therefore, the method for a short-term assessment of maritime reliability and risk (MRRAM) presented in the previous chapter includes variables that reflect a ship's punctuality. However, prediction of a ship's punctuality is a complex task in which various requirements and factors need to be taken into account. In this chapter we present a method for Ship's Punctuality Prediction (SPP).

In a wider context, the SPP method is a solution for a short-term assessment of maritime reliability since it provides dynamic updates on the predicted punctuality during a ship's voyage. The SPP takes into account three main elements that together allow for an assessment of a ship's punctuality: prediction of the route to the destination, estimation of travel time, and additional factors that may influence travel time. For each of these elements, a set of algorithms has been proposed.

8.1. Outline of the method

The method for Ship's Punctuality Prediction (SPP), as its name suggests, predicts whether a ship will punctually arrive at a destination port. It provides an estimation of the time when a ship will arrive at a destination port, compares it with the information declared by the captain of a ship, and based on it predicts whether the ship will be on time or not.

Similarly to MRRAM, SPP also relates to a given voyage (delivery) and to an individual ship, but here the main focus is put on determination of a single characteristic of this voyage, namely punctuality. In general, it encompasses most of the voyage-related variables of MRRAM. The method has been designed and developed taking into account challenges and drawbacks of the existing methods, presented in Section 3.4, as well as the data available in this study, presented in Section 4.5.

SPP predicts the route and travel time for a given ship's voyage, estimates the arrival time at a destination port, and compares it with what is declared by the ship itself. Based on it, SPP provides information whether the ship is able to reach the destination in the declared time period. Moreover, SPP monitors the travel

time along the ship's voyage and provides updates on the predicted punctuality (dynamic approach).

Similarly to MRRAM, in this research we focus on punctuality prediction only for merchant ships that transport cargo. However, SPP might also be used for other types of ships.

The problem of determining whether a ship will arrive punctually at a destination port requires the following three aspects to be taken into account:

(1) Predicted route to the destination.
(2) Estimated travel time.
(3) Additional factors that may influence travel time.

SPP includes all three aspects. To this end, the method analyses worldwide ships movements from actual and historical AIS data (ships trajectories) and based on it predicts the route a ship will follow. Then, based on the historical voyages, it automatically estimates the travel time. Finally, it takes into account additional variables, such as congestion, possible hazard, and forecast weather conditions to predict the time of arrival.

The proposed method is a similar solution to smart car navigation systems. We assume that knowing the current position of a ship, its destination, and a probable route, and having the historical distribution of the ship's speed along that route, it is possible to estimate the travel time and the ship's punctuality. Like in car systems, the estimated travel time can be further improved by taking into account additional information on current congestion at sea or in ports.

Estimation of the punctuality of a ship can be calculated at the beginning of the ship's voyage (as a planning tool). This process can be then repeated when the ship is already underway. Thanks to this, updates on the arrival time can be provided on a regular basis taking into account new information that might arrive (a monitoring tool).

In the research, a set of methods that allow for an automatic determination of a ship's punctuality has been developed. The approach consists of five components that are required for the final determination of the punctuality:

(1) Route prediction: a first step in which a probable route, that the ship will follow to reach the destination port from a given location, is predicted.
(2) Travel time profile: having the predicted route, a travel time to the destination for a given route is calculated. As a result, a travel profile for a given voyage is created that presents a standard travel time for a given route.
(3) Congestion factor: a method for determining the current congestion on the predicted route; this information may lead to an increase or decrease of the predicted travel time.
(4) Hazard index: a measure calculated based on information about potentially hazardous factors on the predicted route, such as accidents, piracy and risk of

the countries through which the ship will be sailing; this information may lead to an increase of the predicted travel time.

(5) Weather and sea state: information about (actual or predicted) unfavorable weather conditions (e.g., heavy rain or wind, dense fog) or weather extremes (e.g., hurricane) on the route that potentially may lead to an increase of travel time.

(6) Average delay: a method for determining the average delay on the route, calculated based on the historical voyages and the detected differences between the declared ETA and the actual arrival time.

The overall idea of the SPP method is presented in Figure 8.1.

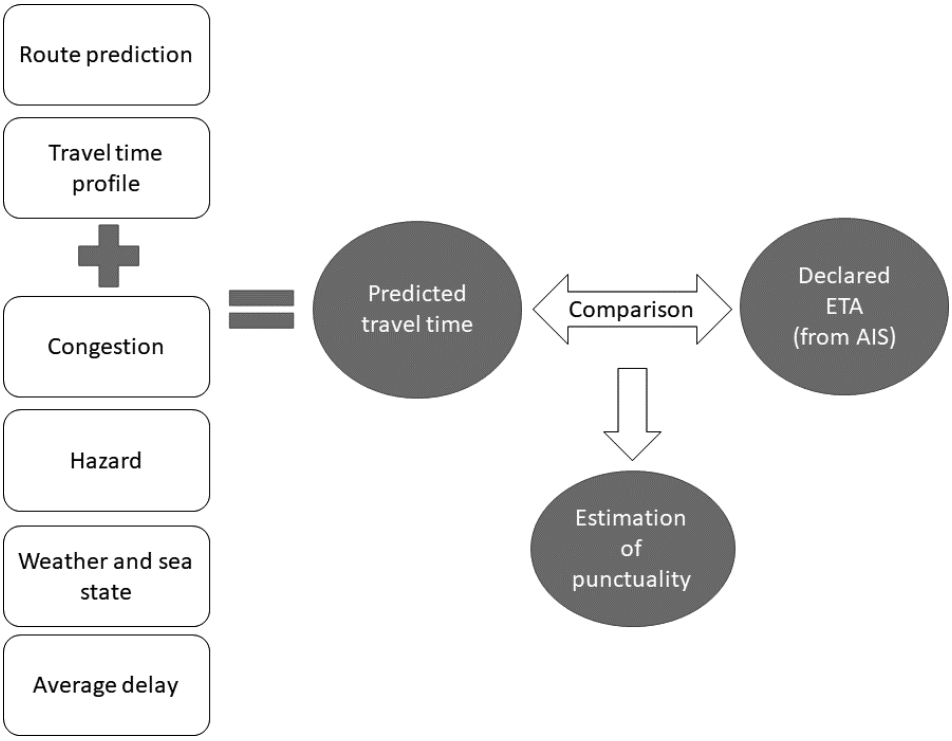


Figure 8.1. Overview of the Ship's Punctuality Prediction method

Source: Own work.

In order to be able to combine the results of the above methods we had to define a common basis when it comes to their calculation. To this end, it was decided to adopt a concept of maritime sectors. This concept assumes that the globe is divided into smaller areas, called sectors. For example, in the evaluation of the SPP method (presented further in Section 8.6), we distinguished 7200 sectors,

rectangles of $3^\circ \times 3^\circ$. However, the size of a sector can be adjusted by the user and all measures of the SPP method can be calculated for both smaller or bigger sectors.

While developing methods for SPP, there was a number of issues that had to be considered. Some of them were also indicated by other researchers (Heywood, Connor, Browning, Smith, & Wang, 2009). They are described in short in the following points:

- The location to be analyzed (either a start point or a destination) will hardly ever be the exact same location as the locations sent by other ships and stored in the database. This problem was solved by implementing a sensitivity radius r around the current location. The AIS messages sent within the defined radius are perceived as sent from the same localization.
- Determination of the exact time of arrival at a destination port (an exact point in time when a ship reached the port). This issue results from the fact that a port covers an irregular area and its boundaries are not strictly defined. Also, a sensitivity radius around the defined port's coordinates is implemented here. The arrival time is set as soon as the ship reaches the area defined by the radius (as soon as it sends an AIS message inside the port area).
- Determination whether a ship travels nearby an object (nearby another ship or a port). In this case, if ship sends an AIS message in the same sector in which the other object is located, it is determined as "being nearby".

In the following sections, the developed methods for each SPP component are presented.

8.2. Route prediction

The first step in estimating a ship's punctuality is determination of a probable route the ship will follow from a starting point to a destination. Here, three alternative types of routes are considered (three conditions are checked): 1) routes of a given ship on the same voyage[1]; 2) routes of other ships on the same voyage; 3) routes of other ships on a similar voyage. Within the algorithm, the following heuristic has been designed to determine the route from a start location to a destination.

Case 1 Determine whether a ship has already traveled on this route in the past; if yes, find the past trajectory(-ies) of the ship and determine the set of sectors which were followed.

Case 2 Determine whether there are other ships that have traveled on this route in the past; if yes, find the past trajectories of these ships and determine the set of sectors which were followed.

1. Voyage means a travel from the start location to the destination.

Case 3 Determine whether there are other ships that have been traveling on a similar route; if yes, find the past trajectories of these ships and determine the set of sectors which were followed.

Final step Based on the results of the previous steps determine the most probable route for the ship.

A similar voyage in case 3 means traveling nearby a given destination (e.g., sailing in close proximity to the destination port but without calling it) or traveling to another port that is located nearby the destination port. The nearby ports should be located in the same sector as the destination port or in the proximity not larger than the size of the sector (e.g., 3°).

The route is understood here as a set of consecutive sectors from the start location to the destination port. The trajectory is a set of consecutive AIS messages (geographical coordinates) sent by a given ship. The designed algorithm is also presented in Algorithm 8.1.

Algorithm 8.1. Route prediction

```
1:  procedure FindTrajectories(AIS messages, destination, start location, reference ship)
    ▷ find trajectories of all ships that travelled to the declared destination
2:      if the reference ship is on the list then                          ▷ Case 1
3:          take trajectories of the reference ship
4:          for all found trajectories do
5:              determine if the ship has been travelling from the start location
6:              if yes then
7:                  check when the ship arrived to destination             ▷ Algorithm 3
8:                  determine sectors that were followed                   ▷ Algorithm 4
9:              else if no then
10:                 take next trajectory
11:     else if the reference ship is not on the list then                 ▷ Case 2
12:         take trajectories of other ships on the list
13:         determine all ships that travelled on this route
14:         for all found trajectories do
15:             check when the ship arrived to destination                 ▷ Algorithm 3
16:             determine sectors that were followed                       ▷ Algorithm 4
17:     find other ports in the same sector as destination
18:     find trajectories of all ships that travelled to the other ports   ▷ Case 3
19:     for all found trajectories do
20:         determine sectors that were followed on similar route          ▷ Algorithm 5
21:         determine the predicted travel time
22:     return Predicted route from start location to destination
```

In each case, additional methods for checking whether a ship is in a port and what its exact arrival time was (Algorithm 8.2) and for determining the set of sectors that were followed (Algorithm 8.3), were designed.

Algorithm 8.2. Check if ship is in port

```
1:  procedure CheckIfShipInPort(AIS messages, ship, destination port, radius, time
    range)
2:      find AIS messages send in defined time range
3:      determine sector of destination port
4:      for each AIS message do
5:          determine sector
6:          check if sector is the same as sector of destination port
7:          if yes then
8:              check if message sent in radius of destination port
9:              if yes then
10:                 set AIS timestamp as time of arrival to destination port
11:             else if no then
12:                 take next AIS message
13:         else if No then
14:             take next AIS message
15:     return Time of arrival to destination port
```

Algorithm 8.3. Determine set of sectors

```
1:  procedure DetermineSetOfSectors(AIS messages, ship, start location, destination
    port, time range)
2:      find AIS messages send in defined time range        ▷ time range defined based
    on timestamp of start location and time of arrival to destination (case 1 and 2) or ETA
    (case 3)
3:      determine sector of start location
4:      set sector of start location as the first sector of the route
5:      for each AIS message do
6:          determine sector
7:          check if sector of message is the same as sector of the previous message
8:          if no then
9:              add sector to ship's route
10:             take next message
11:         else if yes then
12:             take next AIS message
13:     save ship's route
14:     for each sector of the route do
```

```
        determine time and distance
    calculate travel time and distance
    return ship's route
```

Based on all the found routes, the most probable route for a given ship is predicted (Algorithm 8.4). Here, we analyze a sequence of sectors for all the found routes. Each sector in the sequence is a separate segment of the route. Then, for each segment the most common sector is selected (the sector that was followed most often in the past by ships on the same/similar voyage). In the case when for a given segment there are several sectors that were followed with the same frequency, the sector with the highest ship density is selected (the assumption is that a ship will probably follow the sector that is generally most often used).

Algorithm 8.4. Determine the most probable route

```
procedure DetermineMostProbableRoute(past routes, start sector, destination sector)
    add start sector as first segment of predicted route
    for each past route do
        take sequence of sectors that have been followed
        number the sequence as separate segment   ▷ Each sector is a segment of the route
    for each segment of all past routes do
        determine the most common sector
        if there is more then one the most common sector then
            determine ship density for these sectors                ▷ Eq. (8.1) or (8.2)
            take sector with the highest ship density
            add sector as next segment of predicted route
            if added sector is destination sector then
                break
            else if added sector is not destination sector then
                take next segment
        else if there is one the most common sector then
            add sector as next segment of predicted route
            if added sector is destination sector then
                break
            else if added sector is not destination sector then
                take next segment
    save predicted route
    calculate predicted travel time
    return predicted route
```

The next sectors are added to the route as consecutive segments until the destination sector is reached. As a result, the predicted route with a sequence of sectors that will be followed is provided.

Having the route predicted, the travel time can be determined. Here, two approaches are considered. They are described in the next section. Figure 8.2 presents an example of route prediction for a selected ship and voyage; the dark sectors indicate the predicted route.

Route prediction
Sector sequence
1502
1382
1383
1263
1264

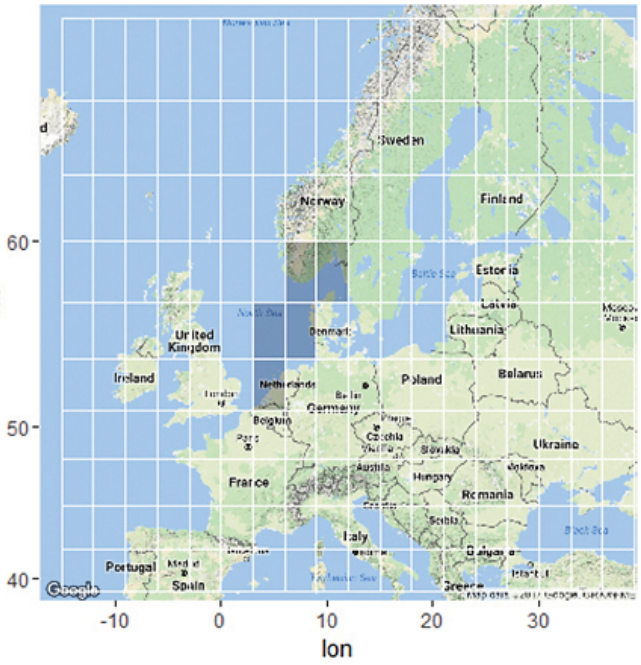


Figure 8.2. Route prediction example. Route prediction for the ship NORTHSEA BETA (MMSI 248970000), Voyage: Maasvlakte (Rotterdam)–Goteborg, Data source: AIS data, 1 year period

Source: Own work.

8.3. Travel time profile

As indicated in Section 3.2.3, travel time can be calculated based on historical data. There are two approaches that can be utilized:

- Based on historical travel time—the average travel time between the departure and destination location is calculated based on historical routes of a ship and other ships.
- Based on the distance to travel and the average speed—having historical AIS data, a distribution of ships' speeds in different maritime areas can be modeled. To this end, for each sector an average speed is calculated that further is used to estimate travel time.

The first approach assumes that travel time is a sum of average travel times of sectors that create the route. The average time for a sector is calculated based on the amount of time that ships have spent in this sector (during past voyages). The time is calculated as the difference between the first AIS message sent in a given sector and the first AIS message sent outside this sector.

The second approach assumes that travel time is calculated as a quotient of distance and speed. For each sector from the predicted route an average speed as well as an average distance are determined. The distance in a sector is calculated as a sum of the distances between consecutive AIS messages sent by a ship in this sector. The sector's average distance can be calculated based on distances traveled by ships on the voyage in the past. The sector's average speed is calculated based on the distance and travel time in a given sector. It can be calculated either only for ships that have traveled the voyage in the past or based on the speed of all ships that have crossed this sector (the general average speed for the sector).

For both approaches, also the standard deviation of travel time can be calculated. This information can be further used in the method, for determining whether a ship will arrive on time—it can be used as a threshold in determination whether there is a significant deviation of travel time in comparison to past voyages.

Using the above approaches, a travel profile from a given location to a destination is built. The profile shows the standard travel time on the route, where on the X axis we have the planned route (a set of sectors) and on the Y axis the corresponding cumulative travel time. An example of a travel profile is presented in Figure 8.3. It is calculated taking into account various conditions which are presented using separate lines. For example, the red line shows the travel profile calculated based on past voyages of the ship (case 1). The average historical travel time is calculated based on past voyages of the ship and other ships (cases 1–3) and the travel time calculated using the second approach. Travel times with congestions and with hazards include updates of travel time taking into account additional information (described in the next sections).

Using the travel profile, it is possible to track whether there are any deviations in the travel time in the current voyage. If so, a dynamic re-calculation of the

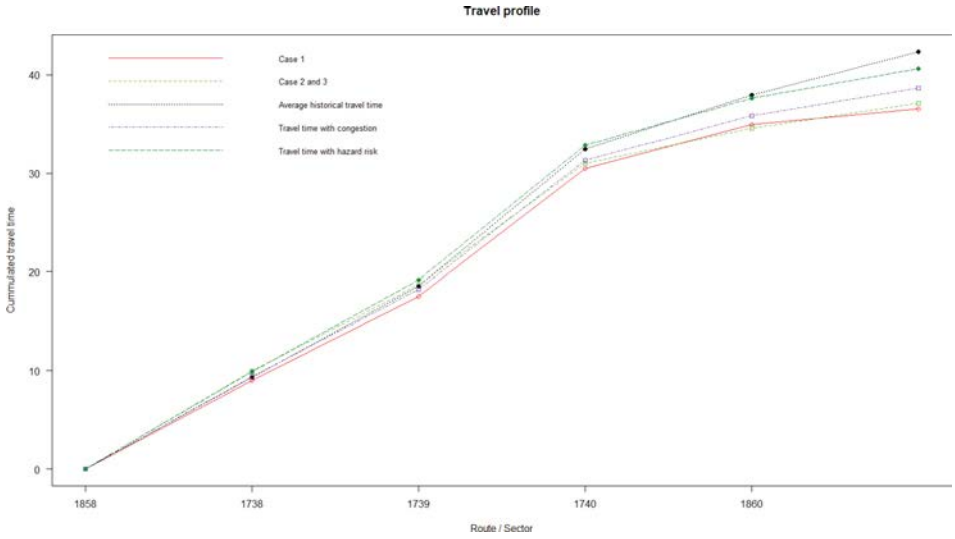


Figure 8.3. Example of a travel profile

Source: Own work.

predicted travel time can be conducted to provide updates of the predicted arrival time, and thus determine the current prediction of a ship's punctuality.

8.4. Additional variables

The basic prediction of travel time can be further updated taking into account additional information that concerns the operational environment on the predicted route. The additional variables proposed in SPP include congestion, hazards, weather and sea state, and past delays. The developed methods for their calculation are presented in the following sections.

8.4.1. Congestion

The analysis and implementation of the congestion factor is conducted in a few steps. First, an average ships density in each maritime sector is calculated. Then, the density is used to determine whether there is a congestion on the route being analyzed. Congestion appears when the ships density in a given area in the last 24 hours is higher than the standard (average) density. If congestion occurs, the

estimated travel time is adjusted taking into account the difference between the current congestion measure (in the last 24 hours) and the average congestion.

Application of congestion on travel time is done using the following heuristic:

(1) Calculate the average monthly ships density for each sector.
(2) Set the moment in time for which congestion is to be determined (e.g., the start of a voyage or a moment in time when a ship is under way).
(3) Calculate the current ships density (in the last 24 hours).
(4) Compare the current and the average density.
(5) If the calculated difference is higher, then a defined threshold (standard deviation), the congestion factor, is calculated. The threshold causes the calculated difference to be treated as congestion only if a significant deviation appears (small differences will always appear since the number of ships in a given area is constantly changing, which is a normal phenomenon).
(6) If the congestion factor is other than null, its value is used to adjust the estimated travel time for a given sector. The travel time may increase if the congestion is higher than the average, or may decrease, if the congestion is lower than the average.

In the course of this research it was assumed that the standard deviation of ships density will be used as a threshold to distinguish a congested sector. This value can be, however, modified and adjusted to suit the user's needs. We are aware that utilization of the standard deviation might be seen as a simplification because the process of determining the congestion on a route is a rather complex task. Examples from other domains (e.g., congestion on roads) show that it might be a non-linear phenomenon. A deeper analysis of this phenomenon was beyond the scope of this research. Nevertheless, as it seems to be an interesting area of future work.

Below we present, in more details, steps for determining the average and the current density as well as a calculation of the congestion factor. All these values can be calculated based on historical positions of ships from AIS.

Ships density calculation

First, we determine the number of ships in each sector in a given time period (in a given month). However, in this calculation we take into account the fact that ships are constantly moving and in a given time period may cross the sector several times. To this end, 1-hour "snapshots" are used. It means that for each hour of a day, the number of distinct ships in each sector is calculated. This makes it possible to take into account the movements of ship between sectors (assuming that during this hour a ship does not leave and return to the same sector). Then,

these hour's statistics are aggregated to determine the average monthly number of ships in each sector.

Finally, ships density is calculated. To this end, we adopted, with some modifications, approaches proposed by (Wu et al., 2017) and (Eiden & Martinsen, 2010)—traffic density can be calculated using two formulas.

In the first case (formula (8.1)), the traffic density in a maritime region (sector *s*) in a given time period *m* is defined as the average number of vessels per unit area. It is calculated as a total number of ships that were present in a given sector *s* in a given period *m* divided by the area of the sector.

$$\mathrm{Density}_s^m = \frac{\mathrm{ShipCountSum}_s^m}{\mathrm{Area}_s} \tag{8.1}$$

where $\mathrm{Density}_s^m$—the density in sector *s* in a month *m*; $\mathrm{ShipCountSum}_s^m$—the number of ships that crossed sector *s* in a month *m*; Area_s—area of sector *s*.

In the second case (formula (8.2)), the traffic density in a given region *s* and in a given time period *m* is defined as a ratio of the total number of ships present in a given sector *s* in a given period *m* to the total population of ships.

$$\mathrm{DensityRatio}_s^m = \frac{\mathrm{ShipCountSum}_s^m}{\mathrm{ShipPopulation}} \tag{8.2}$$

where $\mathrm{DensityRatio}_s^m$—the density in sector *s* in a month *m*; $\mathrm{ShipCountSum}_s^m$—the number of ships that crossed sector *s* in a month *m*; ShipPopulation—the total number of ships.

Congestion factor calculation

Having the average density for each sector, we can start analyzing whether at the time of planning or realizing the voyage there is a congestion on the predicted route. To this end, the difference between the average and the current density (in the last 24 hours) is calculated. The difference can be both positive and negative, and takes values between –1 and 1. The positive difference occurs when the current density is higher than the average. The negative occurs otherwise.

As indicated above, a difference, even a very small one, almost always occurs. It results from the fact that the number of ships in a given area is constantly changing and the probability that the number of ships in a given area in a given moment will be exactly the same as the average is small. Therefore, it was decided that an area will be perceived as 'congested' only if the current density will be higher than the standard deviation. In such a case, the congestion factor is calculated according to formula (8.3):

$$\mathrm{CongestionFactor}_s = \frac{\mathrm{DensityDifference}_s}{\mathrm{AverageDensity}_s} \tag{8.3}$$

where $\text{CongestionFactor}_s$—the congestion factor in sector s; $\text{DensityDifference}_s$—the difference between the current and average congestion in sector s; AverageDensity_s—the average congestion in sector s.

Update of the predicted travel time

Having calculated the congestion factor for all the sectors in the predicted route, in the next step the predicted travel time may be updated. We assume here that if a positive congestion occurs the predicted travel time should be increased while for a negative congestion it might be decreased. To this end, the average travel time for each sector where the congestion occurs is multiplied by the congestion factor. Then, the total travel time for the predicted route is updated.

8.4.2. Hazard index

The aim of the hazard index is to incorporate the geopolitical factors in the travel time prediction. The concept of the hazard index, together with examples of its application, is presented in (Stróżyna, 2017b). In SPP a hazard (risk) level is determined for different maritime areas, taking into account various geopolitical factors. The proposed hazard index includes 3 types of factors:

(1) Maritime accidents, which takes into account the number of maritime accidents which have happened in a given area in the past.
(2) Piracy, which takes into account the reported piracy attacks and armed robberies which have happened in a given area in the past.
(3) Country Risk, which analyses the risk of the departure and destination country as well as the risk of countries a ship will travel through during its voyage.[2]

Maritime accidents

The first variable, which may influence the level of geopolitical risk in a given maritime area, is number of past maritime accidents (*Accident* in short). In order to include this information in the hazard index, first we checked how the accidents were spread out in space, and based on it, whether it is possible to identify maritime areas which are significantly more prone to accidents, and whether a seasonality in number of accidents can be observed. Here, the data about historical accidents acquired from the GISIS database[3] for a period 2005–2016 was used.

In order to calculate the *Accident* measure, similarly to the congestion calculation, the globe was divided into 7200 sectors ($3^\circ \times 3^\circ$). Then, for each sector yearly

2. *Country* means in this case the Exclusive Economic Zone belonging to a given state.
3. https://gisis.imo.org

and monthly statistics on the number of accidents for the period of 2005–2016 were calculated.

Then, it was tested whether there is a trend or a seasonality in the occurrence of maritime accidents. To this end, first a variance analysis (using the Anova test) was conducted, which confirmed that such a trend actually exists (p-value $2.2e-16$). Then, the trend was eliminated in order to see whether there is a seasonality in particular months of the year. The received results confirmed the seasonality of accidents (p-value $= 0.003372$). The results of Anova were then confirmed using the autocorrelation function (ACF), which also confirmed both the trend and the monthly seasonality in the number of accidents.

Taking into account the results of both analyses, it is essential to include the seasonality in the *Accident* measure. As a result, for each maritime sector a monthly *Accident* measure is calculated. The measure relates the number of accidents in a given month and in a given sector to the total number of maritime accidents reported between 2005–2016.

Piracy

The next factor that is assumed to influence travel time is piracy and acts of armed robbery (*Piracy* in short). Its significance can result from the statistics on this phenomenon presented in Section 2.2.

Similarly to *Accident*, maritime areas with a high density of *Piracy* incidents were identified, including their spread over time. First, it was tested whether trends and seasonality appear in occurrences of *Piracy*. The conducted Anova analysis confirmed the existence of a trend in the data (p-value $= 3.858e-10$). However, after eliminating the trend it turned out that there is no seasonality (p-value $= 0.4031$). As a consequence, for each maritime sector, the *Piracy* measure was calculated without a differentiation for time periods (seasonality). The *Piracy* measure compares the number of piracy attacks in a given sector to the total number of the reported attacks in 2010–2016.

Country risk

The geopolitical risk of a country may result from various aspects, especially political and legal risks. Political risk concerns what kind of economy, tax, societal and legal politics is conducted by the government of a given country. The political situation in a country may influence the economy and, thus, may affect various business processes. In the case of maritime transport, the political risk may concern clearance procedures, import/export fees and taxes (e.g., port and canal fees), law (e.g., embargo), time of transport (e.g., closing of borders or ports).

Legal risk results from the different laws in force in different countries. It may concern, for example, different interpretations of the law, a different terminology

used in contracts or different legal systems. The risk here concerns mainly lack of knowledge or ignorance of local and international legal provisions and contractual clauses (T. T. Kaczmarek, 2012, p. 160).

In order to calculate a risk of the countries a ship is sailing through various aspects can be taken into account, starting from government issues (e.g., regime type, corruption, public transparency), through the security of the country (e.g., current and historical conflicts), social aspects (e.g., education) up to economic issues (e.g., income levels, poverty, unemployment, money laundering), and the exposure to natural hazards. All these factors compose the overall assessment of a country from the point of view of safety and security. This, in turn, influences the certainty that shipping through the Exclusive Economy Zone of a country or calling a port located in this country is safe for the ship, its crew, and the transported cargo.

In the proposed method, three country risk indicators are included: INFORM, Basel AML Index and World Risk Index (they were described in detail in Section 3.2.3). Using the values of each indicator the average *Country Risk* index is determined.

Having determined the average country risk index for the three indicators, information about the type of flag for a given country is included. The type of flag is an important maritime risk factor, but none of the selected indicators take it into account. The flags are generally divided into three colors: black, grey and white. Classification to each group is based on a number of inspections and detentions of ships under a given flag and the performance of the classification society the ship is affiliated with. The low risk flags are classified as white, while the high risk ones are black.

In order to include the information about the type of flag, data published by two well-known Memoranda of Understanding—Paris MoU and Tokyo MoU—can be used. Figure 8.4 presents a map with an indication of flags' colors, including the areas of Exclusive Economic Zones (EEZs) of the countries.

Information about the flag is included in *Country Risk* by application of an increasing factor. If a given flag is on the black list, the *Country Risk* measure is increased by 20%, while for the grey list by 10%.[4] As a result, the overall *Country Risk* index can be calculated, which takes a value between 0 (low risk) and 1 (high risk). Figure 8.5 presents a map with values of the *Country Risk* index for different countries (including EEZs) that was calculated using the proposed approach.

Then, the values of *Country Risk* had to be transferred on the sectors level. To this end, the globe was divided into 7200 sectors ($3^\circ \times 3^\circ$) and the country map was overlapped with a grid of the sectors. For the sectors that overlap with the area of

4. Both increasing factors can be changed and adjusted if needed.

a country or its EEZ, the value of *Country Risk* was simply transferred. However, in the case of the sectors that cover more than one country / EEZ it was more

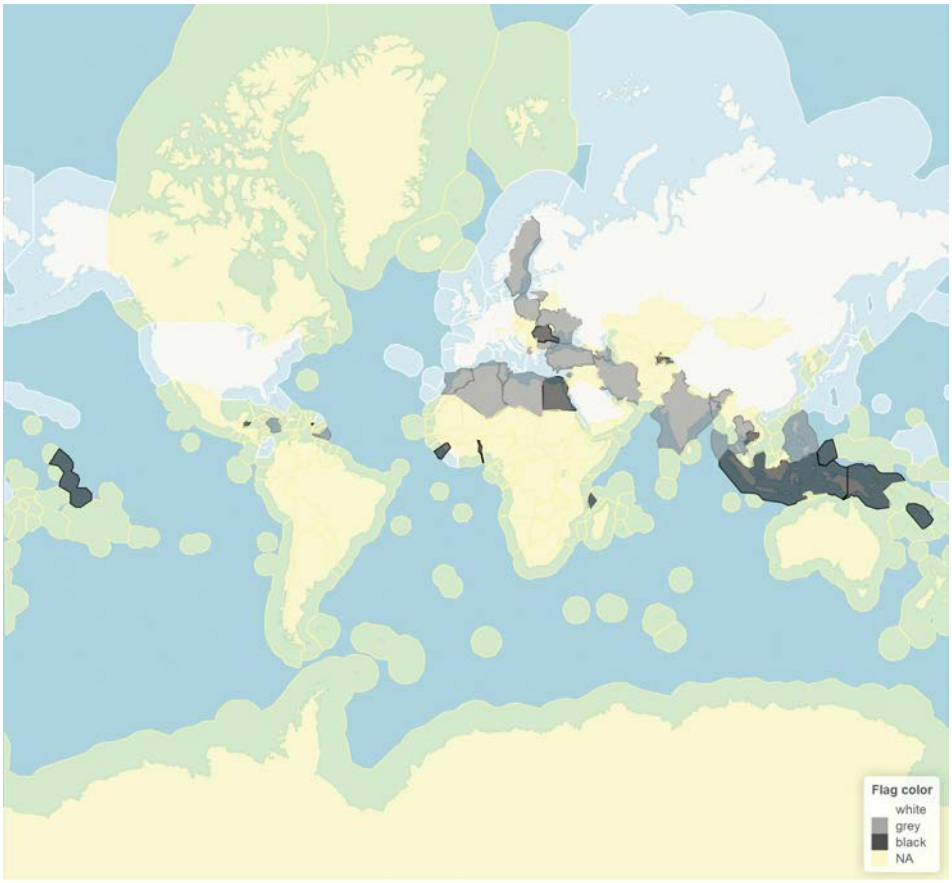


Figure 8.4. Colors of flags

Source: Own work.

problematic[5]. In that case, it was decided to calculate the average country risk based on country risks of individual countries and assign the average value as the *Country Risk* of a sector.

Final Hazard index

5. There are 476 sectors that cover more than one country.

In the final step, the overall *Hazard* index for each maritime sector is determined (formula (8.4)). The value is calculated based on three factors: *Accidents*, *Piracy* and *Country Risk*. Each factor is additionally weighted according to its importance.

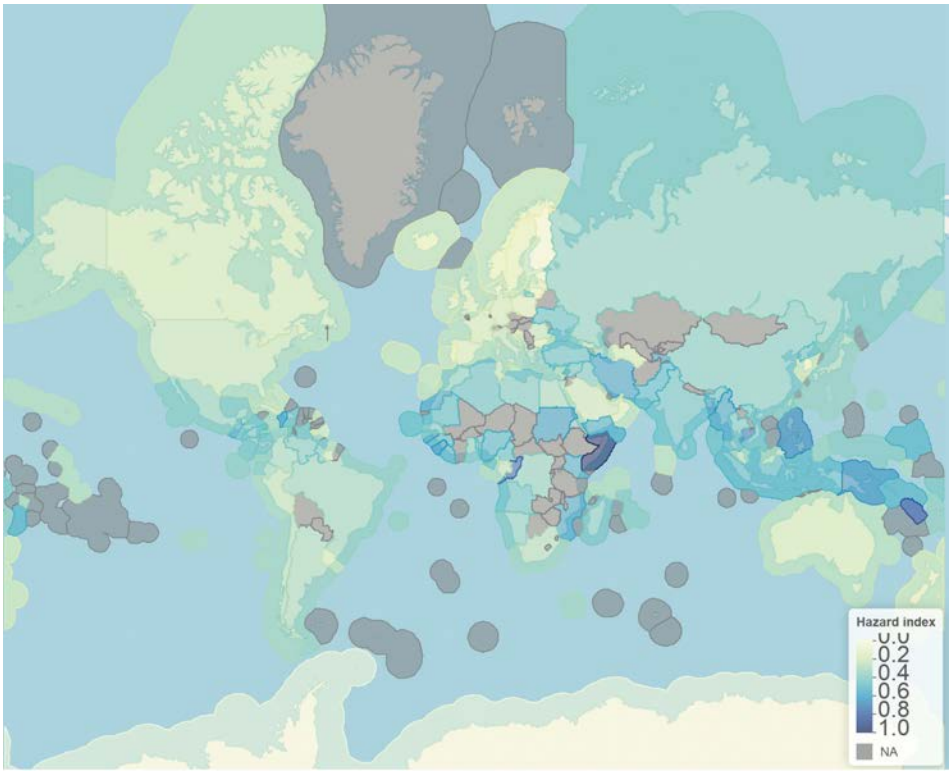


Figure 8.5. Country risk map

Source: Own work.

The weights are subjectively assigned by the author and their value can be adjusted if needed.

$$\text{Hazard}(S)_i = \alpha \times A(S)_i + \beta \times P(S) + \gamma \times CR(S) \tag{8.4}$$

where $\alpha = 0.3$, $\beta = 0.5$, $\gamma = 0.2$, *A—Accident*, *P—Piracy*, *CR—Country risk*, *S—sector*, *i—month*.

Due to the seasonality of the *Accident* measure, the *Hazard* index is calculated for each time period (months in this case) and for each defined maritime area (sector). Having calculated this value, and knowing the predicted route, it is possible to estimate the risk for a given ship's route taking into account the hazards defined in the index.

8.4.3. Weather and sea state

The last factor that may influence a ship's travel time and its punctuality is weather conditions and sea state. The SPP method assumes that unfavorable weather conditions (e.g., heavy rain or wind, dense fog) or weather extremes (e.g., hurricane) on the planned route may lead to an increase of travel time and delays. In the method it was decided to include only information about unusual conditions, and exclude other weather variables (like currents, tides, wind, wave height) that influence the speed of a ship (and the travel time). The decision was made due to the following reasons.

The method for prediction of travel time, presented in Section 8.3, is based on the average speed of ships in different regions and the average travel time. Thus, it might be assumed that this average speed already reflects some average / general weather conditions that prevail in a given region. The authors are aware that this assumption has some drawbacks, however, this aspect is planned to be further developed in future studies. Besides, the weather at sea may change very dynamically so even the existing algorithms that model the relationship between weather conditions and the speed of a ship, have drawbacks and are not very accurate (which was already mentioned in Section 3.4). Besides, many of them are based on averaged values.

For weather conditions, the SPP method includes a similar to one applied for congestion. Having predicted the sequence of sectors a ship will travel through, it is checked whether there are (or are predicted within the time of the ship's voyage) any weather extremes in each of the sectors. Based on (Samson & Ibitoru, 2015), we take the assumption that if a ship encounters an extremely rough weather condition, the captain of the ship can decide to reduce the speed in order to avoid extreme ship motion.

The weather extreme might be determined based on the mean wind speed (Beaufort wind scale) and sea state scale (wave height) in a region. Based on this information, the weather factor that represents the slip of a ship's speed (in %) can be calculated. Here, the approach proposed by (M. M. L. Chen & Chesneau, 2008) can be adopted that creates the possibility to predict how wind and wave values reflect on the slip of a ship's speed. It is presented in Table 8.1. The first row of the presents wind in the Beaufort scale and wave in the Sea State scale. In the second row the probable wave height in meters in given conditions is presented. The last row shows the predicted slip of a ship's speed in % under given conditions.

Table 8.1. Slip of the ship's speed depending on weather conditions

Wind / Wave	2/2	3/3	4/4	5/4	6/5	7/6	8/7	9/7
Wave height [m]	0.5	1.0	2.0	3.5	4.5	5.0	6.0	7.0
Speed slip [%]	8.0	9.0	10.0	12.0	15.0	18.0	23.0	49.0

Source: Based on data in (M. M. L. Chen & Chesneau, 2008).

Using the values from Table 8.1 and the information about the current or predicted weather conditions on the route, it is possible to update the ship's average speed in the sector. To this end, WeatherFactor for each sector is calculated according to Formula (8.5). Finally, the predicted travel time is updated.

$$\mathsf{WeatherFactor}_s = 1 - \mathsf{SpeedSlip}(w/w)_s \tag{8.5}$$

$$\mathsf{WeatherUpdate}_s = \mathsf{AverageSpeed}_s \times \mathsf{WeatherFactor}_s \tag{8.6}$$

where $\mathsf{AverageSpeed}_s$ is the ship's average speed in sector s. $\mathsf{SpeedSlip}(w/w)_s$ is the value of the speed slip depending on the predicted wind / wave in sector s according to the Table 8.1. $\mathsf{WeatherUpdate}_s$ is the updated travel time in sector s.

Although in SPP it was decided to include the weather factor that reflects loss of time in the case of traveling through an area with heavy weather, it should be noted that in practice other approach can also be used. As indicted by (Cai et al., 2014), when sailing in adverse weather conditions, a ship is likely to encounter various kinds of dangerous phenomena which may lead to capsizing or severe roll motions causing damage to cargo, equipment and persons on board. Therefore, instead of sailing through such areas, a captain often just re-designs a ship's route to avoid heavy weather and sea conditions. However, this may not always be possible since the weather may change dynamically during the voyage and such a change may not be possible. This, however, is another area for possible improvements of SPP. In the future, further research is planned to include the option when, in the case of heavy weather, the planned ship's route is updated in order to avoid potentially dangerous areas.

8.4.4. Past delays

The last variable that is included in SPP is the delay factor. This factor reflects the average delay noted in the past on the analyzed voyage. To this end, the past voyages, previously found while determining the predicted route and travel profile (cases 1 and 2), are analyzed. For each past voyage, the difference between the declared ETA at the beginning of the voyage and the actual time of arrival at the destination is determined. Then, the average delay is calculated. First, only voyages for case 1 is taken into account. If case 1 does not prevail, then the voyages from case 2 are used to calculate the average delay.

The delay factor is a numerical value that reflects the average number of hours of delay. It might be positive, which means that on average the ship arrives delayed, or negative, which means that the ship arrives before the declared ETA. It is then used to update the predicted travel time.

8.5. Determination of ship's punctuality

8.5.1. Travel time updates

Finally, having obtained the results provided by all the components of SPP, including the information about the predicted route, the basic travel time from the travel profile, delay, congestion, hazard, and weather factors, it is possible to determine a ship's punctuality.

The overall process of determination of punctuality is depicted in Figure 8.6. The predicted travel time that results from the travel profile is the basis which is further updated according to the order presented in Figure 8.6.

The predicted travel time is calculated based on the past voyages (cases 1–3). We assume here that the travel profile, which is calculated at the beginning of the voyage, is static and does not change during the voyage. Also the predicted route does not change.[6]

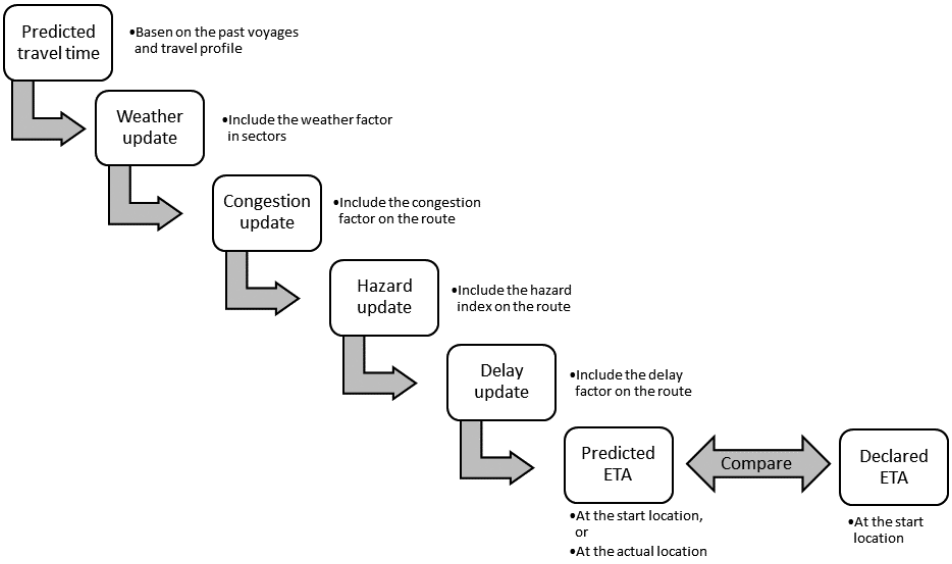


Figure 8.6. Steps of the process of the punctuality determination

Source: Own work.

6. In the case when there is a deviation from the initially predicted route, the SPP method has to be re-started in order to calculate a new predicted route (the current position is treated as the new start position) and a new travel profile, and all the remaining factors have to be re-determined. It basically means that the whole process is repeated, so this case might be treated as a new instantiation of the method for a new start and destination location.

Then, the weather factor is included if any weather extremes are noted. To this end, for each sector of the predicted route the average speed is updated and then the updated travel time is calculated (see Formula (8.5)).

In the next step, the congestion factor is taken into account. For each sector, in which congestion appears, the travel time is updated. Depending on the value of the congestion factor (CongestionFactor), the travel time is proportionally decreased (when the current density is significantly lower than the average), or increased (when the current density is significantly higher) according to Formula (8.7).

$$\begin{aligned} \mathsf{CongestionUpdated} &= \forall_s \mathsf{CongestionUpdated}_s \\ &= \mathsf{WeatherUpdated}_s \cdot \mathsf{CongestionFactor}_s \end{aligned} \tag{8.7}$$

where CongestionUpdated—updated travel time on the route; $\mathsf{CongestionUpdated}_s$—updated travel time in sector s; $\mathsf{WeatherUpdated}_s$—travel time in sector s with weather factor included; $\mathsf{CongestionFactor}_s$—congestion in sector s.

Then, the information about a potential hazard on the route is included. Here, a threshold on the accepted risk level should be defined by a user. Then, the hazard index for each sector is compared with the threshold (the difference is calculated) to determine a travel time update rate. HazardUpdateRate depends on the difference:[7]

- Difference: $\langle 0 - 0.25)$ -> update rate: 1%.
- Difference: $\langle 0.25 - 0.5)$ -> update rate: 2%.
- Difference: $\langle 0.5 - 0.75)$ -> update rate: 3%.
- Difference: $\langle 0.75 - 1\rangle$ -> update rate: 4%.

The updated travel time is calculated according to Formula (8.8).

$$\begin{aligned} \mathsf{HazardUpdated} &= \forall_s \mathsf{HazardUpdated}_s \\ &= \mathsf{CongestionUpdated}_s \cdot (1 + \mathsf{HazardUpdateRate}_s) \end{aligned} \tag{8.8}$$

where HazardUpdated—updated travel time on the route; $\mathsf{HazardUpdated}_s$—updated travel time in sector s; $\mathsf{CongestionUpdated}_s$—travel time in sector s with congestion factor included; $\mathsf{HazardUpdateRate}_s$—hazard update rate depending on the difference between the hazard index and the risk threshold.

Finally, the influence of the delay factor is taken into account. Here, the total travel time is increased or decreased by adding or subtracting the number of hours resulting from the delay factor (Formula (8.9)).

$$\mathsf{DelayUpdated} = \mathsf{HazardUpdated} + \mathsf{DelayFactor} \tag{8.9}$$

7. The values of the update rates are proposed by the authors and can be adjusted if needed.

where DelayUpdated—updated travel time on the route; HazardUpdated—travel time on the route with hazard factor included; DelayFactor—average delay in hours.

8.5.2. ETA prediction

The calculated travel time (DelayUpdated) is the final value that is used to determine the predicted ETA (Formula (8.10)). To this end, this predicted travel time is added to the timestamp, which is either a point in time when the voyage starts or a point in time during the voyage. Thus, determination of the ETA can be conducted at the beginning of a ship's voyage (as a planning tool) and further on when the ship is already under way (based on the current time and location of the ship) in the form of regular updates on the arrival time (a monitoring tool).

$$\text{PredictedTravelTime} = \text{DelayUpdated} \tag{8.10}$$

$$\text{PredictedETA} = \text{TimestampStart} + \text{PredictedTravelTime} \tag{8.11}$$

where DelayUpdated—updated travel time for the route; PredictedTravelTime——final prediction of the travel time for the route; TimestampStart—a point in time for which the analysis is conducted.

In order to determine the ship's punctuality, the ETA declared by the captain at the beginning of the route (provided in AIS) can be compared with the ETA predicted based on the proposed method. If a significant deviation between these two values occurs,[8] it can be reasoned that the ship will not arrive punctually at the destination. The result of SPP allows for determination whether the ship will arrive earlier or later than the declared ETA.

All in all, the SPP method was designed and developed based on the analysis of the existing methods for prediction of a ship's route and determination of travel time and the estimated time of arrival. Besides, the identified gaps and challenges that still might be addressed in this area were taken into account. The final concept of the method also took into consideration the available data sources.

To sum up, the main characteristics of the SPP methods are:

- It relates to a given voyage and to an individual ship.
- It focuses on the determination of a single characteristic of a ship's voyage—the punctuality.
- It consists of six components: route prediction, travel time profile, congestion factor, hazard index, weather and sea state factor, and average delay factor.
- The results provided by each of the components are fused to determine the final prediction of a ship's arrival time.

8. A significant deviation means that the calculated difference is higher than a defined threshold.

- The estimation of a ship's punctuality can be conducted in two forms; it may be calculated at the beginning of the ship's voyage (as a planning tool) or when the ship is already under way to provide updates on the travel time (as a monitoring tool).
- Prediction of a route includes three alternatives (cases) that are checked 1) past routes of a given ship on the same voyage; 2) past routes of other ships on the same voyage; 3) past routes of other ships on a similar voyage (traveling nearby a given destination or to a nearby port).
- Punctuality is determined based on a comparison of the predicted travel time and the travel time declared by a ship.
- Basically, the method is dedicated to the punctuality prediction of merchant ships. However, it might be used also for other types of ships.

The estimations provided by SPP may be utilized by different entities (actors) from the maritime domain, since they might be a valuable information needed in the decision making process, such as monitoring the course of the delivery process and planning further actions when the ship finally arrives at a port. Besides, thanks to SPP ships which probably will be delayed may be quickly identified. This information might be potentially interesting for logistic companies, senders and recipients of goods, port authorities, etc. Thus, a potential application and exploitation of SPP seems to be very wide because it might be incorporated in the existing maritime and logistic systems.

8.6. Application of the SPP method—an example

In order to perform the evaluation of the SPP method and show its applicability, it was implemented and tested using real maritime data and for selected examples of real ships' voyages. Similarly to the MRRAM example (see Section 7.3), the process started with selection of real world examples of past voyages between 25 different European ports. In total, a set of 255 voyages was collected. It was further divided into a training set (consisting of 229 voyages), and a testing set (26 voyages).

The evaluation process consisted of a few steps. In the first step, the algorithms for route prediction, presented in Section 8.2, were verified and tested. To this end, the set of selected destination ports and starting points was used. They are presented in Table 8.2.

According to Algorithms 8.1 and 8.2, three types of routes (three cases) were checked: 1) routes of a given ship on the same voyage; 2) routes of other ships on the same voyage; 3) routes of other ships on similar voyages. Table 8.2 presents the total number of routes that were found from a given starting location for all

Table 8.2. Route prediction method—summary of the evaluation results

Voyage number	Destination port	Start location		Ships in start location	Ships arrived at port	Correctness
		lat	long			
1	Bilbao	30.00	32.50	23	5	Yes
2	Bilbao	35.90	−5.60	128	29	Yes
3	Teesport	54.59	12.28	47	8	Yes
4	Montoir	52.37	3.40	36	7	Yes
5	Kambo	55.90	17.30	9	2	Yes
6	Muuga	51.40	2.00	56	23	Yes
7	Livorno	16.15	41.30	21	5	No
8	Valencia	51.90	4.00	35	10	Yes
9	Algeciras	16.15	41.30	53	13	Yes
10	Istanbul	16.15	41.30	61	13	No
11	La Spezia	36.10	−5.40	9	3	Yes
12	Le Havre	51.90	3.70	24	3	Yes
13	Mersin	12.60	43.10	6	2	No
14	Marseille	36.00	−5.50	15	6	Yes
15	Gdansk	36.00	−4.55	31	6	Yes
16	Gothenburg	40.50	1.75	18	3	Yes
17	Fos	35.92	−6.25	136	30	Yes
18	Aliaga	36.15	−4.70	111	27	Yes
19	Tenerife	51.50	3.01	81	10	Yes
20	Swinoujscie	51.50	2.14	26	8	Yes
21	Sines	54.50	10.50	43	9	Yes
22	Barcelona	35.50	25.50	16	2	Yes
23	Piraeus	36.00	−4.70	57	12	Yes
24	Genoa	52.40	4.00	175	13	Yes
25	Felixstowe	54.00	7.75	26	9	Yes
26	Lysekil	52.40	3.20	69	6	Yes

Source: Own work.

three conditions (ships that were heading to a destination or nearby ports from a given location). However, of all the identified routes only for some of them it was possible to determine that (and when exactly) a ship actually arrived at the port (see column ships that arrived at port).

For all the identified ships, their trajectories from the starting location to the destination port were found (the set of sectors that were followed, see Al-

gorithm 8.3). Then, from the list of identified ships, a single ship was chosen as the reference voyage (the reference ship). Finally, based on the trajectories the most probable route for the reference ship was determined.

Each found route was then analyzed based on the created visualization—it was verified whether it seems reasonable (creates a reasonable series), whether there are any gaps or errors.

8.6.1. Data sources and infrastructure

The evaluation of the SPP method was performed using same set of data and infrastructure as in the evaluation of the MRRAM method (see section 7.3). Table 8.3 presents some statistics regarding the time of data analysis for the different steps of the ship's route prediction method as well as the received results (the number of found AIS messages or vessels). The statistics show that for each port selected in the analysis on average 313 unique ships that were heading to this port were found (ships that declared in AIS that they are heading to a given destination). However, this value differs significantly between ports—there were ports with >500 unique ships as well as ports for which this value was <100. Besides, for each port over 55 additional, unique ships were found that were heading to other ports located nearby a given port (Case 3). To identify all these ships, the one-year set of AIS data was analyzed, and for each port it took about 360 seconds.

After identification of ships that were heading to a given port, their trajectories had to be identified (all AIS messages sent by a ship during a voyage to a destination). This process lasted about 460 seconds on average for each destination port. As a result, a set of 3.4 mln dynamic and 1.4 mln static messages were found and grouped into ships' trajectories.

8.6.2. Analysis results

Exemplary results of the route prediction step are presented in Figure 8.7. The route to Istanbul is an example of a not-fully correct route[9]—there are some redundant sectors that probably might be deleted from the predicted route. Detailed results for other ports are presented in Table 8.2 and in Appendix B.

The results of the route prediction step can be summarized as follow:

- In total 26 routes were identified, among them:
  - 23 routes were correct (created a reasonable series of sectors).
  - 3 routes included missing or redundant sectors.

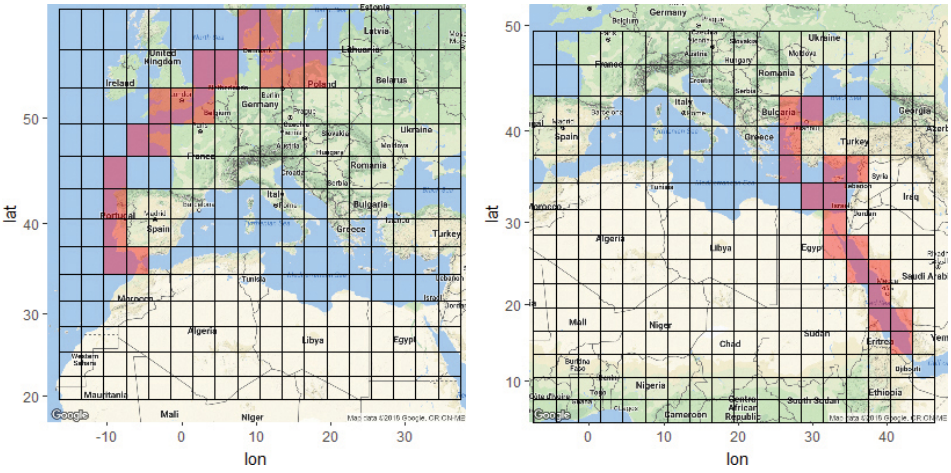
9. It may result from the fact that the ship called another port on its way but did not provide this information in AIS.

Table 8.3. Statistics on data analytics for the selected steps of a ship's route prediction using Microsoft Azure

Step	Avg processing time [sec]	Avg no of ships	
Finding unique ships sailing to a given destination	360.22	313.27	
Finding unique ships sailing to nearby ports	323.86	55.35	
Step	**Avg processing time [sec]**	**Avg no of static msg**	**Avg no of dynamic msg**
Finding and fusing AIS data for ships sailing to a given destination	459.30	1,422,242	3,401,111
Finding and fusing AIS data for ships sailing to nearby ports	416.29	228,558	471,363

Source: Own work.

- Thus, correctness of the route prediction method amounts to 88%.
- Identified routes might be grouped from the point of view of their length:
  - Long voyages (>10 sectors)—14 routes (54%).
  - Medium voyages (between 6 and 10 sectors)—9 routes (35%).
  - Short voyages (up to 5 sectors)—3 routes (11%).



(Route 1—Gdansk) (Route 2—Istanbul)

Figure 8.7. A visual presentation of the identified routes—examples

Source: Own work.

Having identified the routes, the next step was determination of travel time between the starting point and the destination port. Here two approaches were tested, described in Section 8.3. In the first approach, the travel time was calculated based on historical routes of the reference ship (Case 1) and other ships (Case 2 and 3). In the second approach, travel time was calculated based on the distance and the average speed in sectors. Then, the average historical travel time was calculated based on the average travel time for Cases 1–3 and the travel time for Approach 2 (Table 8.4). The results were then summarized in a form of a travel profile from the start location to the destination port (see the example of travel profiles in Figure 8.8; the profiles for the rest of routes are presented in Appendix B).

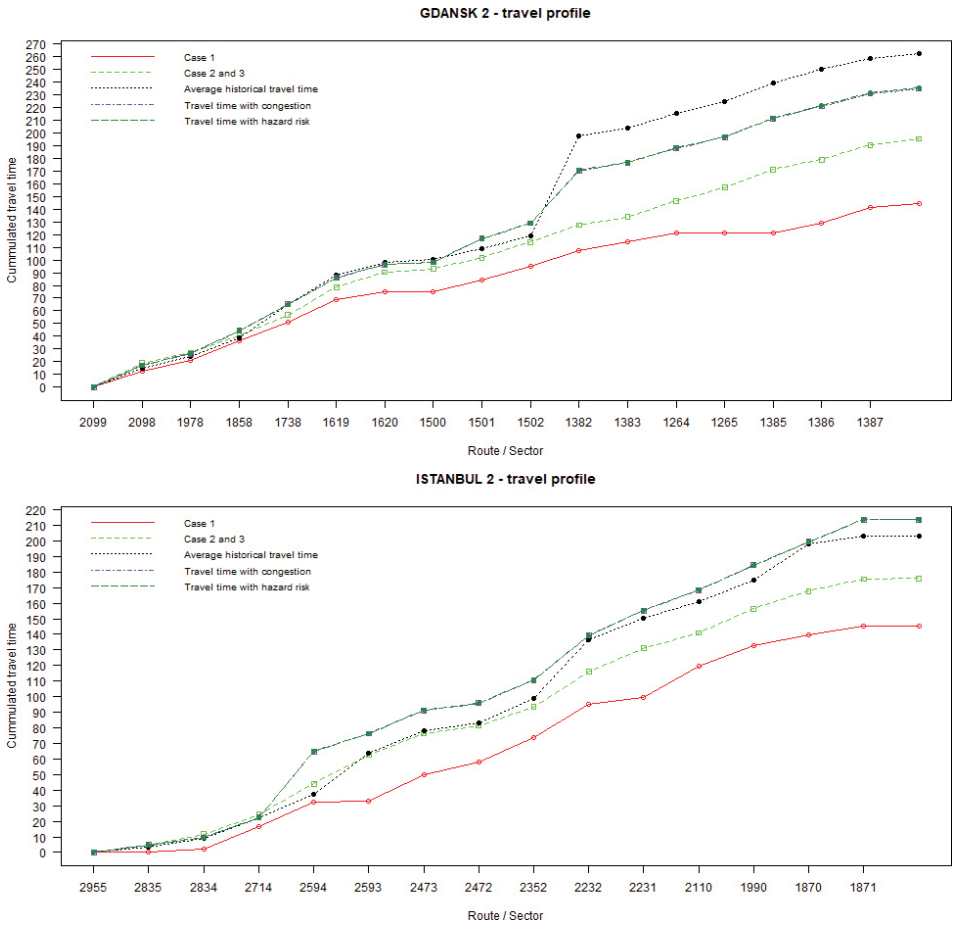


Figure 8.8. Travel profiles for selected voyages

Source: Own work.

Table 8.4. Travel time prediction—summary of the evaluation results

Voyage number	Port	Predicted travel time [hours]				
		Approach 1			Approach 2	Avg travel time
		Case 1	Case 2 and 3	Avg cases 1–3		
1	Bilbao	229.8919	389.2504	309.5712	470.3568	389.9640
2	Bilbao	76.3104	94.8073	85.5588	111.0234	98.2911
3	Teesport	9.8103	78.7788	44.2946	109.7463	77.0204
4	Montoir	38.8156	56.9991	47.9073	119.6381	83.7727
5	Kambo	37.7893	46.8976	42.3434	63.0943	52.7189
6	Muuga	88.5210	121.5447	105.0328	176.9814	141.0071
7	Livorno	172.2152	297.9221	235.0687	472.9906	354.0296
8	Valencia	13.8359	162.5091	88.1725	270.5219	179.3472
9	Algeciras	173.0669	0.0000	86.5335	382.5007	234.5171
10	Istanbul	145.1949	175.8303	160.5126	302.9094	231.7110
11	La Spezia	74.9548	91.9946	83.4747	80.4289	81.9518
12	Le Havre	18.4838	24.7529	21.6184	70.6590	46.1387
13	Mersin	110.1576	234.3767	172.2672	335.6779	253.9725
14	Marseille	66.9062	81.4094	74.1578	96.8173	85.4875
15	Gdansk	144.1444	195.0745	169.6095	316.6610	243.1352
16	Gothenburg	194.4836	205.2322	199.8579	319.7686	259.8133
17	Fos	57.6900	68.1274	62.9087	87.4894	75.1991
18	Aliaga	117.7939	158.7774	138.2857	165.3889	151.8373
19	Tenerife	143.1447	159.6421	151.3934	252.3866	201.8900
20	Swinoujscie	56.3258	56.0069	56.1664	103.1217	79.6440
21	Sines	118.8397	149.7774	134.3086	226.4771	180.3928
22	Barcelona	91.8413	92.1793	92.0103	129.5898	110.8000
23	Piraeus	95.1448	144.4493	119.7971	146.2259	133.0115
24	Genoa	199.8234	228.6362	214.2298	374.9256	294.5777
25	Felixstowe	24.8071	35.9163	30.3617	66.6322	48.4969
26	Lysekil	56.7510	47.6317	52.1913	78.7834	65.4874

Source: Own work.

The predicted travel times were then gradually updated, taking into account the additional information about the predicted route—the calculated congestion factor and the hazard index (described in detail in the next subsections) as well as

the information about past delays on the route. This step was conducted using the approach presented in Section 8.5.

Having updated the travel time, it was possible to finally determine the predicted ETAs for all the reference ships and compare them with the ETAs declared by the captains in AIS at the starting point. The first experiment assumed prediction of a ship's punctuality at the beginning of the route. The second experiment focused on the possibility to update the ETA and determine punctuality when a ship is already under way to the destination.

The results of the ETA prediction are presented in Table 8.5. The table provides the following information:

- Timestamp for which prediction was conducted (column "start time").
- ETA declared by the captain at the starting point (column "declared ETA").
- ETA predicted based on SPP calculations (column "predicted ETA").
- Actual arrival time at the destination port, derived from AIS messages based on a ship's localization (column "arrival time").
- Information if, according to the comparison of the declared and the predicted ETA, the ship will be delayed (column "forecast delay").
- Information if the ship was actually delayed, by comparing the declared ETA and the actual arrival time (column "actual delay").
- Accuracy of the estimation of arrival time for the declared ETA (column "accuracy ais").
- Accuracy of the estimation of arrival time for the predicted ETA (column "accuracy SPP method').
- Information whether the accuracy of the predicted ETA is better than for the declared ETA (column "better accuracy").

The accuracy values both for the declared and the predicted ETA were calculated as a difference between the ETA and the actual arrival time (the difference is provided in hours). It means that the lower the value, the better the accuracy of estimation. The negative value means that the ship arrived before the declared/predicted ETA (arrived earlier), while the positive value means that the ship arrived after the declared/predicted ETA (was delayed). While determining which ETA (declared or predicted) is better the absolute value of the difference was taken into account; it means that the closer the ETA to the actual arrival time, the better (no matter whether ETA was before or after the actual arrival time).

It might be mentioned that during the research a few experiments that aimed at testing different variations of ETA prediction were conducted—the ETA was predicted taken into account different values of the predicted travel times (Table 8.4). For each such variation, the accuracy of ETA predictions was calculated and compared. Finally, the basic travel time with the best accuracy was selected. In general, in 62% of voyages the best accuracy of ETA prediction was received when the value of the average travel time was used (the average of travel time for Cases

1–3 and Approach 2). Therefore, it was decided that the average travel time should be used as the basis for further updates of travel time (when additional factors are included) and for the final prediction of the ETA.

The analysis of the results shows that for 22 out of 26 voyages the accuracy of the predicted ETA is better than the accuracy of the declared ETA. In other words, it means that in 88.42% of cases the estimation of the ETA using the SPP method was better than the ETA provided by the captain of the ship (in a given timestamp). This result proves the effectiveness and usefulness of the SPP method in determining the predicted time of arrival at a given destination port.

Another aspect worth-mentioning is the ability of the SPP method to appropriately predict whether a ship will be delayed or not (in this case, by delay we mean that the ship arrives after the declared ETA)—see columns "forecast" and "actual delay" in Table 8.5. The values in both columns overlap in 23 cases. It means that in 88.35% of cases the results of the SPP method correctly predicted whether the ship will or will not be delayed.

8.6.3. Congestion results

One of the components of the SPP method is the *Congestion* factor on the predicted route. As described in Section 8.4.1, it requires calculation of the average monthly ships density and the actual ships density.

During the research, both the monthly density in all sectors and the actual density in the last 24 hours preceding the starting point of the voyage were calculated. First, ships' movements based on AIS were analyzed—here both the period of one year was used to calculate the monthly density ratio and the last 24 hours to calculate the actual density ratio for each voyage. Then, for all the sectors of the predicted route the *Congestion* factor was calculated, taking into account the difference between the actual and the average density. Finally, for each sector the predicted travel time was updated taking into account the *Congestion* factor. Figure 8.9 visualizes the results of the actual density calculation for the selected voyages and indicates in which sectors the congestion occurs (the black sectors).

Having acquired the information about the congestion on the predicted route, it was possible to evaluate how this information influences the accuracy of the results provided by SPP. In other words, we wanted to check whether, in general, the information about congestion should be taken into account to update the predicted travel time, and thus to calculate the predicted ETA.

To this end, an experiment was conducted in which we analyzed and compared the predicted ETA and the accuracy of the SPP method when the information about congestion is either included or excluded. The results of this experiment are presented in Table 8.6. The table compares travel times, predicted ETAs and

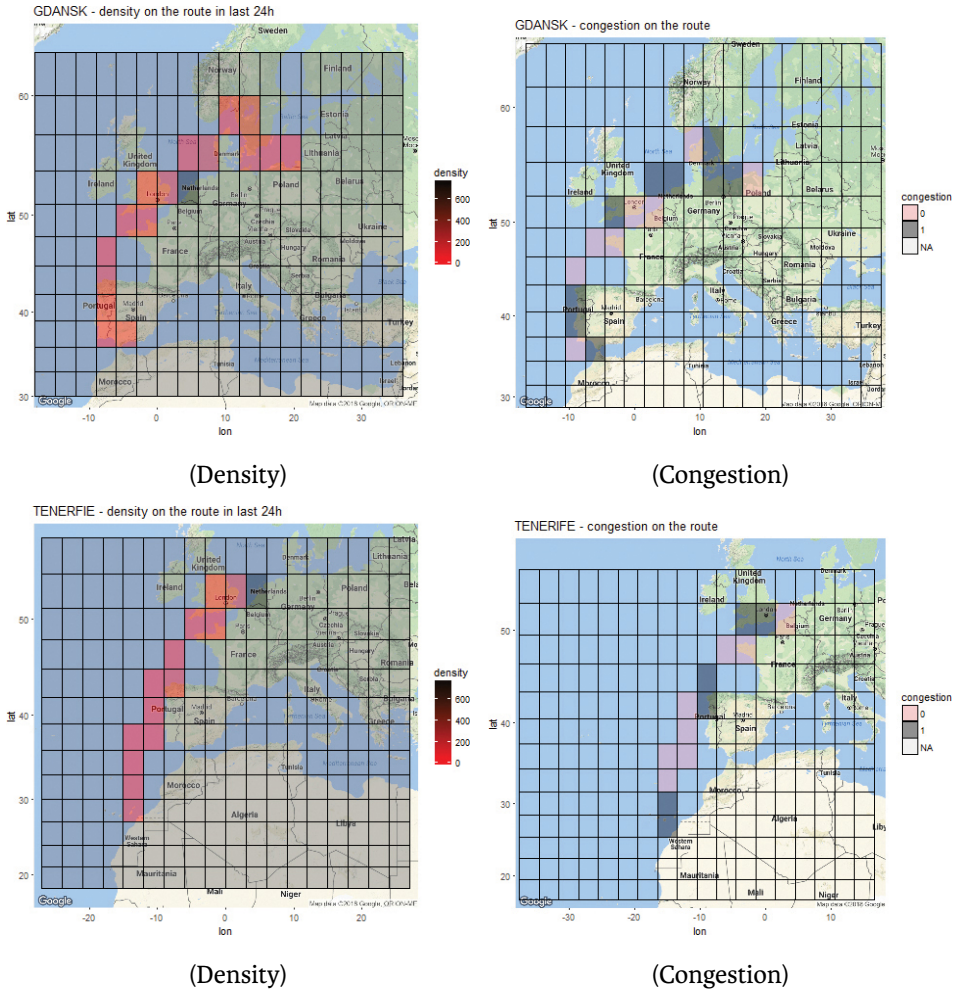


Figure 8.9. Actual traffic density and congested sectors for selected voyages

Source: Own work.

the accuracy of the method before and after inclusion of the *Congestion* factor. The results show that in the case of 18 voyages the accuracy of the predicted ETA improved when the information about the actual congestion has been added, for 2 voyages the accuracy did not changed (because there was no congestion noted on the route), and in the case of 8 voyages the accuracy was not improved. It means that inclusion of information about congestion improved the accuracy in 69% of the cases.

Moreover, in the cases with improved accuracy it increased on average by 7.43 hours, while in those cases were the accuracy was not improved it decreased on average only by 2.41 hours. As a result, it might be concluded that it seems that the

Table 8.5. Ship's punctuality prediction—summary of the evaluation results

Voyage number	Port	Reference ship	Start time	Declared ETA	Predicted ETA	Arrival time	Forecast delay	Actual delay	Accuracy ais	Accuracy SPP method	Better accuracy
1	Bilbao	309072000	28.04.2015 00:50	08.05.2015 20:00	09.05.2015 07:50	08.05.2015 06:59	Yes	No	13.0069	24.1500	FALSE
2	Bilbao	477547300	15.05.2015 23:21	19.05.2015 00:01	18.05.2015 19:21	18.05.2015 20:02	No	No	3.9797	−0.6856	TRUE
3	Teesport	255804960	18.05.2015 01:01	19.05.2015 18:00	20.05.2015 06:01	20.05.2015 11:15	Yes	Yes	−17.2603	−5.2325	TRUE
4	Montoir	246823000	25.04.2015 22:32	28.04.2015 06:00	28.04.2015 04:32	28.04.2015 04:16	No	No	1.7200	0.2547	TRUE
5	Kambo	311000237	17.04.2015 19:20	19.04.2015 14:00	19.04.2015 13:20	19.04.2015 08:55	No	No	5.0803	1.4161	TRUE
6	Muuga	212118000	26.02.2015 16:10	02.03.2015 03:00	02.03.2015 02:10	01.03.2015 23:23	No	No	3.6114	2.7867	TRUE
7	Livorno	258802000	01.02.2015 12:54	09.02.2015 08:00	12.02.2015 00:54	12.02.2015 15:37	Yes	Yes	−79.6211	−14.7047	TRUE
8	Valencia	636014997	27.04.2015 22:48	01.05.2015 21:30	04.05.2015 14:48	01.05.2015 19:41	Yes	No	1.8006	45.1106	FALSE
9	Algeciras	257588000	26.04.2015 12:00	10.05.2015 04:00	09.05.2015 15:00	09.05.2015 19:56	No	No	8.0633	−4.9367	TRUE
10	Istanbul	353992000	12.05.2015 18:00	24.05.2015 11:00	19.05.2015 15:00	22.05.2015 10:54	No	No	48.0883	−24.9117	TRUE
11	La Spezia	548765000	22.04.2015 18:06	27.04.2015 02:00	27.04.2015 05:06	27.04.2015 05:55	Yes	Yes	−3.9175	−0.8053	TRUE
12	Le Havre	249701000	22.01.2015 20:28	04.02.2015 15:00	23.01.2015 16:28	24.01.2015 17:05	No	No	261.9131	−24.6100	TRUE
13	Mersin	538004227	31.03.2015 23:16	07.04.2015 07:30	05.04.2015 10:16	07.04.2015 04:09	No	No	3.3369	−41.8803	FALSE
14	Marseille	10318000	18.05.2015 07:09	21.05.2015 08:00	20.05.2015 23:09	21.05.2015 02:04	No	No	5.9289	−2.9078	TRUE
15	Gdansk	565978000	18.04.2015 15:04	13.05.2015 22:00	28.04.2015 22:04	30.04.2015 20:20	No	No	313.6639	−46.2603	TRUE
16	Gothenburg	212722000	19.01.2015 15:30	16.02.2015 21:00	27.01.2015 17:30	27.01.2015 18:31	No	No	482.4719	−1.0122	TRUE
17	Fos	372016000	21.05.2015 07:00	23.05.2015 17:30	23.05.2015 10:00	23.05.2015 13:26	No	No	4.0656	−3.4344	TRUE
18	Aliaga	538090282	29.04.2015 07:00	04.05.2015 13:00	04.05.2015 04:00	04.05.2015 08:09	No	No	4.8439	−4.1561	TRUE
19	Tenerife	538004027	14.05.2015 06:00	20.05.2015 05:00	20.05.2015 03:00	20.05.2015 01:35	No	No	3.4153	1.4153	TRUE
20	Swinoujscie	538003236	07.04.2015 18:00	11.04.2015 22:00	10.04.2015 11:00	11.04.2015 03:52	No	No	18.1219	−16.8781	TRUE
21	Sines	247226900	28.04.2015 06:00	03.05.2015 16:00	03.05.2015 12:00	03.05.2015 12:09	No	No	3.8406	−0.1594	TRUE
22	Barcelona	371426000	04.05.2015 12:00	08.05.2015 15:00	08.05.2015 20:00	08.05.2015 13:03	Yes	No	1.9361	−4.0639	FALSE
23	Piraeus	636091659	06.02.2015 03:00	10.02.2015 06:00	10.02.2015 01:00	10.02.2015 02:11	No	No	3.8142	−1.1858	TRUE
24	Genoa	220620000	13.04.2015 04:00	21.04.2015 04:00	21.04.2015 02:00	20.04.2015 20:12	No	No	7.7875	5.7875	TRUE
25	Felixstowe	354625000	01.05.2015 14:00	04.05.2015 03:15	04.05.2015 04:00	04.05.2015 04:22	Yes	Yes	−1.1239	−0.3739	TRUE
26	Lysekil	309553000	28.01.2015 12:00	01.02.2015 00:01	30.01.2015 15:00	31.01.2015 06:38	No	No	17.3747	−7.6419	TRUE

Source: Own work.

Table 8.6. Accuracy of the SPP method with and without congestion—comparison

Voyage number	Port	Travel time no congestion	Travel time congestion	Predicted ETA no congestion	Predicted ETA congestion	Accuracy no congestion	Accuracy congestion	Improved accuracy
1	Bilbao	229.8919	233.5327	07.05.2015 03:50	07.05.2015 06:50	−27.15	−24.15	yes
2	Bilbao	75.9948	78.2601	18.05.2015 19:21	18.05.2015 21:21	−0.69	1.31	no
3	Teesport	69.8170	71.2325	21.05.2015 02:01	20.05.2015 17:01	14.77	5.77	yes
4	Montoir	92.9817	94.1355	28.04.2015 04:32	28.04.2015 04:32	0.25	0.25	yes
5	Kambo	47.1174	47.9180	19.04.2015 05:20	19.04.2015 10:20	−3.58	1.42	yes
6	Muuga	88.5210	90.2378	02.03.2015 02:10	02.03.2015 03:10	2.79	3.79	no
7	Livorno	324.2825	331.8735	16.02.2015 01:54	15.02.2015 01:54	82.30	58.30	yes
8	Valencia	158.9274	162.2836	05.05.2015 03:48	03.05.2015 16:48	80.11	45.11	yes
9	Algeciras	319.1557	326.6609	09.05.2015 15:00	09.05.2015 22:00	−4.94	2.06	yes
10	Istanbul	206.7954	213.6645	21.05.2015 01:00	21.05.2015 10:00	−33.91	−24.91	yes
11	La Spezia	109.2900	111.5323	28.04.2015 19:06	26.04.2015 14:06	37.19	−15.81	yes
12	Le Havre	45.2466	45.2466	23.01.2015 16:28	23.01.2015 16:28	−24.61	−24.61	yes
13	Mersin	110.1576	112.7425	05.04.2015 10:16	05.04.2015 12:16	−41.88	−39.88	yes
14	Marseille	66.9062	68.7026	20.05.2015 23:09	21.05.2015 01:09	−2.91	−0.91	yes
15	Gdansk	261.8978	266.8385	28.04.2015 22:04	29.04.2015 03:04	−46.26	−41.26	yes
16	Gothenburg	194.4836	198.2360	27.01.2015 17:30	27.01.2015 20:30	−1.01	1.99	no
17	Fos	57.6900	58.7268	23.05.2015 10:00	23.05.2015 11:00	−3.43	−2.43	yes
18	Aliaga	117.7939	121.8878	04.05.2015 04:00	04.05.2015 08:00	−4.16	−0.16	yes
19	Tenerife	143.1447	146.1490	20.05.2015 03:00	20.05.2015 06:00	1.42	4.42	no
20	Swinoujscie	81.4731	82.8131	09.04.2015 22:59	09.04.2015 18:00	−28.88	−33.88	no
21	Sines	129.5788	132.2236	03.05.2015 12:00	03.05.2015 15:00	−0.16	2.84	no
22	Barcelona	198.4600	106.1483	08.05.2015 07:00	08.05.2015 09:00	14.94	−4.06	yes
23	Piraeus	95.1448	97.0534	10.02.2015 01:00	10.02.2015 03:00	−1.19	0.81	yes
24	Genoa	199.8234	203.7786	21.04.2015 02:00	21.04.2015 06:00	5.79	9.79	no
25	Felixstowe	62.1834	63.1476	03.05.2015 07:00	03.05.2015 04:59	−21.37	−23.37	no
26	Lysekil	56.7510	58.8262	30.01.2015 15:00	30.01.2015 17:00	−15.64	−13.64	yes

Source: Own work.

Congestion factor should be taken into account in the calculation of the predicted ETA. Thus, it is justified that the *Congestion* factor is a part of the SPP method.

8.6.4. Hazard results

Another element included in the evaluation of the SPP method is the information about the hazard on the predicted route. This required determination of the *Hazard* index, using the concept presented in Section 8.4.2.

In this section the results of the *Hazard* index calculation based on the proposed method is presented. The *Hazard* index was determined based on 3 factors: *Accidents*, *Piracy* and *Country risk*. The index was calculated for a given time period (month) as well as for a sector. Table B1 in Appendix B presents the values of particular factors for selected maritime areas and selected time slots.

Then, we obtained the *Hazard* index for a given ship's route/voyage by calculating an average *Hazard* index for all the sectors the ships were sailing through. Moreover, to see the dispersion of *Hazard* levels between maritime areas also a standard deviation was calculated.

The final *Hazard* indexes are presented in Table B2 in Appendix B. Figure 8.10 presents a map of the values of the *Hazard* index for selected maritime sectors.

Finally, three types of experiments were conducted.

The first two experiments were conducted with no relation to the previously presented examples of voyages. Their aim was to evaluate the general usefulness and conformance of the *Hazard* index in the process of route planning.

The first experiment assumed calculation of the *Hazard* index related to the maritime areas a ship plans to sail through in a given time. As a result, the total *Hazard* index for the voyage was calculated along with the information whether a defined risk threshold was exceeded. Then, the planned route was presented on a map with an indication of hazard levels.

In the second experiment, for a given ship two alternative routes to a given destination were planned and using the *Hazard* index a less dangerous route for a given voyage was recommended.

The third experiment concerned the same examples of voyages that had already been used in the evaluation process and aimed at evaluating how the information about potential hazard on the predicted route influences the accuracy of the SPP method.

Experiment 1. Below we present the results of the first experiment that was conducted based on some illustrative examples for two selected ships. For each ship, a planned route and travel period were simulated and a risk threshold was defined. In Table 8.7 and Figures 8.11 and 8.12 we present the results obtained from the input parameters. Having obtained this information, a user can foresee potential

hazards on the planned route as well as see whether a defined risk threshold is exceeded.

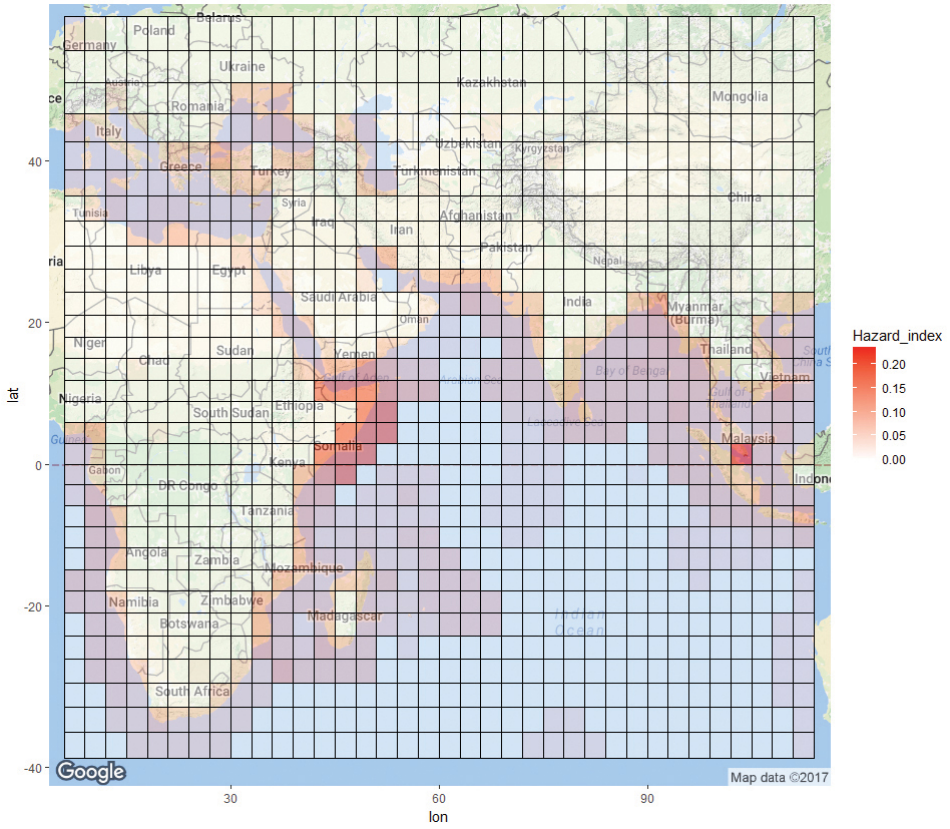


Figure 8.10. Hazard index for selected maritime regions

Source: Own work.

Table 8.7. Experiment 1—results

	Ship 1	Ship 2
Planned route	Rotterdam–Goteborg	Mumbay–Pireneus
Voyage month	April	January
Risk threshold	0.1	0.15
Hazard index for route (avg)	0.0625	0.0716
Hazard index for route (std)	0.0399	0.0335
Threshold exceeded	No	Yes (6 sectors)

Source: Own work.

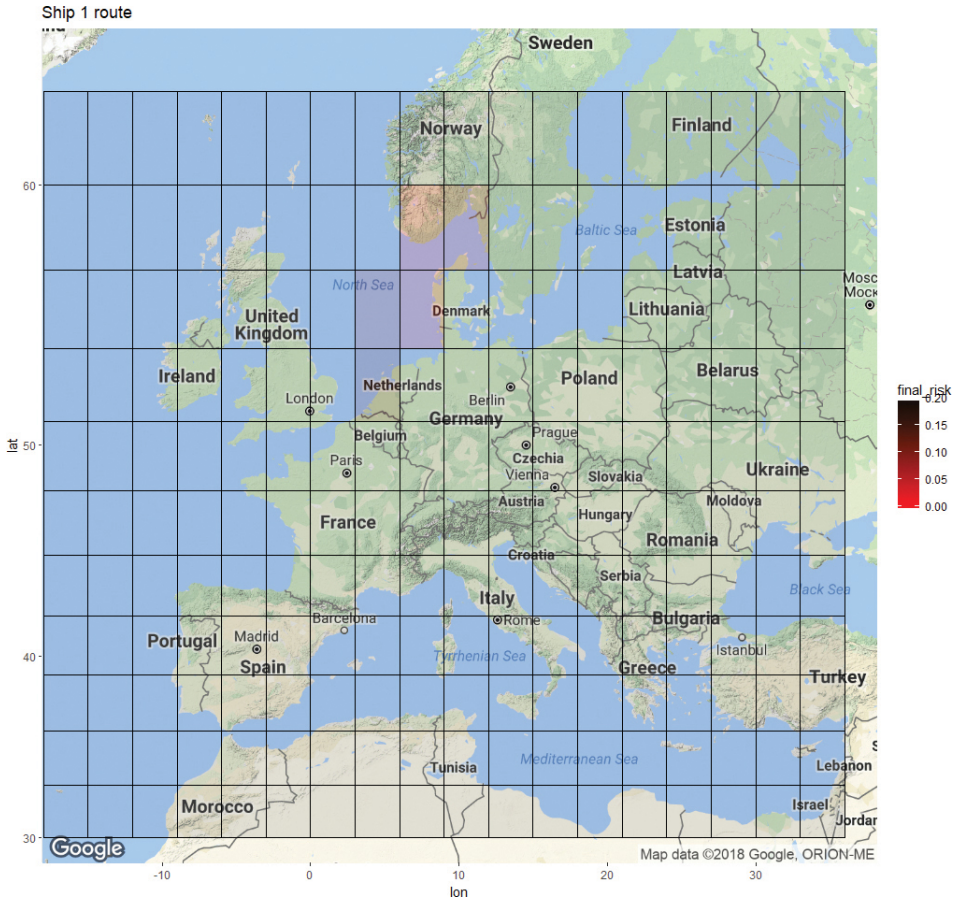


Figure 8.11. Hazard indexes for maritime areas ships are sailing through—Ship 1

Source: Own work.

The results were additionally checked for compliance with what had been observed in reality. In the case of ship 1, the method correctly assigned a higher *Hazard* index for regions near the Netherlands, which is connected with the relatively high number of recorded accidents and a higher *Country risk* in comparison to areas near Denmark or Sweden. In the case of ship 2, the high *Hazard* index near the coast of Somalia is related to the high number of *Piracy* attacks and the *Country risk* index.

Experiment 2. The second experiment was conducted based on an example of a ship that plans its voyage and considers two alternative routes. In Table 8.8 and Figure 8.13 we present the obtained results. They show that although both routes go through potentially dangerous maritime areas, the first route, which

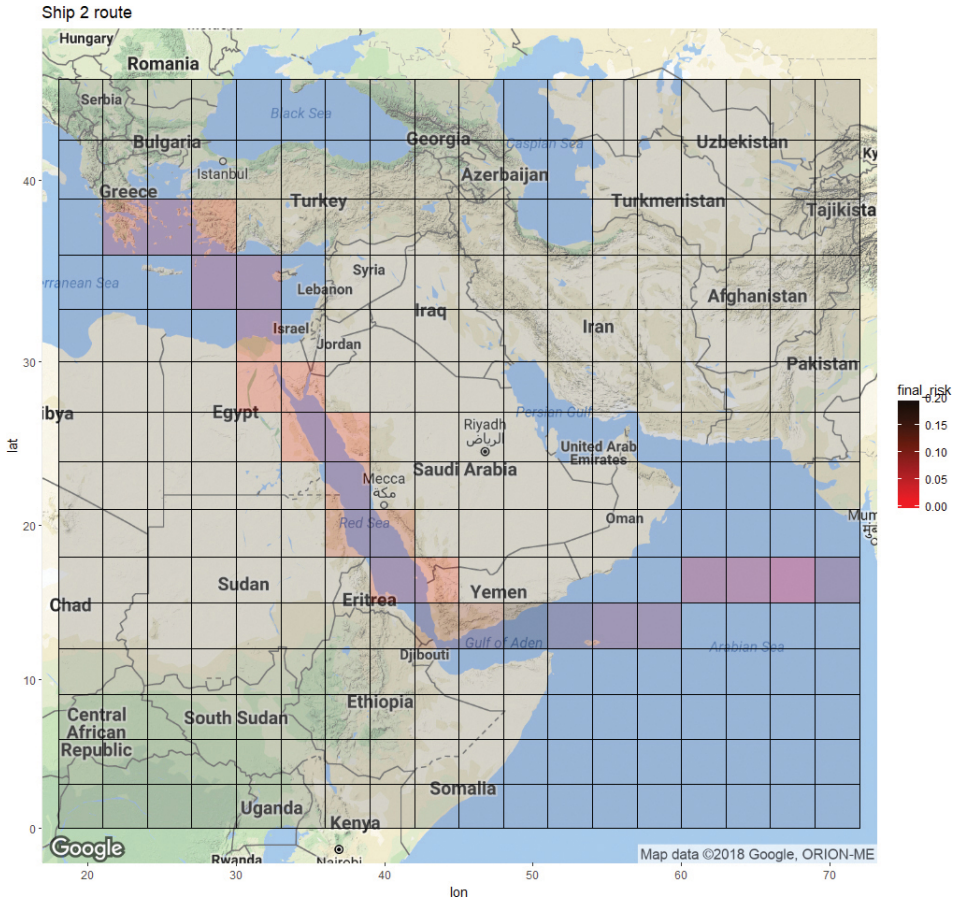


Figure 8.12. Hazard indexes for maritime areas ships are sailing through—Ship 2

Source: Own work.

Table 8.8. Experiment 2—results

Route	Coega (RSA) – Dubai (UEA)	
Voyage month	June	
Risk threshold	0.1	
	Route 1	**Route 2**
Planned route	Along east coast of Madagascar	Along west coast of Madagascar
Hazard index for route (avg)	0.0474	0.0551
Hazard index for route (std)	0.0326	0.0312
Threshold exceeded	Yes (1 sector)	Yes (3 sectors)

Source: Own work.

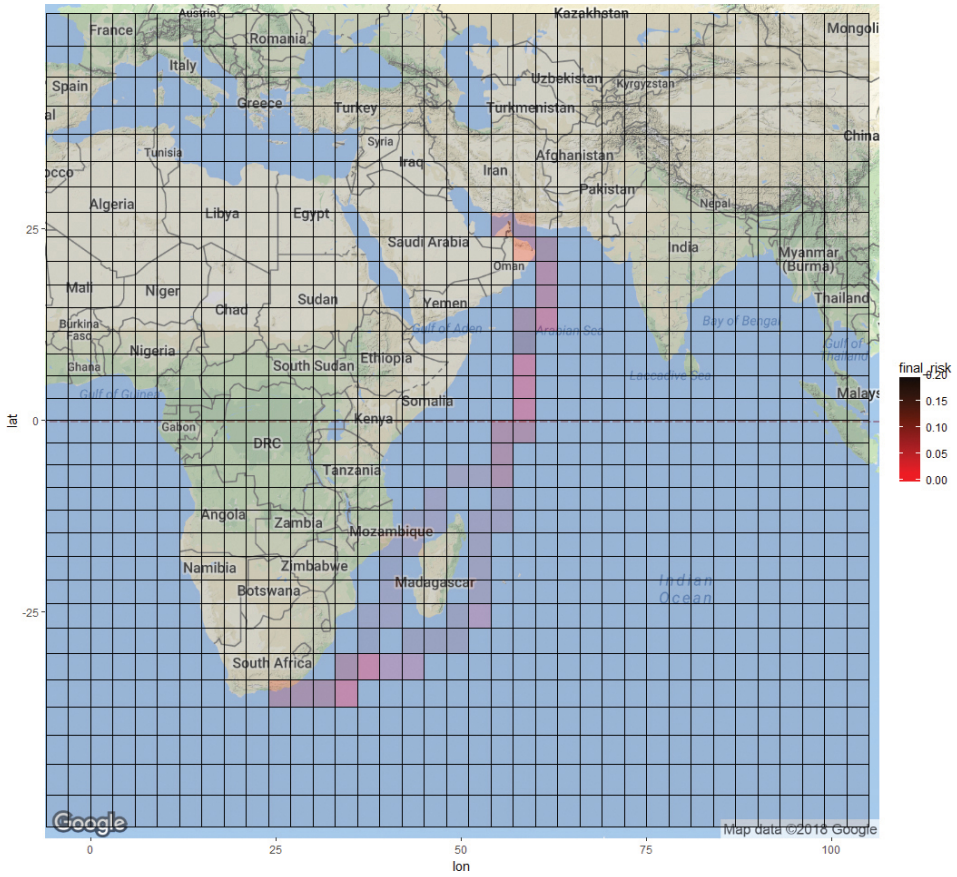


Figure 8.13. Hazard indexes for alternative routes from Coega (RSA) to Dubai (UEA)

Source: Own work.

goes through the east coast of Madagascar, is less hazardous and requires sailing through a smaller number of dangerous sectors. Having obtained this information, a user that plans a voyage can select a safer route.

The analysis of the results also confirms what is observed in the real world, where countries like Mozambique and Tanzania are noted as riskier than Madagascar. Tanzania is additionally listed as a black flag. There were also cases of piracy attacks.

Experiment 3. Having calculated the *Hazard* index for particular sectors of the world, it was possible to evaluate how this information influences the accuracy of the SPP method. In other words, we wanted to check whether, in general, the information about the *Hazard* index should be taken into account to update the predicted travel time and, thus, to determine the predicted ETA.

To this end, similarly to the evaluation of the *Congestion* factor, in the third experiment we analyzed and compared the predicted ETA and the accuracy of the SPP method when the information about hazard on the predicted route is either included or excluded. The basis for this experiment were the results obtained after calculating the predicted travel time that included the *Congestion* factor.

The results of this experiment are presented in Table 8.9. The table compares the accuracy of the method before and after inclusion of the *Hazard* factor and for different values of the accepted risk level on the route. Three levels of the accepted risk were selected for the evaluation purposes: 0.3, 0.5, and 1 (the lower the value, the lower the risk propensity).

In general, the results show that along with an increase of the risk threshold the number of cases for which the accuracy had improved also increased. For the risk threshold 0.3 there are 15 voyages with improved accuracy; for the risk threshold 0.5 it is 21 voyages; for the risk threshold 1 the accuracy did not change in any case (because due to the high level of the accepted risk level there were no risk sectors and no updates of the predicted travel time). It means that inclusion of information about hazard on the predicted route improved the accuracy, for the thresholds 0.3 and 0.5, in 57.69% and 80.77% of cases respectively.

8.6.5. Delay factor results

The last step of the SPP method is inclusion of the information about past delays on the predicted route and taking into account. Similarly to the previous stages of the evaluation, the influence of the *Delay* factor on the accuracy of the ETA prediction was also verified.

In this case, the conducted experiments aimed at comparing the accuracy of the predictions before and after adding the information about past delays on the route. The starting point for this comparison was the predicted travel time that includes the *Hazard* factor for the risk threshold 1. The *Delay* factor was calculated based on the analysis of the past voyages that were found while determining the predicted route and travel profile (Cases 1 and 2). For each past voyage, the difference between the declared ETA at the beginning of the voyage and the actual time of arrival at the destination was determined. Then, the average delay on the route was calculated and the predicted travel time was updated taking into account the delay.

The results of this experiment are presented in Table 8.10. The table compares the accuracy of the SPP method before and after inclusion of the *Delay* factor. The results show that in the case of 17 voyages the accuracy of the predicted ETA improved when the information about the past delays was added and in one case it did not change (because there were no delays on the route in the past). It means that the inclusion of the information about delays improved the accuracy in 65%

Table 8.9. Accuracy of the SPP method with and without hazard—comparison

Voyage number	Port	Avg hazard index	Accuracy congestion	Threshold = 0.3			Threshold = 0.5			Threshold = 1		
				No of risk sectors	Accuracy with hazard	Improved accuracy	No of risk sectors	Accuracy with hazard	Improved accuracy	No of risk sectors	Accuracy with hazard	Improved accuracy
1	Bilbao	0.0517	−24.1547	23	−22.1547	no	12	−23.1547	no	0	−24.1547	no change
2	Bilbao	0.0368	1.3144	7	2.3144	yes	0	1.3144	yes	0	1.3144	no change
3	Teesport	0.0365	5.7675	4	5.7675	yes	2	5.7675	yes	0	5.7675	no change
4	Montoir	0.0565	0.2547	3	0.2547	yes	2	0.2547	yes	0	0.2547	no change
5	Kambo	0.0220	1.4161	1	1.4161	yes	1	1.4161	yes	0	1.4161	no change
6	Muuga	0.0377	3.7867	6	4.7867	yes	3	3.7867	yes	0	3.7867	no change
7	Livorno	0.0528	58.2953	22	61.2953	yes	16	60.2953	yes	0	58.2953	no change
8	Valencia	0.0489	45.1106	12	46.1106	yes	5	45.1106	yes	0	45.1106	no change
9	Algeciras	0.0509	2.0633	18	5.0633	yes	13	4.0633	yes	0	2.0633	no change
10	Istanbul	0.0469	−24.9117	11	−22.9117	no	8	−23.9117	no	0	−24.9117	no change
11	La Spezia	0.0533	−15.8053	7	−14.8053	no	5	−14.8053	no	0	−15.8053	no change
12	Le Havre	0.0670	−24.6100	3	−24.6100	yes	1	−24.6100	yes	0	−24.6100	no change
13	Mersin	0.0570	−39.8803	12	−38.8803	no	10	−39.8803	yes	0	−39.8803	no change
14	Marseille	0.0635	−0.9078	7	−0.9078	yes	4	−0.9078	yes	0	−0.9078	no change
15	Gdansk	0.0423	−41.2603	13	−39.2603	no	4	−40.2603	no	0	−41.2603	no change
16	Gothenburg	0.0463	1.9878	15	3.9878	yes	5	2.9878	yes	0	1.9878	no change
17	Fos	0.0635	−2.4344	7	−1.4344	no	4	−1.4344	no	0	−2.4344	no change
18	Aliaga	0.0646	−0.1561	13	0.8439	yes	11	0.8439	yes	0	−0.1561	no change
19	Tenerife	0.0473	4.4153	10	5.4153	yes	3	4.4153	yes	0	4.4153	no change
20	Swinoujscie	0.0677	−33.8781	5	−32.8781	no	4	−33.8781	yes	0	−33.8781	no change
21	Sines	0.0462	2.8406	10	3.8406	yes	4	2.8406	yes	0	2.8406	no change
22	Barcelona	0.0535	−4.0639	8	−3.0639	no	5	−3.0639	no	0	−4.0639	no change
23	Piraeus	0.0607	0.8142	12	1.8142	yes	10	1.8142	yes	0	0.8142	no change
24	Genoa	0.0486	9.7875	16	11.7875	yes	7	9.7875	yes	0	9.7875	no change
25	Felixstowe	0.0721	−23.3739	5	−23.3739	yes	3	−23.3739	yes	0	−23.3739	no change
26	Lysekil	0.0522	−13.6419	4	−13.6419	yes	2	−13.6419	yes	0	−13.6419	no change

Source: Own work.

Table 8.10. Accuracy of the SPP method with and without delay factor—comparison

Voyage number	Port	Accuracy with hazard	Accuracy with delay	Improved accuracy
1	Bilbao	−24.15	−3.43	yes
2	Bilbao	1.31	−4.16	no
3	Teesport	5.77	24.85	no
4	Montoir	0.25	−0.69	no
5	Kambo	1.42	1.42	yes
6	Muuga	3.79	−16.88	no
7	Livorno	58.30	−5.23	yes
8	Valencia	45.11	0.25	yes
9	Algeciras	2.06	1.42	yes
10	Istanbul	−24.91	−0.16	yes
11	La Spezia	−15.81	−4.06	yes
12	Le Havre	−24.61	2.79	yes
13	Mersin	−39.88	−14.70	yes
14	Marseille	−0.91	45.11	no
15	Gdansk	−41.26	−4.94	yes
16	Gothenburg	1.99	−1.19	yes
17	Fos	−2.43	−24.91	no
18	Aliaga	−0.16	5.79	no
19	Tenerife	4.42	−0.81	yes
20	Swinoujscie	−33.88	−24.61	yes
21	Sines	2.84	−41.88	no
22	Barcelona	−4.06	−2.91	yes
23	Piraeus	0.81	−46.26	no
24	Genoa	9.79	−0.37	yes
25	Felixstowe	−23.37	−1.01	yes
26	Lysekil	−13.64	−7.64	yes

Source: Own work.

of cases. As a result, it might be concluded that in most cases the *Congestion* factor improves the accuracy of ETA prediction and it might be taken into account. Thus, it is justified that the *Congestion* factor is a part of the SPP method.

8.7. Summary of the results

The aim of the evaluation process of the SPP method was to show its accuracy, compliance with real world examples, efficiency (how fast the results can be pro-

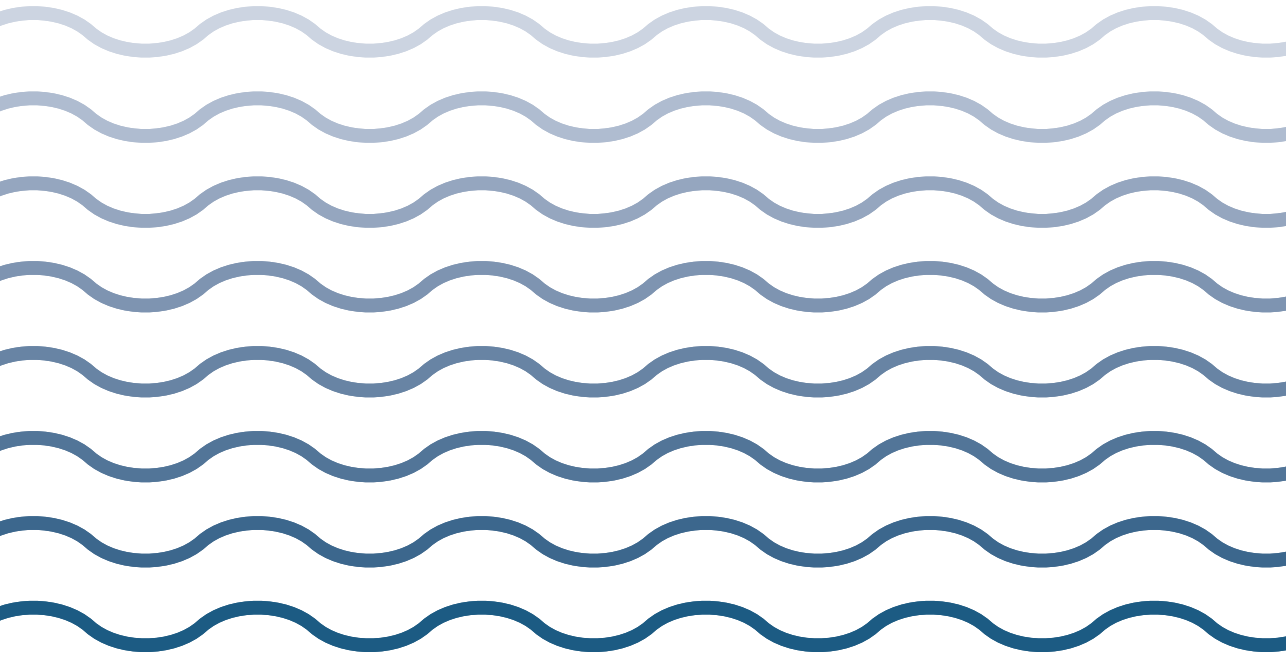
vided) and usefulness of the method in supplying a potential user with up-to-date information regarding a ship's punctuality and ETA.

The results of experiments presented in the previous section prove the quality of the SPP method and can be summarized as follow:

- Based on the conducted experiments, in which the whole SPP method was verified and tested, it was confirmed that the method is effective in 88.42% of cases. It means that in general the method provides more accurate estimation of the ETA than the ETA provided by a captain. This result also proves the effectiveness of the method in determining the predicted time of arrival at a given destination port, and thus its usefulness in supplying a potential user with more accurate and up-to-date information regarding estimated time of arrival of ships to a given destination.
- The evaluation confirmed that it is justified to include additional information about the operational environment of the ship as well as historical information while determining the predicted travel time and the ETA. This concerns especially the information about congestion, potential hazard on the route and past delays:
  - Inclusion of the information about congestion improved the accuracy of estimations in 69% of cases.
  - Inclusion of the information about potential hazards improved the accuracy of estimations in 79.49% of cases on average.
  - Inclusion of the information about the past delays on the route improved the accuracy of estimations in 65% of cases.
- The conducted experiments proved that the method is able to appropriately indicate delayed ships. The method in 88.35% of cases correctly predicted whether a ship will or will not be delayed.
- The results of the additional experiments conducted to evaluate the *Hazard* index proved that the proposed method for determining the *Hazard* index based on three hazard factors is useful and confirms real-life observations. The experiments confirmed the usefulness of the method in supporting a potential user in decision-making regarding which route to choose for a given voyage and indicating potential hazardous areas that require special attention.
- The *Hazard* index is a factor that should be taken into account in calculating the predicted ETA, especially when there is a low propensity for risk. In the case of a high propensity for risk, the index may not influence the predicted ETA.
- All the experiments were conducted based on real, historical data, and real examples of ships' voyages. Thus, it might be said the obtained results are in compliance with what is observed in reality and the method can be used in real business processes.

It is worth mentioning that the evaluation process was a real challenge when it comes to the analysis of the huge amount of available AIS data. The process required an appropriate infrastructure that helped to deal with this challenge; here the services provided by the Microsoft Azure platform, and especially the capabilities of the Hadoop cluster and the distributed computing based on Spark, did a really good job. By using a relatively small cluster, it was possible to significantly speed up all the analysis. The processing of millions of AIS messages collected for a one-year period took just a few minutes instead of a few hours or days, like it was observed in other research (Marine Management Organisation, 2014; Shelmerdine, 2015; Wu et al., 2017). Probably, along with an extension of the cluster and providing better processing capabilities, it would be possible to decrease the processing time even more. In conclusion, it might be said that the proposed SPP method is much more efficient than other existing solutions when the cloud solutions are used.

Chapter 9



9. APPLICATION OF BIG DATA TECHNOLOGIES FOR MARITIME DATA ANALYSIS

In the recent years, big data has drawn huge attention from researchers in information sciences, as well as from policy and decision makers in governments and enterprises. This results mainly from the fact that huge volumes of data provide a great potential and may enable the discovery of highly useful information which originally is hidden in data. Big data encapsulates tools that may help to process and analyse these huge volumes of data in a fast and efficient manner. Therefore, big data has been one of the current and future research frontiers, resulting in a new data-intensive scientific paradigm, also known as big data paradigm. Data-intensive science is especially concerned with data-intensive computing and aims at providing tools to handle big data problems. The maritime domain is one of the sectors where big data technologies might be utilized, especially with regard to analysis of vast amount of satellite, radar, and other sensor-based data. This concerns inter alia data from Automatic Identification System (AIS) that, compared to other sources, generates a huge amount of data about the movement of vessels every day. Analysis of this data (especially real-time and retrospective analysis) requires utilization of appropriate technologies that can deal with big data challenges.

This chapters presents how big data technologies may be applied to analyse maritime data. Two case studies, showing the potential stemming from big data processing, are presented. One that focuses on maritime anomalies detection (Section 9.1), and another one that is related to the generation of maritime traffic networks (Section 9.2).

9.1. Application of big data technologies for maritime anomalies detection

As already described in Chapter 6, the maritime domain has nowadays been facing a problem of detection and anticipation of various maritime anomalies. As a result, anomaly detection has become one of the main issues of the Maritime Surveillance. Detection of maritime anomalies requires collection and analysis of maritime-related data, such as AIS. However, before leveraging AIS data for the purpose of anomaly detection, first they have to be pre-processed (decoded), stored, and analysed. After a few years, they stack up to terabytes of data.

The analysis of literature shows that there is a growing number of maritime surveillance systems which offers threat detection capabilities (see Section 2.5). However, they are based on traditional architectures and approaches for data processing—centralized, relational database systems, SQL-based applications for managing and accessing data, clearly defined structured formats, static schemas, applications that require loading data from a disk into the memory to process data, etc. These approaches and architectures are costly and known for their inefficient and poor scalability when large volumes of data need to be processed (Trujillo et al., 2015). As AIS datasets are large and complex, traditional data processing tools are inefficient in processing them within a tolerable time. A solution for this challenge is application of big data technologies. These technologies assume acquisition of vast amounts of data from various sources and in different formats, which is further processed, fused, and analysed in (near) real-time. Moreover, in the case of anomaly detection, a relatively long period of time needs to be analysed in order to detect the standard behaviour of vessels and find patterns such as the main routes that are followed by most ships or by ships of a given type. Besides, when applied to security and safety purposes, anomaly detection needs to be performed online—it is crucial to reduce delays between an anomalous event and its detection. This is another argument for the application of big data technologies.

There are scholars who have already recognized the importance of big data for AIS data processing. X. Wang, Liu, Liu, de Souza, and Matwin (2014) have carried out an extensive study on vessel route anomaly detection with the MapReduce algorithm. Notably, they presented Density-based Spatial Clustering of Applications with Noise considering Speed and Direction (DBSCAN_SD) and Parallel Meta-Learning (PML) in big data settings. As it turned out, the accuracy and time complexity results improved with the numbers of nodes in their cluster. A distributed DBSCAN_SD method was also used in the work of K. Qi (2016), who compared a Velocity OLAP (vOLAP) with Hadoop for analysis and discovery of vessel traffic patterns from trajectory data. Another research was conducted by Mestl, Tallakstad, and Castberg (2016), who focused on previously disregarded parameters in AIS data and presented an approach that, based on the rate of turns at a maximum time resolution, detects (near) collision situations. They used HBase, which is a popular distributed database running on Hadoop Distributed Filesystem (HDFS). Also, Chatzikokolakis, Zissis, Vodas, Spiliopoulos, and Kontopoulos (2019) proposed a distributed architecture for detecting possible collisions, groundings and travel patterns deviations based on AIS. Their system follows the Lambda architecture paradigm, in which one part of data processing is executed in batches, while the other one in the streaming mode with the goal of detecting deviations of vessel behaviour in real-time. For anomaly detection, they use the Kafka producer / consumer distributed platform and the Akka system as a stream computation engine.

Taking into account the aforementioned challenges and the fact that there is still a need for efficient solutions for maritime big data processing, we developed a set of methods for the detection of selected maritime anomalies relying on an analysis of a vast number of AIS messages. The conducted analyses concern anomalies related to the movement of ships and were performed using big data technologies. The process was focused on the efficiency of calculations. As a result, we were able to compare the big data approach for AIS data analysis with the traditional, SQL-based one.

In the following sections we present our research method, describe the dataset used in the research, present the obtained results for anomaly detection and finally show the efficiency comparison of performing such an analysis in a traditional, SQL-based setting and a big data one.

9.1.1. Methodology

We used two types of data sources in our research: AIS messages and selected open internet sources. The AIS dataset used within our analysis consists of class A position reports (AIS message types 1, 2, and 3) and static and voyage-related data reports (AIS message type 5) from satellite and terrestrial receivers. The data scope of this study was limited to messages sent by tankers[1] in 2015. In total, we collected 569,079,486 class A position reports matching these criteria (19 GB with Parquet's compression[2]). Due to the focus on anomalies of a dynamic nature, we later narrowed the dataset down to reports with the navigational status set to 0 or 4 (*under way using engine* and *constrained by her draught* respectively), which resulted in 313,747,021 position reports (message type 1–3). These messages formed trajectories of vessels.

Following other studies (Marine Management Organisation, 2014; Shelmerdine, 2015; Wu et al., 2017), we adopted the grid-based approach. Therefore, the whole world was tessellated into 64,000 segments of $1^\circ \times 1^\circ$ each. These segments were further described by parameters derived from ship position reports. Segments with no received messages were excluded from further calculations, which resulted in 33,123 active segments. Then each point of each trajectory of each vessel was evaluated with regard to the occurrence of selected types of anomalies. In 2015, there were 34,662 tankers with distinct MMSI numbers, and 32,262 of them had meaningful trajectories, i.e., their maximum reported speed over ground (SOG) was higher than zero.

The second dataset consisted of data collected from selected open internet sources. The purpose of this step was to enrich the information about a given vessel

1. A tanker is a vessel whose reported two-digit *ship type* field starts with 8.
2. https://github.com/apache/parquet-format

with additional data. However, internet sources are known for their issues with data quality. Data might be incomplete, inconsistent, or may contain contradictory facts. Therefore, appropriate methods were developed to alleviate these quality issues. The process of internet sources selection and data fusion was described in detail in Sections 4.6.2–4.6.3. As a result, various ancillary data for tankers were acquired, such as tonnage, dimensions, detailed type, build year, builder, home port, detentions and inspections data as well as classification statuses and affiliation to a classification society. Notably, we collected 57,193 maritime companies, 85,652 classification surveys from 134 classification societies, and 29,011 events of bans and detentions of vessels. Regarding MID country codes, we collected 23 black-listed flags, 30 grey-listed flags, and 50 flags marked as the Flag of Convenience.

AIS data forms a continuous data stream—therefore, traditional methods relying on one physical machine might be computationally inefficient. Some studies reported that processing of one month's AIS data takes one day (Marine Management Organisation, 2014; Shelmerdine, 2015). Wu et al. (2017) covered a few years of AIS data on a global scale in their research, though they did not leverage big data techniques. Namely, the solution proposed by Wu et al. (2017), based on the MySQL technology, did not scale well. This problem was addressed in our research in which we replaced our legacy Microsoft SQL Server database with a big data solution. In general, a reliable big data cluster should provide a near-linear scalability, in-memory computing, stream processing, and efficient data compression. We chose the Hadoop-compliant processing framework—Apache Spark (Zaharia et al., 2012), which enables fast in-memory computing and facilitates a number of analytical tasks. As a result, we developed a scalable solution that enables efficient analysis of a huge amount of AIS data. In the case of the presented research, 257.5 GB RAM memory was used to conduct the analysis in an in-memory manner and a set of 40 cores was responsible for a parallel task execution. Moreover, AIS messages were stored using space-efficient column-oriented data format—Parquet. Technically speaking, maritime anomaly detection requires appropriate processing of vast amounts of immutable data (AIS) in order to infer correct findings. Therefore, we applied the Lambda architecture in our solution, in which static anomalies are processed in a stream, while traffic analysis and loitering detection are conducted using batch processing. Similarly to other big data solutions, adding more worker nodes to a cluster is expected to increase computational power and the storage capacity in a linear manner. The former is the main bottleneck in legacy maritime surveillance systems.

To compare the new solution with the legacy one, we ran a series of practical tests—namely, the movement-related anomaly detection methods, which are presented in the next section.

9.1.2. Anomaly detection

This section describes the obtained results of the developed anomalies detection methods. Maritime anomalies were selected in the course of the SIMMO project (see Section 4.6). They included:

- Inconsistent or missing AIS data (i.e., incomplete static and dynamic information).
- Ambiguous identification (i.e., duplicated or implausible MMSI or IMO number).
- Sudden change of ship's identity (i.e., change of name, call sign, type).
- Flying a black- or a grey-listed flag.
- Suspended/withdrawn classification status.
- Registration of a ship's owner/manager as a poor-performing maritime company.
- Banned or detained ship.
- Loitering on the high sea (i.e., anomalies in ship's behaviour with regard to speed or course).

These anomalies were further grouped into three types of analysis: traffic analysis, static anomalies, and loitering detection. Traffic analysis identifies the busiest routes, determines the average speed, the average relative speed and its standard deviation. This part does not detect anomalies, though it is necessary for further analysis. Static anomalies concern detection of ships that fly a black or a grey flag or a Flag of Convenience, have an IMO number in a banned or detention list, have a certificate issued by a low-performing Recognised Organisation, belong to a low-performing company, and have a withdrawn or suspended classification status. These anomalies rely on data from open internet sources. Loitering-related anomalies are divided into 7 subcategories: invalid coordinates, location or speed, a sharp change of course, an unpredicted location, and an unusually low or high speed.

9.1.3. Traffic analysis

In order to be able to detect loitering behaviour of ships (definition of loitering has been introduced in Section 6.4), we have to define the notion of *normal speed*, which then can be used as a reference point to indicate the anomalous speed in a given area. Such a speed should be location-specific, i.e., it should be defined for a certain geographical area. In general, a normal speed can be inferred from historical AIS data. To this end, in our research we divided the globe into segments of the size $1^\circ \times 1^\circ$.

First, an overview of a number of messages sent within each segment is presented (Figure 9.1). In total, there were 313,747,021 position reports received from tankers with the navigational status set to 0 or 4 in 2015. It is important to notice that the colours are based on a logarithmic scale, so the supremacy is even higher than visually interpreted from the figure. Moreover, it's worth mentioning that in areas with a dense ship traffic we observed problems with the synchronisation of time between various AIS devices. Therefore, some anomalies that were detected in such regions might be false positives.

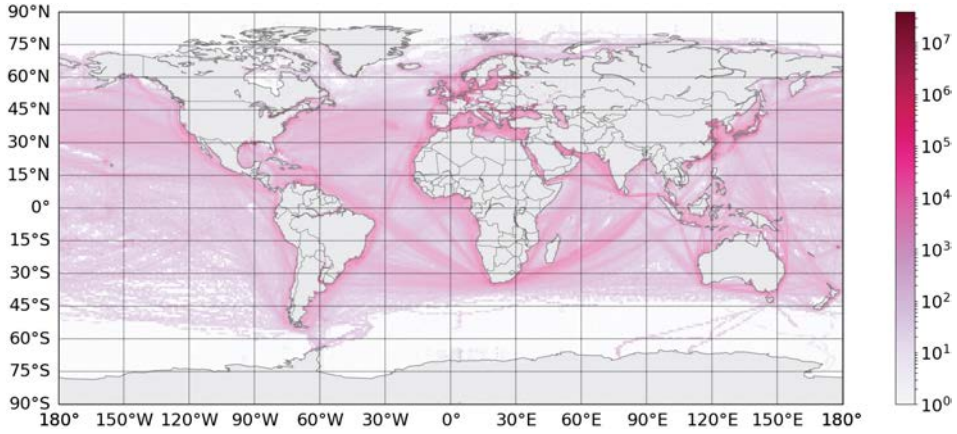


Figure 9.1. Number of received AIS position reports per segment (log scale)

Source: (Filipiak, Stróżyna, Węcel, & Abramowicz, 2018).

Then, we calculated the average speed, the relative speed and the standard deviation of the relative speed in all segments (similarly to the study presented in Section 6.4). The results are presented in Figures 9.2, 9.3, and 9.4 respectively. The analysis of the charts does not reveal anything specific about the results. In general, vessels rarely travel at full steam–usually, it is 61% of their maximum speed and the variability of speed is higher at coastlines and in regions near to ports. Table 9.1 gathers basic statistics about ships' speeds in all the segments of the world.

9.1.4. Static anomalies

In order to detect static anomalies, data from selected internet sources were used.

In calculations presented in this section, the number of detected anomalies in a given segment was divided by the number of AIS position reports received in this segment. Such an approach made it possible to spot anomalies also in the areas

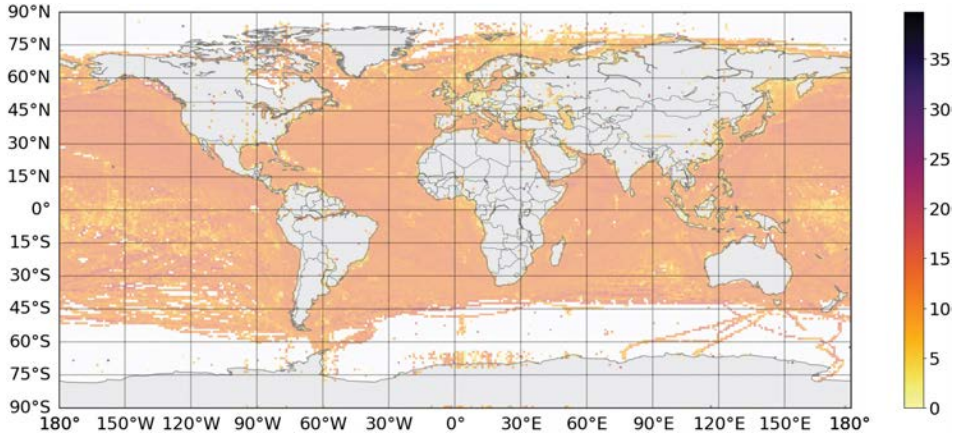


Figure 9.2. Average speed over ground in knots per segment

Source: (Filipiak et al., 2018).

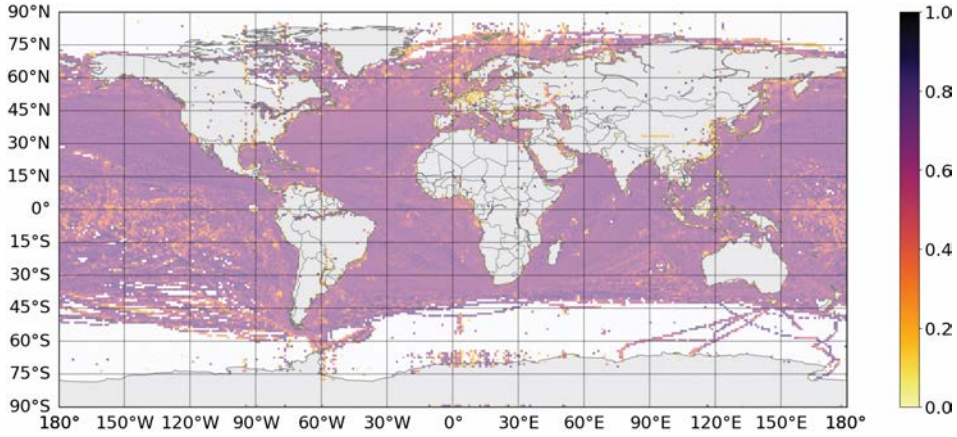


Figure 9.3. Average relative speed per segment

Source: (Filipiak et al., 2018).

with dense ship traffic and resulted from the fact that if we used a nominal number of ships with a given anomalous characteristic, the map would be biased due to the high standard deviation of the messages received in segments with dense traffic. We refer to this approach as *relative*.

The analysis of static anomalies started with classification of flag states into the three categories: black (high risk), grey, and white (low risk). The colours of flags were assigned based on data published by well-known maritime organisations, such as: the Paris MoU, the Tokyo MoU, and the US Coast Guard. These

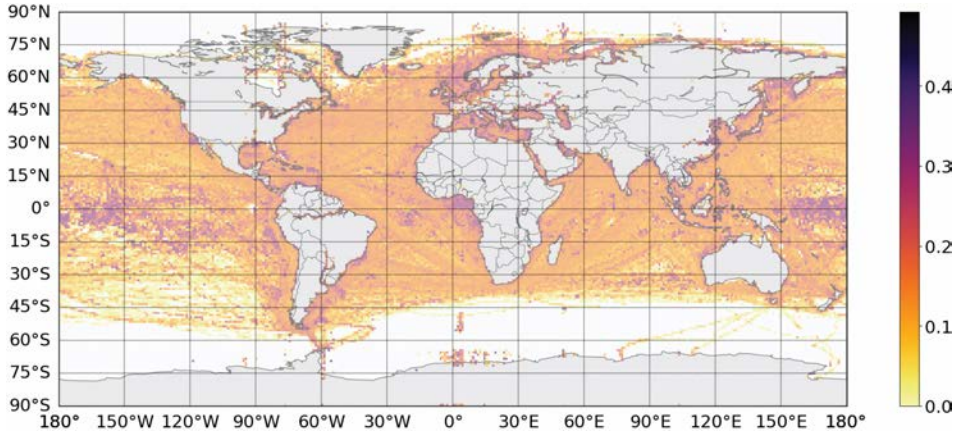


Figure 9.4. The standard deviation of a relative speed per segment

Source: (Filipiak et al., 2018).

Table 9.1. Statistics of tanker traffic in 2015

	Mean	σ	Min	25%	50%	75%	Max
No. of messages	9449.12	240, 904.28	0.0	149.00	554.00	1670.00	39, 260, 666.00
Mean SOG	11.63	3.10	0.0	10.97	12.53	13.31	39.70
Max SOG	17.70	4.91	0.0	14.70	18.20	20.50	40.00
Rel. SOG	0.61	0.16	0.0	0.59	0.66	0.70	1.00
Rel. SOG (std)	0.14	0.08	0.0	0.10	0.14	0.17	0.49

Source: (Filipiak et al., 2018).

organizations determine colours of flags based on risk assessment that reflects the safety performance of ships registered to each flag state, measured as the number of port state inspections and detentions recorded over a three-year period. If a ship is flying a black or grey flag, it is considered as an anomaly.

Spatial distributions of black flags (Anomaly S1) and grey flags (Anomaly S2) are presented in Figures 9.5 and 9.6 respectively. Blacklisted tankers are particularly active in the area between the Indian Ocean and the Pacific Ocean. It results probably from the localization of their home ports—for example, we observed 555 tankers from Indonesia, which is a black-listed country. The highest activity of grey flags was observed on the coasts of the Indian Ocean, particularly around Thailand, the Philippines, and India—probably for the same reason as above. Interestingly, near Madagascar, which is not a grey-listed country, a very high activity of such vessels was observed.

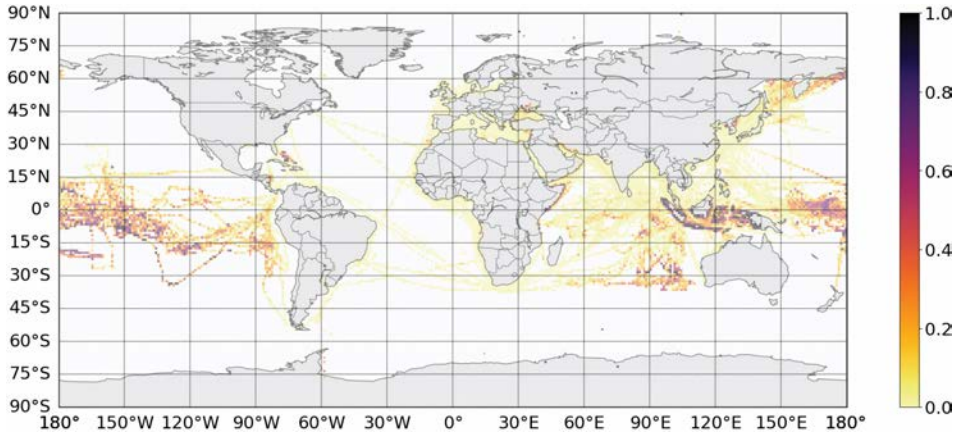


Figure 9.5. Anomaly S1—traffic of tankers with black-listed flags (relative)

Source: (Filipiak et al., 2018).

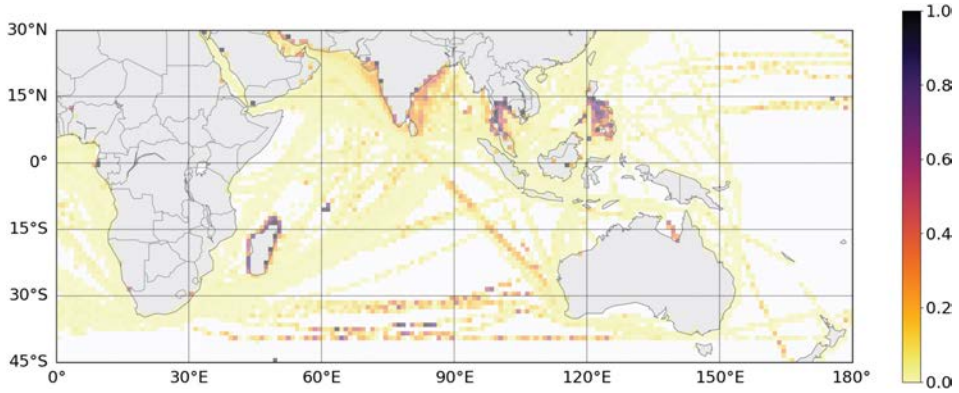


Figure 9.6. Anomaly S2—traffic of tankers with grey-listed flags (relative)

Source: (Filipiak et al., 2018).

The next analysis concerned tankers being registered under the so-called "Flag of Convenience" (FoC). Flag of Convenience (Anomaly S3) is a business practice of registering a ship in a sovereign state, different from that of the ship's owners. FoC allows shipowners to be legally anonymous, which hinders prosecution in civil and criminal actions. Apparently, the spatial distribution of FoC tankers is very high across the whole world, since they constitute nearly 20% of all analysed vessels (Figure 9.7). The Marshall Islands, Liberia, and Panama were the most popular FoC countries.

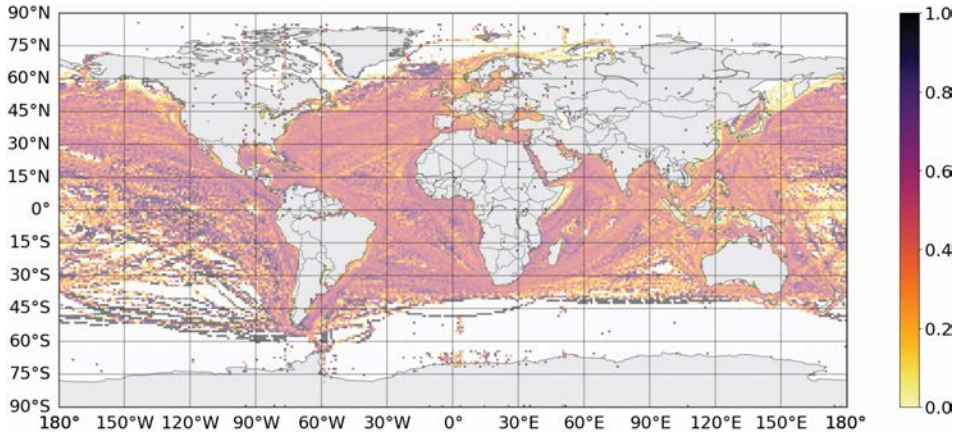


Figure 9.7. Anomaly S3—AIS position reports sent by FoC tankers (relative)

Source: (Filipiak et al., 2018).

Further on, we analysed the spatial distribution of tankers being on a detention or banned ships list. A ship can be subject to Port State Control (PSC), after which, in the case of an occurrence of any deficiencies that are clearly hazardous to the safety of a state or to the environment, a ship can be detained (Anomaly S5). If a ship was detained three or more times by a maritime authority during the last 12 or 24 months, it is classified as banned or added to the list of under-performing ships by a given MoU (Anomaly S4). In the course of the analysis just a single banned tanker was found in the area of the Gulf of Oman (Figure 9.8). On the other hand, detained tankers were found across the whole globe. However, it seems that they were active mostly near Micronesia and the Marshall Islands (Figure 9.10).

Then, we analysed classification certificates issued by the so-called low-performing Recognized Organizations / classification societies (RO). Classification societies are non-governmental organizations that establish and maintain technical standards for construction and operation of marine vessels. The primary role of a classification society is to validate if a design and technical equipment of a ship are in accordance with the published standards. If a ship meets all the requirements, a classification society issues a classification certificate. However, among the classification societies, there are ones that do not perform a minimum number of inspections in a 3-year period and are called Recognized Organizations (RO). If ROs do not meet the criteria for their ships to qualify as Low Risk Ships, they are listed as low-performing ROs (Anomaly S6). Thus, ships having a classification certificate issued by a low-performing RO are potentially dangerous. Our analysis showed that in 2015 such tankers concentrated mostly at the Chinese coast and particularly near Taiwan (Figure 9.9).

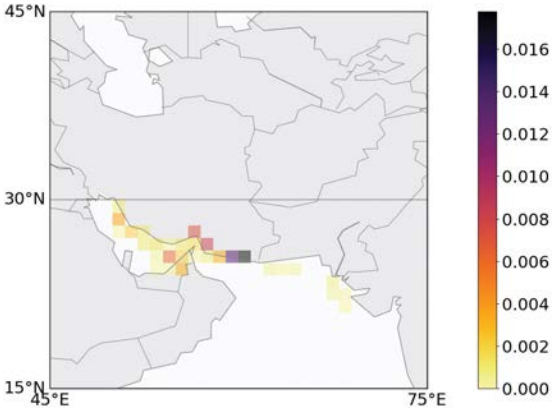


Figure 9.8. Anomalies S4 / S8—AIS position reports sent by a banned and withdrawn or suspended tanker (relative)

Source: (Filipiak et al., 2018).

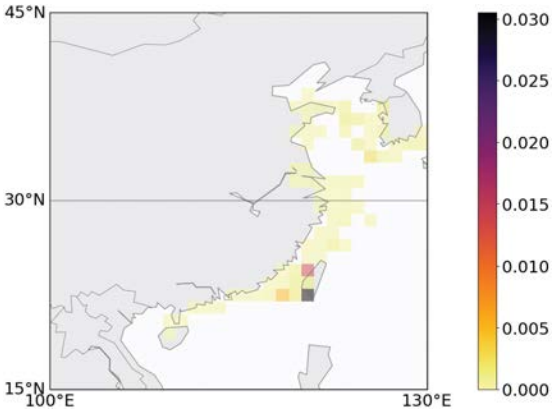


Figure 9.9. Anomaly S6—AIS position reports sent by tankers belonging to low performing ROs (relative)

Source: (Filipiak et al., 2018).

Classification societies are also responsible for granting a classification status for ships. This status is designated based on a periodical survey of a ship and it ensures that a ship meets the classification standards. There are five classification statuses that may be granted: delivered, suspended, reinstated, withdrawn, or reassigned. The ships with the withdrawn and suspended status may be regarded as an anomaly (Anomaly S8). We detected only one tanker that matched this criterion—it was the same vessel as presented in Figure 9.8.

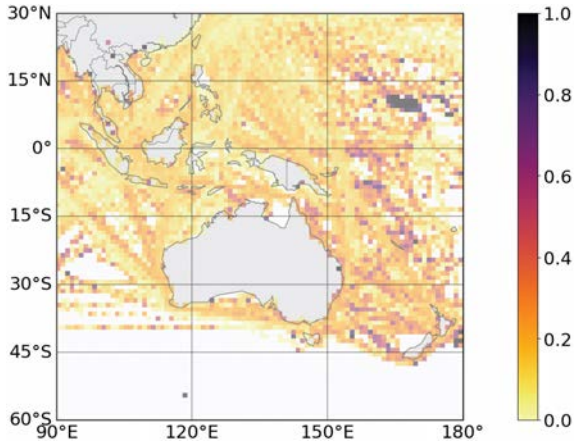


Figure 9.10. Anomaly S5—AIS position reports sent by tankers marked as detained (relative)

Source: (Filipiak et al., 2018).

The final static anomaly concerned tankers being owned / managed by a poor--performing company. The European Maritime Safety Agency (EMSA) publishes a list of such poor-performing companies (Anomaly S7). In the course of the analysis, 24 tankers matching that criterion were identified. They were particularly active in some parts of the Pacific Ocean, south of Hawaii (Figure 9.11).

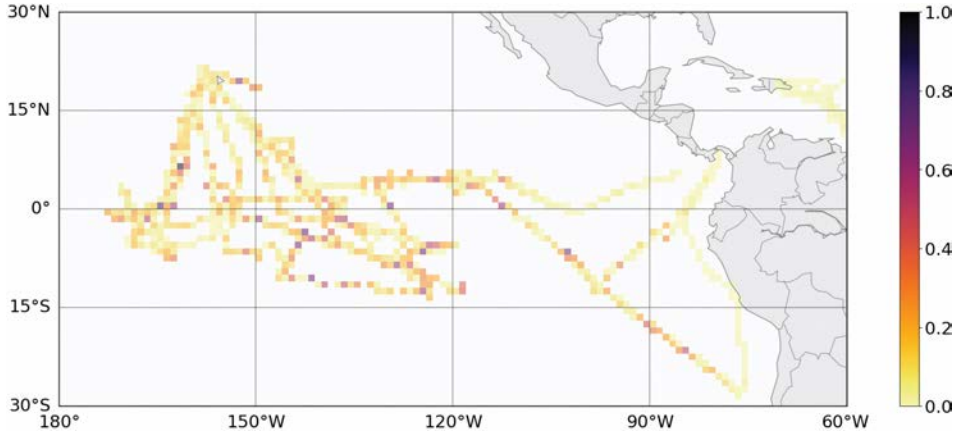


Figure 9.11. Anomaly S7—AIS position reports sent by tankers belonging to low performing companies (relative to the number of all considered position reports in a segment)

Source: (Filipiak et al., 2018).

Table 9.2 summarizes the detected static anomalies for tankers.

Table 9.2. Static anomalies related to tankers in 2015

ID	Anomaly	No. of tankers	%
S4	IMO in banned list	1	0.003
S8	Withdrawn or suspended	1	0.003
S6	Low performing RO	5	0.014
S7	Low performing company	24	0.069
S1	Black-listed flag	1512	4.362
S2	Gray-listed flag	1521	4.388
S5	IMO in detention list	1983	5.721
S3	Flag of Convenience	7097	20.475
	Tankers without static data anomalies	24345	70.235
	Tankers total	34662	100.000

Source: (Filipiak et al., 2018).

9.1.5. Loitering detection

As already indicated in Section 6.4, loitering is mainly related to an anomalous speed of a vessel. In our research, loitering-related anomalies were divided into seven categories: invalid coordinates (L1), location (L2) or speed (L3), sharp change of course (L4), unpredicted location (L5), and unusually low (L6) or high speed (L7).

The first three types of anomalies (L1–L3) result from the verification of the correctness of AIS data values. First, we checked if correct coordinates are provided in an AIS message. If not, Anomaly L1 is reported. Then, whether the reported speed over ground is within expected limits (Anomaly L2). We set a threshold at 25 knots, meaning that a speed above this value will be perceived as an anomaly. Thanks to this, segments with the highest relative number of reports of invalid speeds were identified (Figure 9.12). In the next step, we checked whether an actual position of a ship is reliable considering its potential speed over ground (Anomaly L3). This method makes it possible to eliminate problems with incorrect AIS reading since it filters out cases of sudden *teleportation* of a ship (Figure 9.13).

The next method concerns an angle anomaly (Anomaly L4), which detects a sharp change of course (over 90 degrees). If a ship changes its course so rapidly, it might be interpreted as loitering (Figure 9.14).

Anomaly L5—unpredicted location—concerns a situation when a ship is found in another location than inferred from its previous course. The expected location is predicted based on two previous locations of a ship (points and times), assuming that a vessel should continue its trajectory. A location other than the predicted

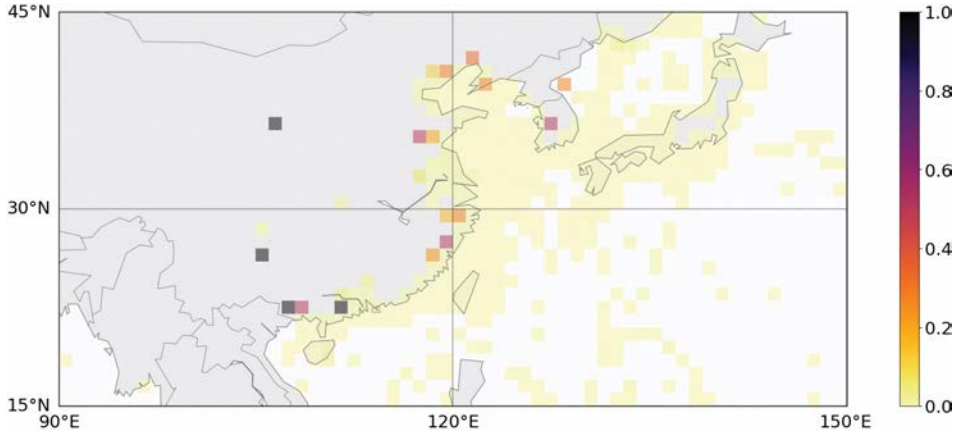


Figure 9.12. Anomaly L2—AIS position reports with an invalid speed (relative)

Source: (Filipiak et al., 2018).

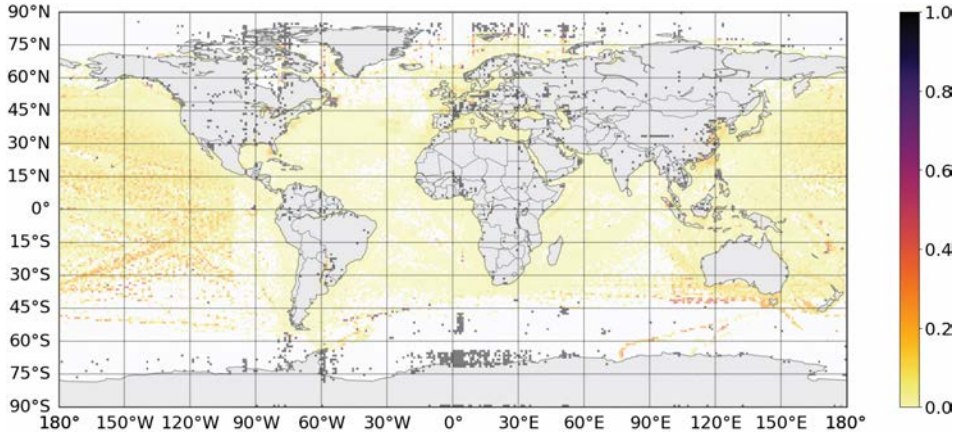


Figure 9.13. Anomaly L3—AIS position reports with an invalid location (relative)

Source: (Filipiak et al., 2018).

one with a margin of 3 miles is considered as anomalous. However, ships that do not move for over 1 hour are excluded. The detected anomalies with regard to unpredicted location are presented in Figure 9.15.

The last method tests whether a ship is sailing with an unusually low or high speed (Anomalies L6 and L7). Loitering occurs when a ship being on the high sea starts sailing with a low speed. This method compares the ship's relative speed in a given segment with the average relative speed and its standard deviation calculated for this segment. If the difference exceeds a defined threshold (value of 2 standard deviations), this position report is considered as anomalous. The results for these two types of anomalies are presented on Figures 9.16 and 9.17.

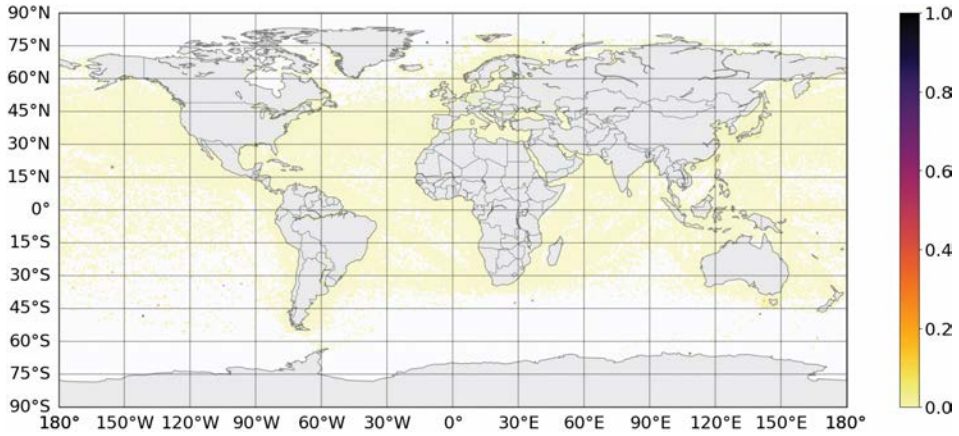


Figure 9.14. Anomaly L4—AIS position reports with an anomalous angle (relative)

Source: (Filipiak et al., 2018).

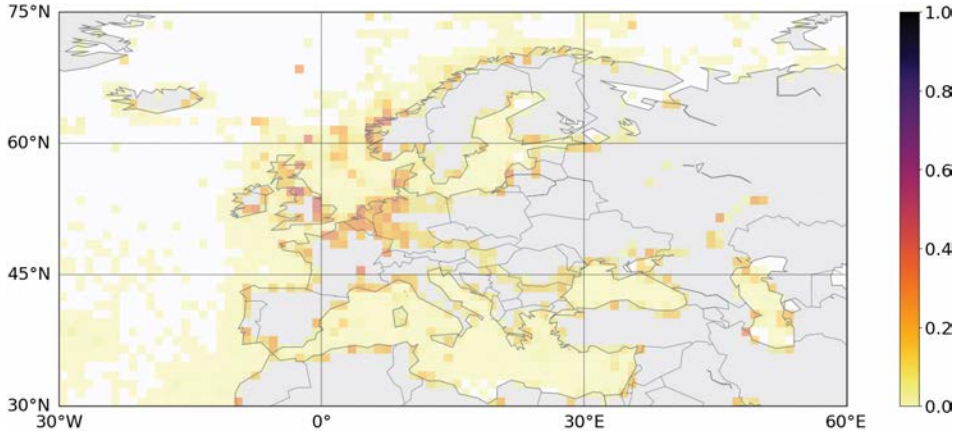


Figure 9.15. Anomaly L5—AIS position reports with unpredicted location (relative)

Source: (Filipiak et al., 2018).

Table 9.3 presents summary statistics of all the loitering-related anomalies detected for tankers in 2015.

9.1.6. Benchmark

As indicated at the beginning of the chapter, the aim of this study was to compare the big data approach to maritime anomalies detection with the traditional (SQL--based) one. In this section the benchmark for the analysis procedure and our

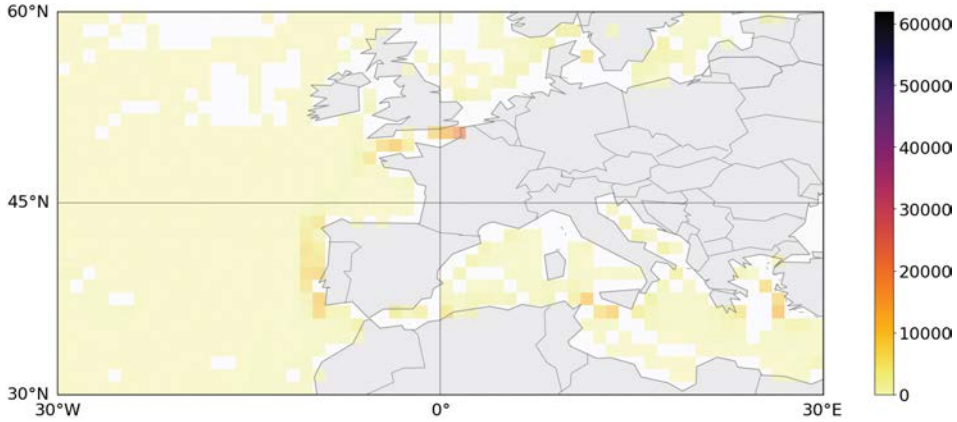


Figure 9.16. Anomaly L6—AIS position reports with unusually low speed (nominal)

Source: (Filipiak et al., 2018).

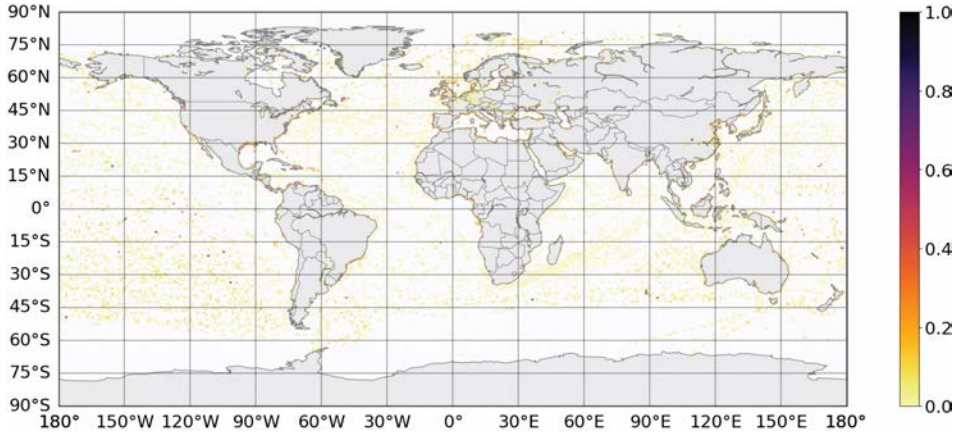


Figure 9.17. Anomaly L7—AIS position reports with unusually high speed (relative)

Source: (Filipiak et al., 2018).

findings are presented. The result analysis consists of three stages. First, we investigate the general results. After that, we compare the traditional and the big data approach. Finally, we investigate the scalability of Apache Spark in terms of adding more vCPUs.

In the test settings, we used three virtual machines in total. The first one (40 GB RAM, 8 virtual CPU cores) was controlled by Microsoft Windows Server 2012 R2 with Microsoft SQL Server 2012. The second and third VMs were built on a standard CentOS 7 Linux distribution with Spark 2.3.1 with bundled Hadoop

Table 9.3. Loitering-related anomalies detected for tankers in 2015

ID	Anomaly	Reports	%
L1	Invalid coordinates	126, 483	0.040
L3	Invalid speed	808, 339	0.258
L5	Unpredicted location	2, 746, 783	0.875
L7	Speed unusually low	3, 434, 848	1.095
L2	Invalid location	11, 849, 397	3.777
L6	Speed unusually high	24, 495, 390	7.807
L4	Sharp course change	37, 105, 095	11.826
	Messages without anomalies	235, 543, 722	75.074
	Messages total	313, 747, 021	100.000

Source: (Filipiak et al., 2018).

distribution—having 8 and 40 virtual CPU cores respectively. It also featured 40 GB of RAM. The dataset was stored in a column-oriented Parquet format. To facilitate the comparison, it was set up as a one-node pseudo-cluster (standalone mode)—therefore, there was no network bandwidth penalty, which normally occurs in distributed environments.

We tested both approaches with two groups of queries: 1) a *speed statistics* calculation, and 2) the proper *anomalies* detection—the former is needed for the latter. Speed statistics is a set of counts, averages, and maxima for tessellated latitudes and longitudes across the whole globe (the already mentioned segments). We used 5 degrees tessellation in our tests. The anomaly detection queries the dataset in order to find unpredicted behaviour defined in Section 9.1.5 as L1–L7, all at once. We queried for random samples of 10, 100, and 1,000 vessels. It is worth mentioning that for MSSQL, the statistics query was written in pure SQL, while the query for anomalies was a mixture of SQL and Python code. Regarding the big data approach, we used PySpark.

The obtained comparison results are presented in Table 9.4. It is clearly visible that the considered technologies vary significantly in terms of processing speed. The general picture emerging from the analysis is that Spark absolutely outperforms the iterative based traditional SQL approach. Firstly, the calculation on Spark was nearly 5 times faster than in the traditional approach. Regarding anomaly detection, Spark was 1.38 times faster for 10 vessels. More significant differences can be observed with the increase of the number of the analysed vessels. For 100 tankers, Spark was nearly 7 times faster, whereas for 1,000 it was approximately 10 times faster. In other words, MSSQL was an order of magnitude slower in the last case. The superiority of Apache Spark in the bulk data analysis is shown in Figure 9.18. To improve readability, we also provided an additional bar chart in a log scale in Figure 9.19.

Table 9.4. Anomaly detection speed for 10, 100, and 1,000 vessels in seconds (5 degrees tessellation)

Technology	Query time [s]			
	Statistics	Anomalies		
		10	100	1000
MSSQL (8 CPUs)	2241.032	132.631	1785.051	16434.894
Spark (8 CPUs)	410.570	95.683	269.796	1707.539
Spark (40 CPUs)	90.896	24.343	87.239	593.511

Source: Own work.

An immanent feature of nearly every big data solution is the promise of near linear scalability. In order to verify this claim, we increased the number of available vCPU cores five times (from 8 to 40) on the machine running Spark. The time of calculating the statistics was close to 5 times faster. However, the gain at the anomaly detection process was slightly smaller—nearly 4 times faster for the 10-vessel test and approximately 3 times faster for 100 and 1,000 vessels. A possible reason for this discrepancy might be connected to increasing the number of vCPUs, instead of adding nodes to the cluster, which would have constituted a true test of scalability. On the other hand, our code was migrated from the legacy solution and perhaps might still have been optimised better in order to fully take advantage of Spark.

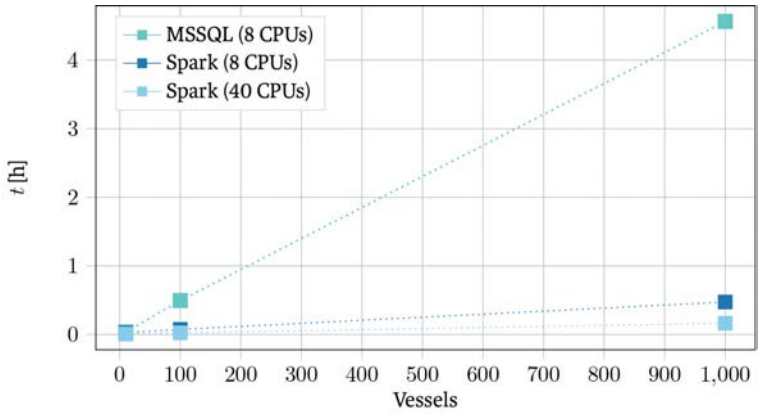


Figure 9.18. Anomaly detection speed for 10, 100, and 1,000 vessels in hours (5 degrees tessellation)

Source: Own work.

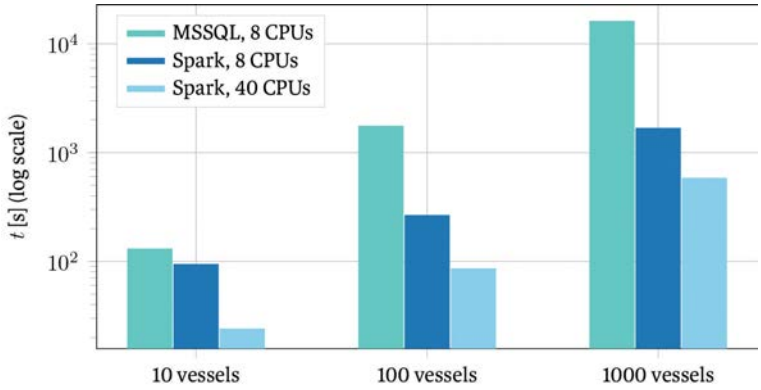


Figure 9.19. Anomaly detection speed for 10, 100, and 1,000 vessels in seconds (log scale, 5 degrees tessellation)

Source: Own work.

All in all, our findings are consistent with the previous results showing a great potential for big data technologies in the maritime domain. Although our dataset was limited to tankers in 2015, these findings might be generalized to other types of vessels. While the advantages of switching to big data in the maritime domain seem obvious, it is advised to follow the common practices in enrolling such systems. It is important to bear in mind the disks read/write throughput, RAM, and network bandwidth speed. Since almost all operations involve reading and writing data, it is necessary to provide fast disks within distributed file systems, such as HDFS. Random Access Memory plays a key role in Apache Spark since it is used to cache resilient distributed datasets (RDDs—Spark's native data format) in memory. This provides a significant speed boost, compared to reading data from hard drives—especially during the iterative calculations. In distributed environments, the network bandwidth can constitute a serious bottleneck for map-reduce calculations (notably in aggregations, or more generally in so-called *shuffle* operations). It is advised to use at least 10 Gbps network cards to reduce this effect.

9.2. Maritime traffic network analysis

In this section the second case study for application in the maritime domain is presented, namely the methods developed for generation of maritime traffic networks based on historical AIS data. The general concept of the method is

presented in Section 5.3. Here, implementation details of the methods are presented, followed by the results of their evaluation. The calculations presented in this section were performed using historical AIS data from three selected maritime areas, and for three types of ships—cargo ships, passenger ships and tankers.

The problem of maritime traffic network analysis can be considered as a classical operational research (OR) problem. There are, however, several features that make our research problem different. First, the task is not about pure routing in the network but about the discovery of the network itself, which might be then used for route planning. Second, we need to consider huge volume and velocity of data. Here we speak about 1 GB (gigabyte) of data daily. It is particularly challenging as classical OR algorithms do not scale. Also, the developed methods are to be used in the HANSA system which is designed to implement more complex scenarios. It requires that calculations be repeated many times with additional constraints, hence the importance of efficiency. By applying the techniques and methods from the research field of big data, the process of extracting maritime traffic patterns might be streamlined (Cazzanti & Pallotta, 2015). Therefore, our solution applies big data technologies to assure efficiency and scalability and take advantage of fast in-memory computation.

9.2.1. Methodology

Extraction of context-sensitive traffic patterns requires both historical vessel movement data and historical weather data. In addition, the sea weather data must cover the same period and area as the historical vessel movement data. In our analysis historical AIS data from 2017 to 2018 was used. The data covered the area of the German Bight and the Baltic Sea. The appropriate historical sea weather data was obtained from services offered by COPERNICUS (see Section 4.1.2). The required data was obtained using a Rest-API and was then fused into a single dataset in order to be combined with the historical AIS data.

As presented in Section 5.3, generation of maritime traffic networks consists of four main methods:

(1) CUSUM algorithm for finding waypoints candidates,
(2) Spatial partitioning of AIS data,
(3) Genetic algorithm for discovery waypoints for each partition,
(4) Mesh generation for discovery edges between waypoints and creation of the traffic network.

In the following subsections details on the implementation and evaluation of each method are provided.

9.2.2. CUSUM

The CUSUM algorithm was implemented in the Scala language. In order to achieve optimal efficiency and parallel AIS data processing, we used the Spark engine with the respective Scala API.[3]

The first step was to define an input structure of AIS data. It is necessary to provide information about the date and time format or to determine the header. Then, the CUSUM method was implemented. In the listing below, the code contains an executable class:

```
        class CusumExec(var mmsi: String, var h: Double, var nsma: Int)
↪  extends CusumMethod {

                this.mmsi = mmsi
                this.h = h
                this.nsma = nsma

                var gp, gn = 0.0
                var prevGP, prevGN = 0.0
                var v_2: Double = _

                private var upperQuantile: Double = _
                private var smaList: ListBuffer[Double] = new
↪  ListBuffer[Double]

                def setUpperQuantile(upperQuantile: Double): Unit = {

                        this.upperQuantile = upperQuantile
                        this.v_2 = this.upperQuantile / 2

                }
                def calculateMi0(yk: Double): Double = {

                        var mi0: Double = 0

                        if (this.smaList.length == this.nsma) {
                                mi0 = UDFs.avg(this.smaList)
                                this.smaList.append(yk)
                                this.smaList.remove(0)

                        } else {
                                if (this.smaList.length > 0) {
                                        mi0 = UDFs.avg(this.smaList)
                                        this.smaList.append(yk)
                                } else {
                                        this.smaList.append(yk)
```

3. http://spark.apache.org/docs/latest/api/scala/index.html

```
                                        mi0 = UDFs.avg(this.smaList)
                                }
                        }
                        return mi0
                }
                def updateDecisionFunction(yk: Double, mi0: Double): Unit =
↪ {

                        val number_gp = this.prevGP + yk - mi0 - this.v_2
                        val number_gn = this.prevGN - yk + mi0 - this.v_2
                        this.gp = Math.round(number_gp * 100.0) / 100.0
                        this.gn = Math.round(number_gn * 100.0) / 100.0
                }
                def reachThreshold(): Boolean = {
                        if (this.gp >= this.h || this.gn >= this.h) return
↪  true else return false
                }
                def lessThan0(decisionFunction: Double): Boolean = if
↪  (decisionFunction < 0) return true else return false

                def reset(): Unit = {
                        this.smaList.clear()
                        this.gp = 0
                        this.gn = 0
                        this.prevGP = 0
                        this.prevGN = 0
                }

                def resetGP(): Unit = this.gp = 0
                def resetGN(): Unit = this.gn = 0

        }
```

There was an instance of the above class created for each trajectory of each unique vessel. If the *reachThreshold*() function returned *true* value, i.e., the threshold was reached, the algorithm stopped and stored the current observation. This situation is presented in Figure 9.20.

The next step included execution of instance methods in order to calculate the decision function in the current combination of a vessel's identifier (MMSI) and timestamp. The last step was to create a Spark data frame, which makes it possible to:

- filter only relevant AIS signals, e.g., by excluding area with certain coordinates or select only specific types of ships;
- sort AIS signals, which is important from the further processing point of view;
- calculate the new column dynamically;

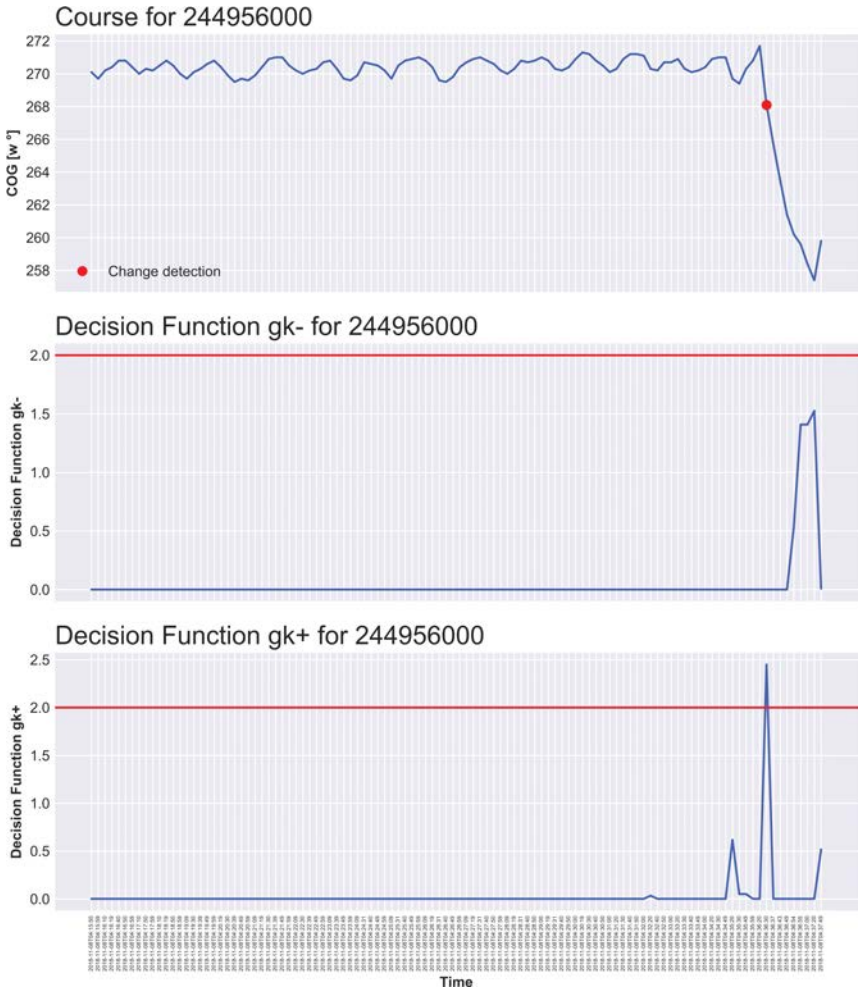


Figure 9.20. Example of a single manoeuvre detection and visualization of the decision function

Source: Own work.

- group by simple keys and divide unordered data into ordered vessel tracks sorted by date and time.

The results can be mapped to a structure containing coordinates and then written into flat files. An example of manoeuvres detected by CUSUM based on course changes is presented in Figure 9.21.

The performance of the change detection algorithm was evaluated in several steps. The main purpose was to find optimal parameters, such as the threshold h and the number of historical data that should be taken into account in the moving

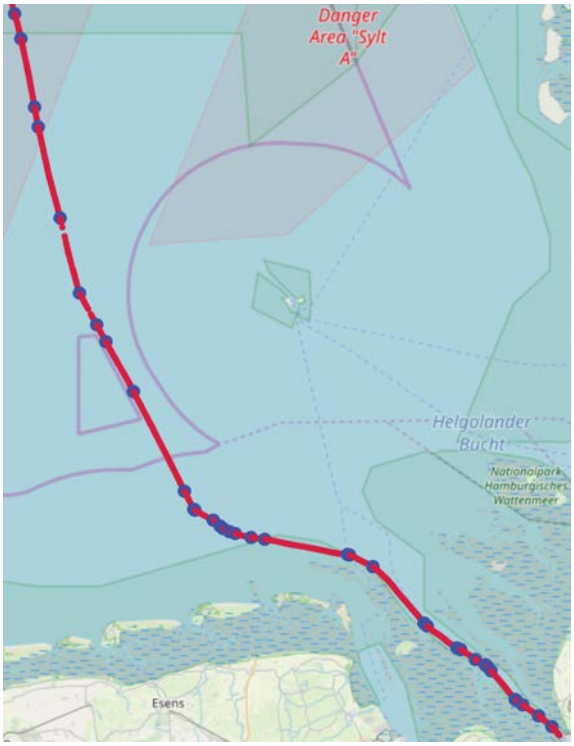


Figure 9.21. Visualization of manoeuvres detected by CUSUM

Source: (Filipiak, Węcel, Stróżyna, Michalak, & Abramowicz, 2020).

average *nsma*. During this process, only individual voyages, limited by fixed coordinates were considered. The key was to find different tracks in terms of change of course and speed. Having a set of MMSI numbers and their tracks, we manually assigned 'expert' points on the map that should be alerted by the algorithm. A single expert point was a circle with a radius of 500m. The main idea was to perform a classification of waypoints returned by the CUSUM and to calculate some measures based on it. In the next step, we provided a range of evaluated hyper-parameters. For the threshold h, we were choosing between 1 and 6 and for the hyper-parameter *nsma* we were choosing between 2 and 10. Then, the algorithm was executed for each combination of set of hyper-parameters. In a single iteration, K-Means algorithm for the list of waypoints returned by the CUSUM was calculated, with the number of clusters equalling to the number of expert points. The following measures were calculated in subsequent iterations:

(1) Distance from the cluster centroids to the nearest expert point. Having all distances for each expert point, it was possible to calculate the mean distance.

(2) Confusion matrix that classifies waypoints to the nearest expert point. We assume that if a waypoint is within the radius of an expert point, it is assigned to it.
(3) Number of unassigned expert points. It means that there were no waypoints detected within the radius of 500m of this expert point.

Based on the confusion matrix, the respective measures were calculated including accuracy, recall, specificity and precision. This method allows for a detailed analysis of the resulted waypoints, e.g., the precision shows the number of unassigned alarm points while the recall shows a number of points that should be classified as waypoints but the algorithm skipped them.

To sum up, we aggregated all single, manually collected tracks with the respective results and took the highest-rated parameters among all samples. It turned out that the most optimal combination of CUSUM parameters was the threshold of 1.25 and the number of historical observations of 8. Finally, the results were collected in the local files.

9.2.3. Spatial partitioning

In the next step two approaches for spatial partitioning were tested: k-d B-tree and QuadTree. We used an implementation available in GeoSpark[4]. Sample results of the aforementioned spatial partitioning methods in the area of the German Bight are presented in Figures 9.22 (a) and 9.22 (b), for k-d B-tree and QuadTree respectively. The blue dots mark AIS points after being filtered with the CUSUM method, the orange circles are the waypoints obtained using the genetic algorithm, and the red rectangles denote separate partitions. Spatial partitioning is used by the genetic algorithm (see the next section), in which each partition is treated separately. In GeoSpark, specifying a desirable number of partitions (variable numPartitions) does not entirely fix the resulting number of partitions and it should be treated rather as an approximate number of partitions. Interestingly, the behaviour is various for different partitioning schemes. This behaviour affects the results of the next step—the genetic algorithm. Since the population and other parameters of the genetic algorithm are set per single partition, the denser partitioning will result in more waypoints at the end. This behaviour can be observed in Figures 9.22 (c) and 9.22 (d).

The goal was to select a method capable of finding a uniform distribution of the CUSUM-generated AIS points in partitions. As a desired consequence, hyper--parameters of the genetic algorithm would control its behaviour better—the more uniformly distributed AIS points in partitions, the better. On the other hand, large deviations in the number of AIS points between partitions makes it hard to choose

4. https://datasystemslab.github.io/GeoSpark/

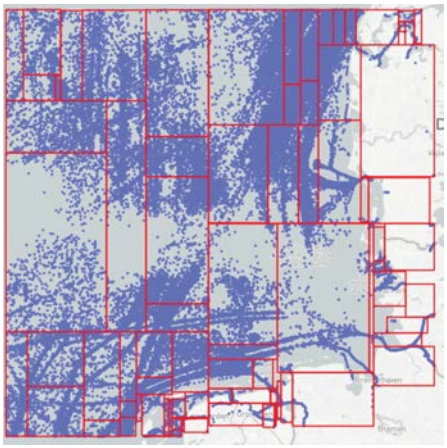


Figure 9.22 (a) *k*-d B-tree (without waypoints)

Figure 9.22 (b) QuadTree (without waypoints)

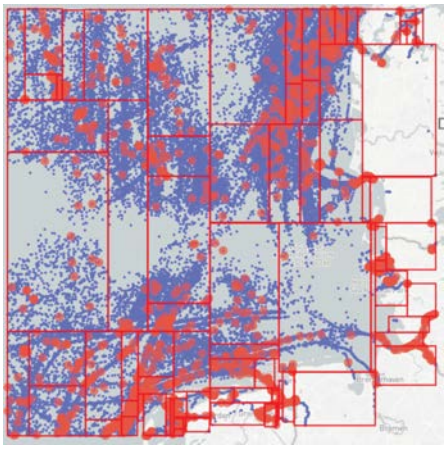


Figure 9.22 (c) *k*-d B-tree (with waypoints)

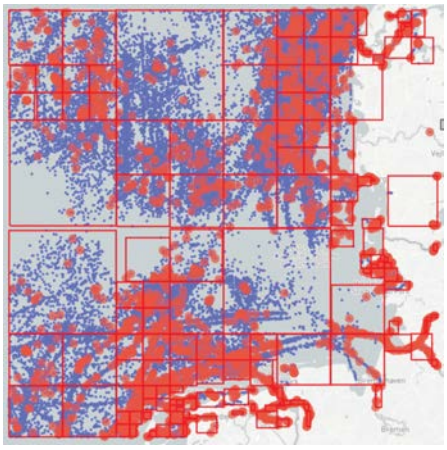


Figure 9.22 (d) QuadTree (with waypoints)

Figure 9.22. Spatial partitioning methods in the area of the German Bight. The blue dots mark AIS points after being filtered with the CUSUM algorithm, the orange circles are the waypoints obtained using the genetic algorithm, and the red rectangles denote separate partitions

Source: Own work.

the right hyper-parameters, as they would impact particular partitions differently. Therefore, we performed a series of experiments, in which the two partitioning methods were compared (each for numPartitions= $\{64, 128, 256\}$). The results for 4-weeks' data are presented in Tables 9.5 and 9.6 for the German Bight and the Baltic Sea consecutively. Finally, we chose *k*-d B-trees, since the standard deviation

Table 9.5. Partitioning evaluation for the German Bight for passenger, cargo, and tanker vessels

	$p = 64$		$p = 128$		$p = 256$	
	k-d B-Tree	QuadTree	k-d B-Tree	QuadTree	k-d B-Tree	QuadTree
count	99	232	188	583	370	1426
mean	2164.35	923.58	1139.74	367.53	579.11	150.26
stddev	674.99	1023.61	441.94	499.20	256.38	219.36
min	1054	0	309	0	67	0
25%	1537	26	818	1	378	0
50%	2084	501	1062	112	550	31
75%	2620	1604	1436	569	731	249

Source: (Filipiak et al., 2020).

Table 9.6. Partitioning evaluation for the Baltic Sea for passenger, cargo, and tanker vessels, 4-weeks data

	$p = 64$		$p = 128$		$p = 256$	
	k-d B-Tree	QuadTree	k-d B-Tree	QuadTree	k-d B-Tree	QuadTree
count	96	211	190	442	369	922
mean	36573.70	16640.17	18479.34	7943.61	9515.11	3808.11
stddev	6916.45	15678.11	4083.42	7673.66	2258.85	3879.84
min	25470	0	11897	0	4342	0
25%	30518	1961	15217	894	7803	343
50%	35315	12474	17682	5904	9243	2533
75%	40303	28029	21343	13074	11018	6301
max	59001	54505	30314	27759	16501	15220

Source: Own work.

of the number of points in each partition tends to be smaller in numerous settings than in the other method. It is worth mentioning that QuadTrees have a visible tendency to produce empty (or scarcely populated, in general) partitions.

9.2.4. Genetic algorithm

We use the genetic algorithm to discover waypoints from AIS data, as it was previously used in the literature (Dobrkovic et al., 2018). Having the AIS points partitioned, the genetic algorithm can be used for each partition to detect the waypoints.

The genetic algorithm is run on each isolated partition separately, and the results are concatenated at the end. This means that this process can be parallelized. Taking the advantage of the latest big data technologies, it can also be distributed. Both of these features are enabled due to the usage of Apache Spark, a well-known in-memory distributed data processing engine built on top of Hadoop. Therefore, this engine was chosen to implement the genetic algorithm. The overall idea of the algorithm is summarised in Algorithm 9.1. The consecutive steps of the genetic algorithm are described in details in Section 5.3.2. After the partitioning step, the algorithm is run for each partition by passing the function DiscoverWaypoints(). After the initial population is generated, the process of generating new offspring is repeated $n-1$ times, where n is the number of epochs.

Algorithm 9.1. Parallel genetic AIS waypoints discovery

1: **function** Genetic-AIS-Waypoints-Discovery(rdd, n_{part})
2: $rdd \leftarrow$ Partition(rdd, n_{part})
3: $w \leftarrow rdd$
4: .MapPartitions(DiscoverWaypoints) ▷ Distributed and parallel
5: .Distinct()
6: **return** w
7: **function** DiscoverWaypoints(AIS, hyperparams)
8: $p \leftarrow$ InitialisePopulation(AIS, hyperparams)
9: **for** $i \leftarrow 2$ **to** n_{epochs} **do**
10: $p \leftarrow$ generateOffsprinf(AIS, p)
11: **return** Iterator(p)

The first step is setting the hyper-parameters and reading the data obtained from the CUSUM algorithm. The genetic algorithm may work with raw AIS data, though the CUSUM-refined data is expected to yield better results.

```
        GeneticAlgorithm.setHyperParameters(epochs, chromosomeLength,
↪  population,
        radius, mutationFactor)
        var numPartitions = partitions
        val pointRDDInputLocation = "/home/jovyan/notebooks/data/1-cusum/"
↪  + fileName
        val pointRDDOffset = 2 // The column offset point long/lat starts
↪  from in CSV file
        val pointRDDSplitter = FileDataSplitter.CSV
        val carryOtherAttributes = false
        var pointsRDD = new PointRDD(sc, pointRDDInputLocation,
        pointRDDOffset, pointRDDSplitter, carryOtherAttributes)
        val buildOnSpatialPartitionedRDD = true // Set to TRUE only if run
↪  join query
        val roundingConstant = 5000
```

```
        val fileNameOut = fileName + "-qt-part" + partitions + "e"
        + epochs + "cl" + chromosomeLength +"p" +population
        + "r" + radius + "mf" + mutationFactor
```

Up next, the spatial partitioning is performed. The data is also cached by Spark to speed up the calculations.

```
        pointsRDD.analyze()
        pointsRDD.spatialPartitioning(GridType.QUADTREE, numPartitions)
        pointsRDD.spatialPartitionedRDD.cache()
```

Finally, the actual algorithm is run on the spatially partitioned data—each partition runs its own genetic algorithm. As they are independent, this results in a massive speedup thanks to the parallel computations. We used a 48-core server in our research—all of them remained busy during the convergence process, which suggests a good parallelization of the algorithm. he results from each partition are combined, rounded up, and cleaned from duplicate values. At the end, the results are transferred to the Spark driver and saved as a CSV file with waypoints coordinates ready to be processed by the edge detection method.

```
        pointsRDD
        .spatialPartitionedRDD
        .rdd
        .mapPartitions(GeneticAlgorithm.discoverWaypoints, true)
        .filter(x => x != null)
        .flatMap(x => x.asInstanceOf[ArrayBuffer[Chromosome]])
        .map(x => x.genes)
        .flatMap(x => x)
        .map(x => ((x._1 * roundingConstant).round.toDouble /
↪   roundingConstant,
        (x._2 * roundingConstant).round.toDouble / roundingConstant)) //
↪   rounding up
        .distinct
        .toDF("lat", "lon")
        .coalesce(1)
        .write
        .mode("overwrite")
        .csv("/home/jovyan/notebooks/data/3-ga/" + fileNameOut)
}
```

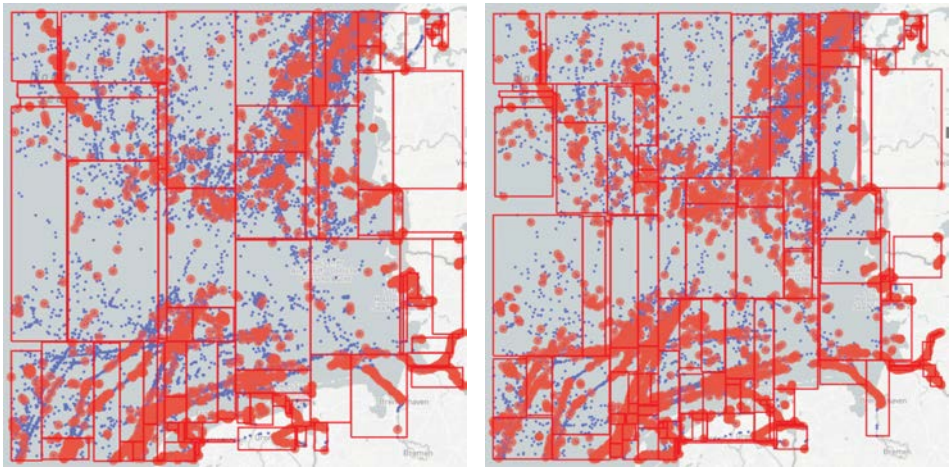
The developed genetic algorithm was then evaluated based on the qualitative approach. Basically it concerns the appropriate selection of hyper-parameters. There are a number of hyper-parameters to control in the algorithm—Table 9.7 sums them up.

All the tests were conducted using the CUSUM-filtered data and were restricted to tankers, passenger, and cargo ships in the vicinity of the German Bight. These are real-world data, with AIS-specific problems, such as lack of coverage in some areas and some spoofed points. Figures 9.22 (a) and 9.22 (b) show the results of tests for

Table 9.7. Hyper-parameters of the genetic algorithm

Symbol	Name	Description
cl	chromosome length	The number of genes within the chromosome
r	radius	The so-called radius of the waypoint, which marks its area of influence
$\min_{\text{div}}$	minimal diversity	The percent of required non-overlapping waypoints in a chromosome
$\|P\|$	population size	The number of chromosomes in the whole population
E	epochs	Total number of cycles for generating new offspring
mf	mutation factor	The percentage chance of a random change in a chromosome
n_{part}	number of partitions	The approximate number of partitions
d	distance function	The distance function used (Euclidean or haversine)
w	weeks	The number of weeks of input data

Source: Own work.



(a) partitions = 64, population = 100 (b) partitions = 128, population = 100

Figure 9.23. Different test settings for 1-week data—testing different number of partitions with 100 chromosomes in populations

Source: Own work.

different numbers of partitions and for 100 chromosomes in the population (one week's data). The initial observation, based on the tests' results is that the most noticeable changes can be observed in the densest sea areas, making it appear more continuous (resulting rather in a smooth route, as opposed to long gaps

between the waypoints), whereas areas scarce in waypoints did not change much. Similar results are produced for a smaller population (40). However, 512 partitions seem to generate too many waypoints.

Then different values of epochs and radii (in degrees—for the Euclidean distance) were also tested. Empirical tests showed that the algorithm quickly converges to the solution—perhaps due to the fact that the population is drawn from the existing AIS points, contrary to (Dobrkovic et al., 2018). Usually, 200–300 epochs seemed to be enough. Setting the correct radius is tricky, since values which are too high are handled poorly in dense areas (the diversity condition can't be met). The experiments also concerned setting different values for the mutation factors. This value must be high due to the fact that the algorithm can "get stuck" in small and dense partitions, so random noise is needed.

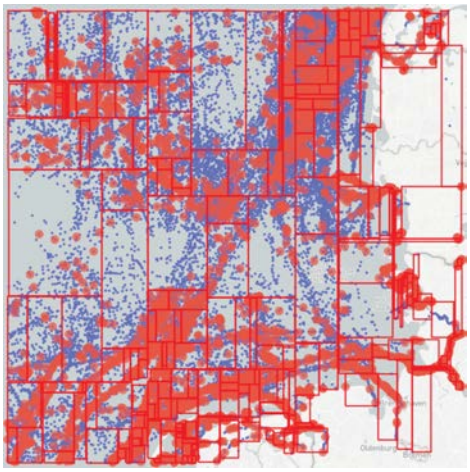
To partially mitigate the problem with poor AIS coverage in some areas, we also performed tests on 4-week data (Figures 9.24 (a), 9.24 (b), 9.24 (c), and 9.24 (d)). The results are not ideal, but the difference is noticeable, as new and desired waypoints emerged in previously empty spaces, merging dense areas. However, the advantage stemming from using 8-week data is not that clearly visible. Finally, using the great circle distance (the haversine function) yields more accurate results at the cost of being an unnoticeably slower solution.

To sum up, the results of the performed tests suggest that more partitions with smaller chromosomes are better than fewer partitions and longer chromosomes, if one wants to prevent having numerous waypoints in small areas. From our experience, the choice of hyper-parameters is area-specific and general values applicable in all areas can't be derived. If GA is used with the CUSUM results, it seems that at least 4-week data should be used. All in all, the algorithm is very prone to missing data (in terms of AIS coverage)—it just won't generate waypoints in such areas. Therefore, a pre-processing with trajectory reconstruction algorithms might be considered in future work. Nevertheless, the algorithm deals quite well with single bad AIS points (spoofed or misread).

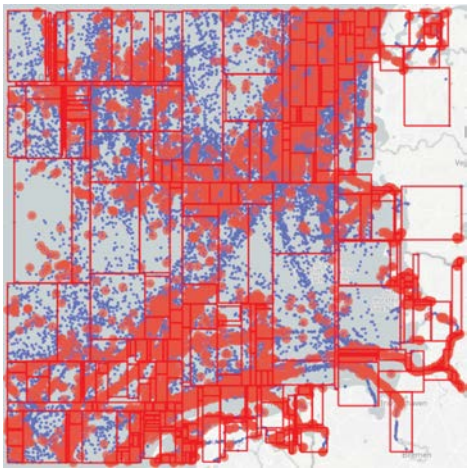
9.2.5. AIS enrichment

Having the waypoints identified, the next phase of maritime traffic network creation is generation of the mesh. In this section this process will be described. Several methods were tested in an iterative approach and their results are presented along with validation.

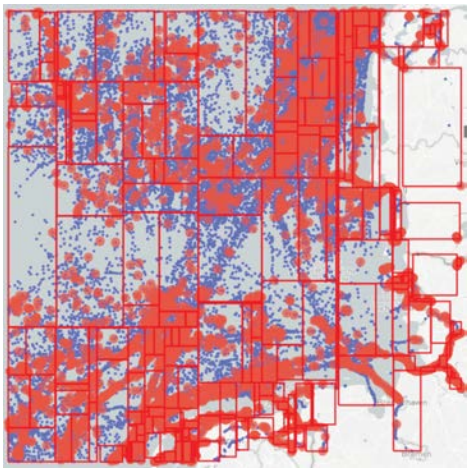
The genetic algorithm described in the previous section generates a set of waypoints. Waypoints are equivalent to nodes of the generated mesh. What is necessary, is to discover the edges, i.e., which waypoints should in fact be connected. It will be conducted based on historical AIS data. By looking at every single trajectory of all vessels (which pass an area of interest) we can track which waypoints they



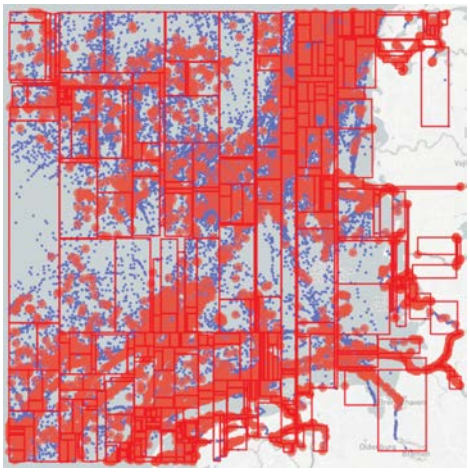
(a) partitions = 256, population = 30, r = 0.3, epochs = 500, mf = 25%, weeks = 4, cl = 2, Euclidean distance



(b) partitions = 384, population = 30, r = 0.3, epochs = 500, mf = 25%, weeks = 4, cl = 2, Euclidean distance



(c) partitions = 256, population = 40, r = 0.3, epochs = 500, mf = 25%, weeks = 4, cl = 2, Euclidean distance



(d) partitions = 512, population = 30, r = 0.3, epochs = 500, mf = 25%, weeks = 4, cl = 2, Euclidean distance

Figure 9.24. Different test settings for 4-week data

Source: Own work.

'visited'. As defined in Section 5.3.3, the first step of the mesh generation process is "AIS enrichment". It is about adding both the identifier of the nearest waypoint and the distance to each AIS message.

The algorithm that was designed for further implementation is given in pseudocode below (Algorithm 9.2). The core function is FunctionKnn.

Algorithm 9.2. AIS enrichment

```
1: function AssignNearestWaypoints(AIS)
2:     AIS_w ← AIS.withColumn('waypoint', NeighborKnn(AIS.lat, AIS.lon))
3:     return AIS_w
4: function NeighborKnn(lat, lon)
5:     w ← ReadWaypoints
6:     nnModel ← NearestNeighbors(algorithm, metric)
7:         .fit(w)
8:     n ← nnModel.kneighbors(lat, lon, n_neighbors = 1)
9:     return n
```

Later, for the purpose of experiments, we also added a separate function for a brute-force calculation of the distance between an AIS point and the nearest waypoint. The FunctionMinkowski calculation is presented in Algorithm 9.3.

Algorithm 9.3. Edges discovery

```
1: function AssignNearestWaypoints(AIS)
2:     AIS_w ← AIS.withColumn('waypoint', NeighborMinkowski(AIS.lat, AIS.lon))
3:     return AIS_w
4: function NeighborMinkowski(lat, lon)
5:     minDist ← ∞
6:     minId ←
7:     w ← ReadWaypoints
8:     for i ← 1 to n do                                  ▷ For all waypoints
9:         w_i ← w[i]
10:        lon_i ← w_i.lon
11:        if Abs(lon − lon_i) > minDist then
12:            continue
13:        lat_i ← w_i.lat
14:        if Abs(lat − lat_i) > minDist then
15:            continue
16:        dist ← Sqrt((lon − lon_i)^2 + (lat − lat_i)^2)
17:        if dist < minDist then
18:            minDist ← dist
19:            minId ← w.id
20:    return minId
```

Taking into account the number of rows in AIS data, the process of assigning waypoints to AIS data proved to be very time consuming. Therefore, we introduced many optimization techniques to make the task feasible. We describe the evolution of our approach below.

KNN method (baseline). Being aware of the complexity of the problem, we decided to use the well-established methods from the standard library. First, we chose the kNN method (k-nearest neighbours algorithm) from the SciKit Learn machine learning library (also known as sklearn)[5]. The idea was that such methods should have already implemented optimized calculations. The most promising were fast indexing structures such as Ball Tree or KD Tree.

This part of the processing was implemented in Python in the PySpark environment.

Our initial benchmarks showed the weakness of this approach—a lot of time was needed for iterations. Processing of the first 10,000 rows took on average 21.4 min. The throughput was then 7.8 rows per second. It seemed not to be a promising perspective for processing millions of rows. Indeed, the test on only 25 rows showed that the method needed 3.42 seconds on average (7.3 rows per second).

The whole AIS data subset for the south Baltic area contained 485,323 rows. The expected processing time would be 66,483 s, i.e., 18.47 h. This time may be sped up by using more partitions in the Spark technology for data processing. If the AIS dataset is split into 16 partitions, the expected time is 1h 9m 15s. In reality, the processing time took on average 1 h 4 min 46 s, but in this case more partitions for Spark had to be used.

KNN approach with UDF. One of the possibilities to speed up data processing in Spark is to improve the way we iterate over rows. The proposed design pattern is to leverage user-defined functions (UDF) which are applied to each row in a manner controlled by Spark. We therefore implemented kNN using such a function. As a result, an increase in efficiency was clearly visible. The average processing time was 20 m 17 s for the whole dataset of 485,322 rows, giving throughput of 399 rows/s.

Brute-force approach with UDF. Another approach that was tested is brute force with UDF. Distance calculations with Euclidean measure do not seem a very complex task. Looking for a weak point, we eliminated the kNN implementation from sklearn, suspecting that the function from this package requires multiple conversions between Python and Scala (which may negatively influence the processing time). Therefore, we implemented the algorithm by 9.3. This approach made it possible to reduce the calculation time even more. It took only 29.4s on average to process all 485,322 rows, yielding throughput of 16,508 rows/s.

KNN approach with Pandas UDF. Looking for further optimization methods, we came across Pandas UDF. It is one of specific features of PySpark, i.e., it is available only in Python, and not in Scala.[6]

5. https://scikit-learn.org/stable/modules/neighbors.html

6. More details are described for example in the blog https://databricks.com/blog/2017/10/30/-introducing-vectorized-udfs-for-pyspark.html

Normally user-defined functions operate on a one-row-at-a-time basis; therefore we suffer from high serialization and invocation overhead. One of the solutions would be to implement these fragile functions in Java or Scala. Apache Spark 2.3 brought a change in API—Pandas UDF, also known as vectorized UDF. It is built on top of Apache Arrow[7] and offers the ability to define high-performance UDFs entirely in Python.

There are two types of Pandas UDFs: scalar and grouped map. Scalar Pandas UDFs are used for vectorizing scalar operations. The function takes in pandas.Series as arguments and returns another pandas.Series of the same size.

Taking this into account, we re-implemented our kNN algorithm with Pandas UDF. In this approach, AIS data is stored in weekly batches. A single batch contains tens of millions of records for the whole world.

To benchmark the approach, we used all AIS data from week 39 of year 2019, which contains 71,799,915 rows. The number of waypoints that we analysed was 5319. Reading a CSV file of 7.5 GB took 12 seconds and it was then split into 60 partitions. The calculation was performed on all 48 CPUs. The operation to assign waypoints to AIS points and write down the data took only 47.3 s on average. Thus, the achieved throughput was 1,517,969 rows/s. This result seemed to be efficient enough for the purpose of our method.

Area filtering and haversine distance. The further experiments were conducted using AIS datasets representing an arbitrary sequence of 8 weeks from 2019. Sample sizes are provided below.

In order to speed up the processing, we additionally restricted the dataset by the analysed area:

- the Baltic Sea + the North Sea + the Norwegian Sea: $50.65 < latitude < 71.50$ *and* $-4.65 < longitude < 35.5$
- the German Bight: $53 < latitude < 57$ *and* $2.5 < longitude < 9.5$

For four weeks from w40 to w43 of 2019, the total number of AIS points was 268,845,453 (Table 9.8). The achieved reduction in number of points (and thus in processing time) is presented in Table 9.9.

Thanks to detailed introspection of distances between close waypoints we discovered that the Euclidean distance used in the kNN calculation from the sklearn library is overly simplified. It is also a matter of the projection used. Therefore, we decided to compare the two approaches.

The first one is the simplified calculation with the Euclidean distance (Minkowski, where $p = 2$) and a cylindrical projection, which is presented in Figure 9.25.

7. Apache Arrow is a cross-language development platform for in-memory data. It specifies a standardized language-independent columnar memory format for flat and hierarchical data, organized for efficient analytic operations on modern hardware.

Table 9.8. Sample sizes of data used in the experiments

2019w36.csv	71,402,966
2019w37.csv	71,472,275
2019w38.csv	72,761,790
2019w39.csv	71,799,915
2019w40.csv	67,340,332
2019w41.csv	69,036,293
2019w42.csv	63,977,977
2019w43.csv	68,490,851

Source: Own work.

Table 9.9. Number of AIS points, by vessel types and filtered areas

Vessel type	Baltic + North + Norway	German Bight
Passenger	9, 918, 152	487,526
Cargo	17, 503, 134	926,288
Tanker	6, 191, 930	284,682

Source: Own work.

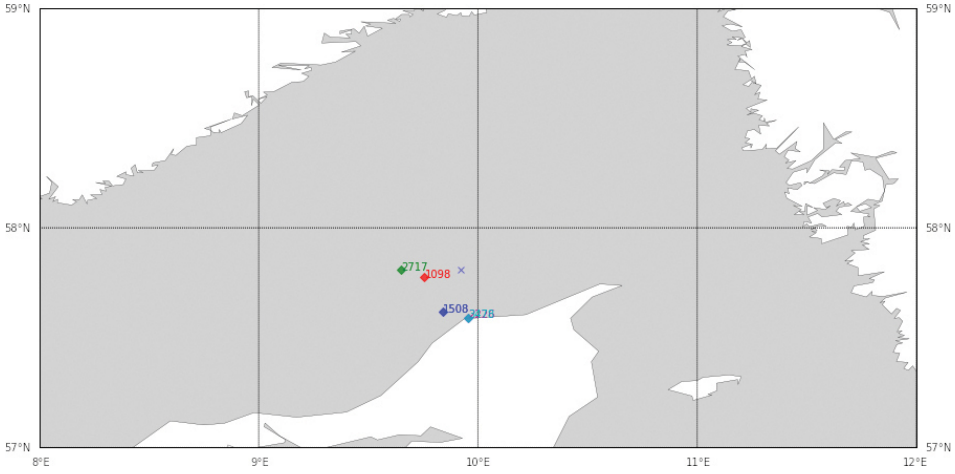


Figure 9.25. Cylindrical projection in distance calculation

Source: Own work.

The aim is to find the waypoints which are the closest neighbour of the point marked with a cross. The order of such waypoints by Euclidean distance is: 1098,

1508, 2226, 3478, 2717. However, due to the cylindrical projection, for high north latitudes the error in calculation of the Euclidean distance can result in a wrong order of waypoints. Indeed, the same coordinates drawn using the Albers Equal Area (AEA) projection show a different situation—see Figure 9.26.

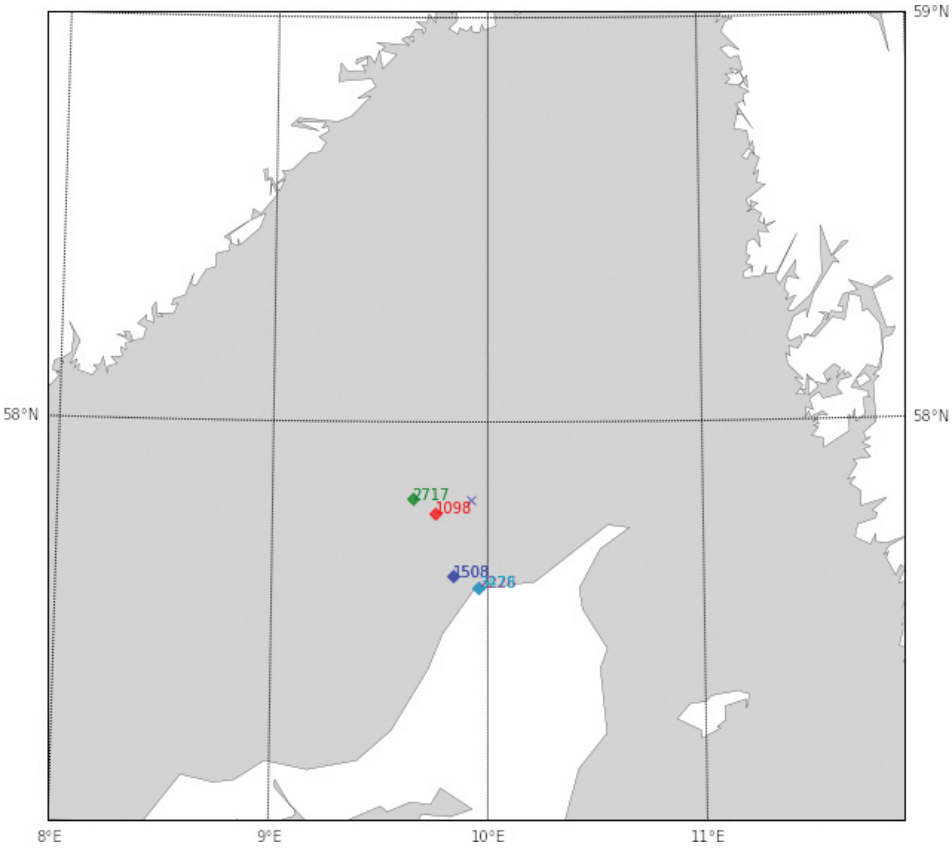


Figure 9.26. Albers Equal Area projection in distance calculation

Source: Own work.

One of the metrics available in sklearn's Nearest Neighbor function is haversine.[8] It provides a much more precise distance between two points on the globe than the Euclidean distance, at the cost of efficiency, though. For the same example as above, the order of waypoints by the haversine metric is: 1098, 2717, 1508, 2226, 3478. The greatest difference concerns waypoint 2717, which advanced from 5th place to 2nd. The distance in km for the respective waypoints is: 10.516, 16.105, 21.754, 24.154, 24.156. We also confirmed the distances by precise geopy.distance[9]

8. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html
9. https://geopy.readthedocs.io/en/stable/#module-geopy.distance

calculation, which gives the following results: 1098–10.550 km, 2717–16.162 km, 1508–21.791 km, 2226–24.193 km, 3478–24.195 km. As a result, it confirmed that the haversine metric provides more precise distance calculation than Minkowski's Euclidean-based metric.

However, the application of haversine metric instead of Minkowski's resulted in a significant drop of processing performance. In order to use the haversine metric we had to switch to another neighbour search algorithm—the BallTree.[10] We also observed the dependency of calculation time on the number of waypoints—the more waypoints, the longer it took to find the nearest waypoints. So, the drop in performance was caused both by the change in the metric and partly by the partition algorithm. Table 9.10 presents sample benchmarking results.

Table 9.10. Performance of haversine distance with filtering by vessel types and areas

Filter	No. of AIS points	Processing time	Performance
German Bight, cargo	926, 288	16m 7s	958 rows/s
German Bight, passenger	487, 526	10m 31s	773 rows/s
Baltic + North, cargo	17, 503, 134	5h 59m 25s	812 rows/s
Baltic + North, tanger	6, 191, 930	2h 17m 7s	753 rows/s

Source: Own work.

KNN summary. Let us summarize the evolution of the methods that were applied to optimize AIS data processing while searching for the nearest waypoint and calculating the distance—see Table 9.11.

Table 9.11. Performance of haversine distance with filtering by vessel types and areas

Method	Troughput (rows/s)
kNN, iteration over RDD with flatMap	7.3
kNN, with UDF	399
brute-force search, with UDF	16,508
kNN Minkowski, with Pandas UDF	1,517,969
kNN haversine, with Pandas UDF	824

Source: Own work.

Summing up, it seems that the most efficient approach for assigning the nearest waypoint to each AIS row and calculating the distance to it is kNN Minkowski, with

10. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.BallTree.html#sklearn.neighbors.BallTree

Pandas UDF. Nevertheless, for the sake of precision in the further experiments kNN with the haversine metric and BallTree partitioning were used.

9.2.6. Reconstruction of edges

Having all AIS points annotated with the nearest waypoints, the next step is reconstruction of the edges between these waypoints. The algorithm for this approach is presented in Section 5.3.3. In this case, due to incomplete distribution and often low quality of AIS data, several measures need to be undertaken to achieve good results. Fortunately, generation of edges proved to be a less challenging task from the performance point of view. The only optimization step that had to be applied was materialization of the enriched AIS dataset. For some reason even caching was not helpful—grouping the edges spawned a re-calculation of the closest waypoints. Therefore, our process is divided into two steps:

(1) from raw AIS data to enriched AIS—results are stored in CSV files;
(2) from enriched AIS data (read from the CSV file) to the edges—results are stored in two files: nodes.csv and edges.csv, representing the mesh.

Nevertheless, there were other challenges concerning the output mesh. A visual introspection of maps, which show the generated mesh, proved that the method generated 'impossible' or 'inappropriate' connections between some waypoints which further on had to be eliminated. It was caused partly by the low AIS data quality. However, other means were undertaken to improve the final mesh. Some of the applied techniques are presented below.

For all the tasks presented below we used AIS data from 8 consecutive weeks (2019 w36–w43). AIS data was filtered, so that only AIS from the German Bight for tankers, cargo and passenger ships were included. For this input data, 8,809 waypoints were identified. Input AIS data contained 3,639,631 rows, in which 4,857 distinct MMSIs were found. The key manoeuvre points identified with the CUSUM method contained 414,824 rows, in which 4,609 distinct MMSIs were found.

Edges calculated based on the full AIS data (border points). When a vessel is moving along its trajectory, it passes many waypoints. We know which points are passed by, as AIS data is already annotated with the closest waypoint (see Section 9.2.5). Sometimes there are several consecutive AIS messages with the same waypoint, especially if the distances between the waypoints are long. We need to identify only the places where the 'borders' between affiliation of AIS to different waypoints are crossed, i.e., a given message has a different waypoint from the previous message.

In the implementation of the algorithm, the effect described above is achieved by using the so-called window functions. In these functions it is possible to refer to the previous value with function lag. We are then able to identify the 'changed' rows as described in the listing below:

```
from pyspark.sql.window import Window

w = Window.partitionBy("mmsi").orderBy("timestamp_ais")
sdf_ais_with_waypoint_changed = sdf_ais_with_waypoint_idx \
.withColumn("from_waypoint", F.lag('to_waypoint', 1, 0).over(w))
sdf_ais_with_waypoint_changed = sdf_ais_with_waypoint_changed \
.withColumn("changed",
↪  (sdf_ais_with_waypoint_changed['from_waypoint'] !=
sdf_ais_with_waypoint_changed['to_waypoint']).cast('int'))
sdf_ais_with_waypoint_filtered = sdf_ais_with_waypoint_changed \
.withColumn("timestamp_delta",
↪  sdf_ais_with_waypoint_idx.timestamp_ais-F.lag(
'timestamp_ais', 1, 0).over(w)) \
.where('changed=1') \
.where('from_waypoint<>0')
```

By applying the above procedure, we reduced the initial 3,639,631 messages to the filtered 1,494,227 messages. They contain only the points where a current waypoint (to_waypoint) is different from the previous waypoint (from_waypoint). We can construct a dataset with edges using grouping by from_waypoint and to_waypoint, as illustrated in the code below. We also calculate group statistics like the number of vessels traversing specific edges or time-related stats.

```
sdf_edges = sdf_ais_with_waypoint_filtered \
.groupBy("from_waypoint", "to_waypoint") \
.agg(F.count("*").alias("cnt"),
F.avg("lon").alias("lon"),
F.avg("lat").alias("lat"),
F.avg("timestamp_delta").alias("avg_time"),
F.min("timestamp_delta").alias("min_time"),
F.max("timestamp_delta").alias("max_time"),
F.stddev("timestamp_delta").alias("stddev_time"))
```

In this specific example we generated 170,644 edges between 8,809 waypoints. The visualization of this mesh on the map is presented in Figure 9.38 (p. 294).

Analysis of distance on edges. By looking at the Figure 9.38 (p. 294), we observe a big number of edges that span long distances. Having been visualized on the map, they very often cross the land. Therefore, we decided to study in detail the lengths of the edges to identify and possibly eliminate the problem.

In Figure 9.27 we demonstrate the histogram of edges lengths. Please note that they y-axis is logarithmic. There are almost 50 edges that span two waypoints that are at least 500 km apart. It reveals the weakness of the approach.

Therefore, we had to adjust the approach to eliminate the longest edges. It was done by adding a function FilterEdges (see Algorithm 5.1, as applied for the visualization of the meshes presented in Figures 9.39 (p. 295) and 9.40 (p. 296).

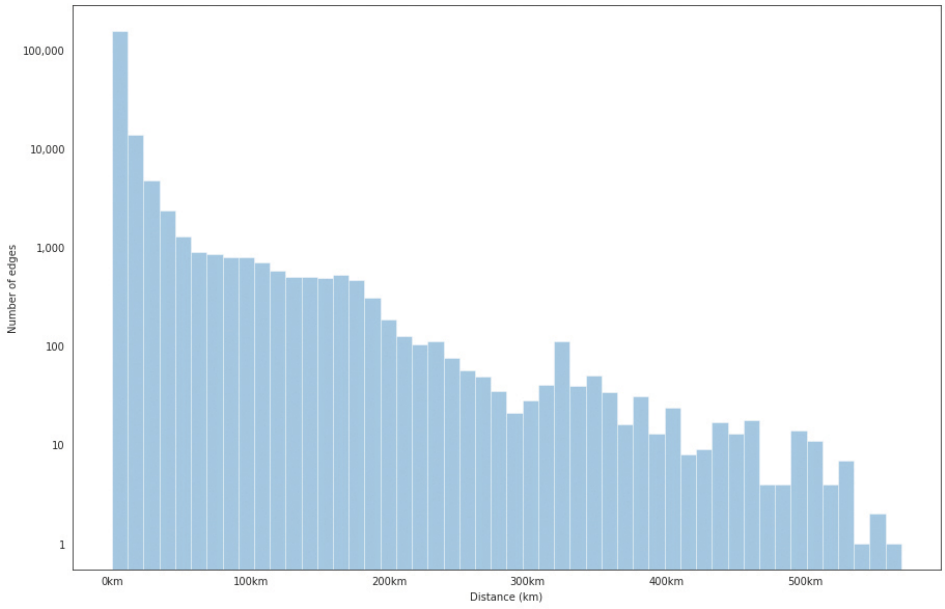


Figure 9.27. Distance between edges in a mesh for all ships in the German Bight

Source: Own work.

Analysis of timestamp delta. AIS data is timestamped. When we analyse a specific trajectory of a given vessel, we can measure the time that passed between two consecutive messages. It is called a timestamp delta, in code referred to as timestamp_delta.

Having calculated the time necessary to pass from one waypoint to another (column timestamp_delta), it should be possible to propose the fastest route. Unfortunately, vessels do not go from waypoint to waypoint. Instead, they go between some locations that are nearby the waypoints. Moreover, when aggregated, there is no guarantee that time will be measured between the same points.

We conducted an analysis of timestamp deltas. To present the results of this analysis, below we show a series of histograms, as a single chart is not able to provide enough details. Figure 9.28 presents the overall histogram for all the data from the 8-week period. We see that there are several trajectories that contain gaps of more than 1,000 hours between the messages. The number is not significant but it still can be filtered. The majority of deltas, i.e., more than 1,000,000, still concentrate around zero.

In order to see the details, we need to zoom in the x-axis and show data only for 24 hours. Figure 9.29 presents the results with an increased resolution. We can observe that after filtering longer deltas the remaining messages are not very separate. We need to increase the resolution once more.

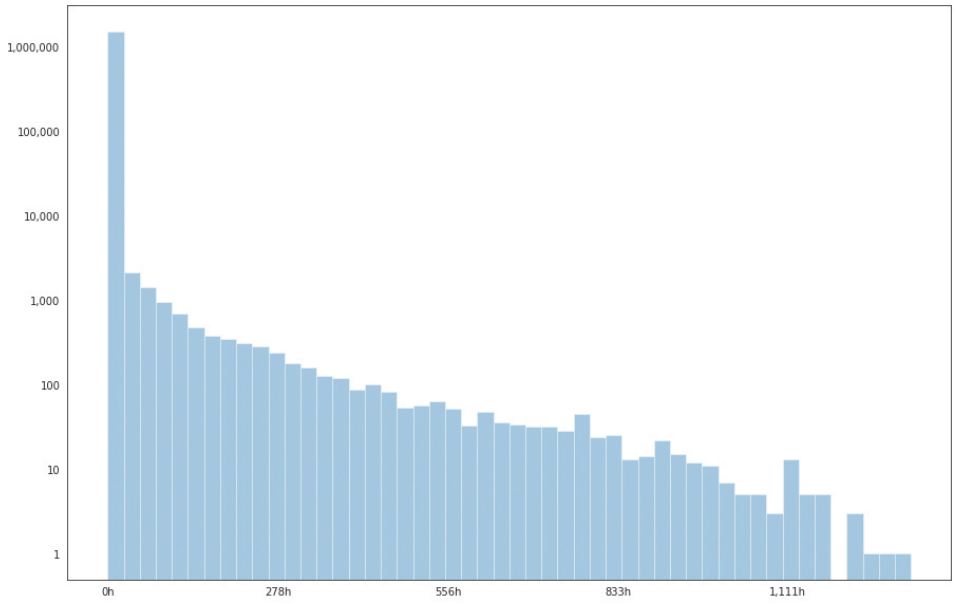


Figure 9.28. Timestamp delta calculated for the whole 8-week period

Source: Own work.

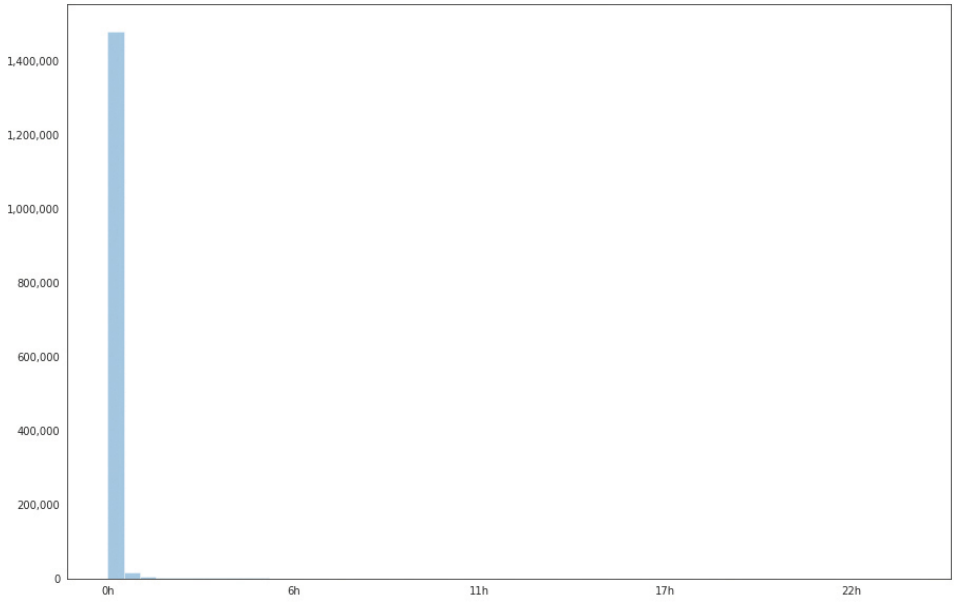


Figure 9.29. Timestamp delta restricted to 24 hours

Source: Own work.

Figure 9.30 presents timestamps deltas for messages that appeared within a period of one hour. The chart reflects the expected distribution of time differences between messages.

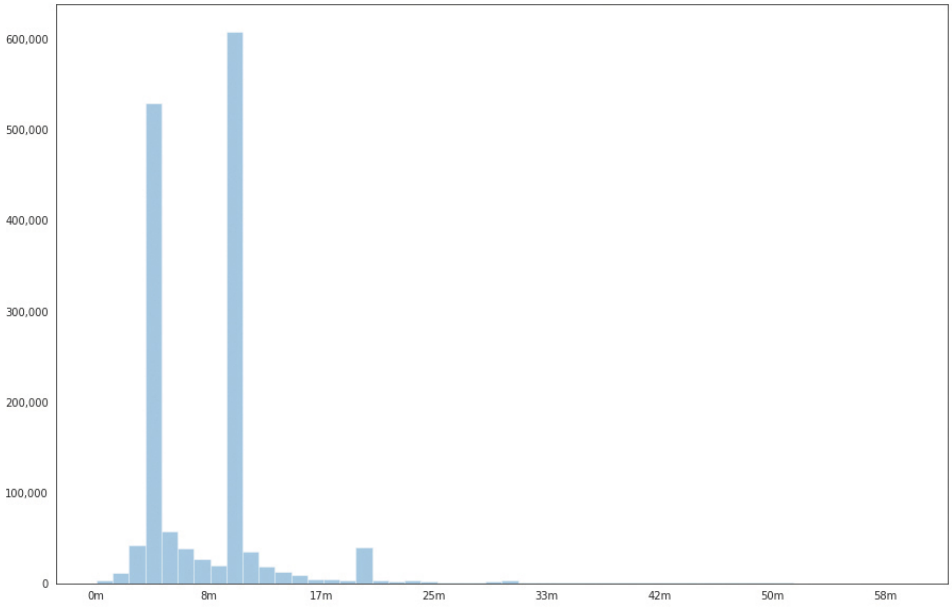


Figure 9.30. Timestamp delta restricted to 1 hour

Source: Own work.

The final chart shows which edges span long distances—see the dark lines in Figure 9.31. These are mostly edges close to the boundaries of the considered area, so it may mean that a vessel left the area and then came back. Thus, a more careful filtering or segmentation is necessary.

To conclude, such a distribution of timestamps suggests that we can safely filter out outliers, i.e., AIS messages that are too far away from each other to form a trajectory. Thus, we can also avoid joining the waypoints that are too far away (or at least are not neighbours).

If we combine two phenomena—imprecise calculation of time deltas and long-distance edges—we also observe anomalies in the average speed as it is calculated as distance divided by time. Figure 9.32 presents the histogram of the average speed. The calculation was conducted for all waypoints.

Edges between minimum distance points (mindist). The analysis conducted in the previous paragraph revealed that more realistic time deltas between waypoints are needed. Our previous approach correctly identified the transition from one waypoint (a waypoint segment to be more precise) to another. We referred to them

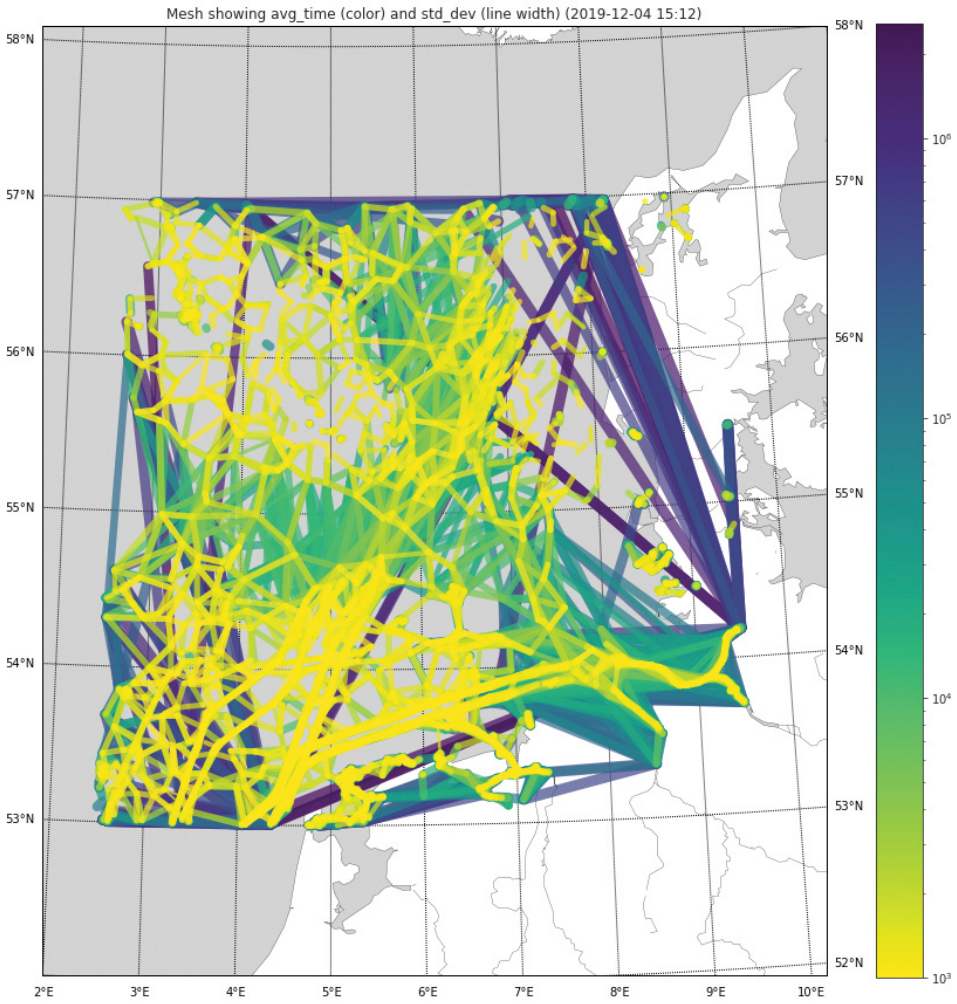


Figure 9.31. Mesh showing an average traveling time in seconds (colour) and standard deviation (width)

Source: Own work.

as border points because we identified only pairs of points located close to the border.

The idea behind the next approach was to find other points that would be more suitable to serve as representatives of waypoints. Out of all the possible representatives in a given segment, we do not choose ones that are in the vicinity of the border, but those that are the closest to the waypoint, hence mindist—minimum distance.

The two approaches are illustrated in Figure 9.33. We track a vessel through the segments. Both subfigures show the trajectory of a vessel, with empty circles

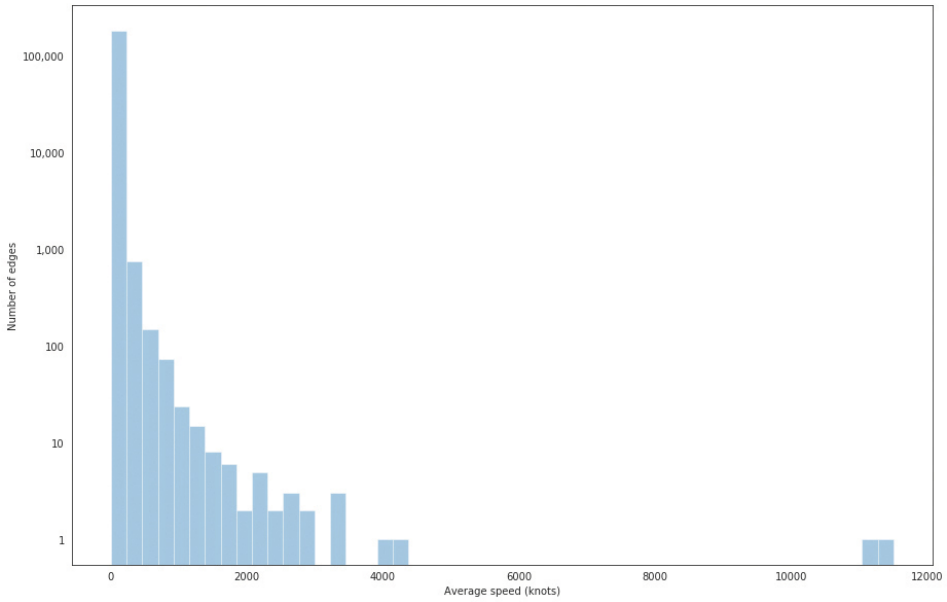


Figure 9.32. Average speed as calculated from generated edges

Source: Own work.

marking positions as found in AIS data. The vessel is moving left to right from segment 1 to segment 2. The segments are delineated based on a distance to the waypoints. The left subfigure shows the initial idea, referred to as border points. All points within a single segment are annotated with the same waypoint. When we detect a change of a segment, we select green and red points as representatives of waypoints and calculate the time difference ('Delta 1'). The right subfigure illustrates a refined approach. Basically, the transition between the same waypoints is determined. However, now we have other points for measuring the difference in time ('Delta 2'): the green point is the closest point to waypoint 1, and the red point is the closest point to waypoint 2.

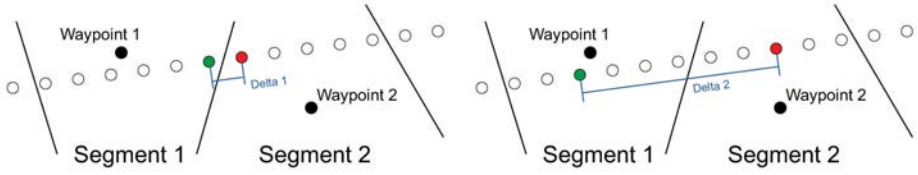


Figure 9.33. Two approaches to determine representative points within segments

Source: Own work.

The revised approach is implemented in the listing below. The closest point is marked by a boolean column mindist.

```
from pyspark.sql.window import Window

w = Window.partitionBy("mmsi").orderBy("timestamp_ais")
sdf_ais_with_waypoint_changed = sdf_ais_with_waypoint_idx \
.withColumn("from_waypoint", F.lag('to_waypoint', 1, 0).over(w))
sdf_ais_with_waypoint_changed = sdf_ais_with_waypoint_changed \
.withColumn("changed",
 ↪  (sdf_ais_with_waypoint_changed['from_waypoint'] !=
sdf_ais_with_waypoint_changed['to_waypoint']).cast('int')) \
.withColumn("segment", F.sum("changed").over(w))

w2 = Window.partitionBy("mmsi",
 ↪  "segment").orderBy(F.col("dist_to_wp_km"))
sdf_ais_with_waypoint_mindist = sdf_ais_with_waypoint_changed \
.withColumn("row",F.row_number().over(w2)) \
.where("row==1").drop("row") \
.withColumn("timestamp_delta",
F.col("timestamp_ais")-F.lag('timestamp_ais', 1, 0).over(w)) \
.where("from_waypoint<>0 and timestamp_delta<1500000000")
```

After applying the revised approach, the resulted dataset contains 1,525,419 rows. The number of generated edges is 183,368.

Edges from time-bound messages (tbound). As a result of the analysis on timestamp deltas, we introduced an additional filtering of vessel-related trajectories (implementation of function FilterTrajectory—see Algorithm 5.1. We observed that many problematic edges stem from the long delays between AIS messages. At this stage, the aim is to choose only those passes of vessels between waypoints where the time between the messages is restricted, i.e., the edges will be time-bound.

As the focus is on more realistic time deltas between waypoints, we assumed that the time period between two consecutive messages connecting two waypoint areas should not be longer than 15 minutes. Implementation of the approach is presented in the listing below.

```
sdf_ais_with_waypoint_changed = sdf_ais_with_waypoint_changed \
.withColumn("changed",
 ↪  (sdf_ais_with_waypoint_changed['from_waypoint'] !=
sdf_ais_with_waypoint_changed['to_waypoint']).cast('int')) \
.withColumn("tbound",
 ↪  (sdf_ais_with_waypoint_changed['timestamp_delta'] <=
TBOUND).cast('int')) \
.withColumn("segment", F.sum("changed").over(w))
sdf_ais_with_waypoint_tbound = sdf_ais_with_waypoint_changed \
```

```
.where('changed==1 and tbound==1') \
.where('from_waypoint<>0')
```

By applying the above approach, we further reduced the number of rows used to generate edges. We obtained 1,408,451 rows compared to 1,494,227 rows of the baseline method (border points). The reduction does not seem to be significant—just 85,776 rows. Nevertheless, the number of resulting edges was reduced significantly from 170,644 (border points) to 156,103 (tbound). The reconstruction indeed avoids too long hops. The visualization of the mesh on the map, generated using the tbound approach, is presented in Figure 9.41 (p. 297).

Edges based on CUSUM. The construction of edges based on the CUSUM result is very similar to the construction of edges based on the full AIS data. One workaround was necessary though. The results provided by the CUSUM algorithm (see Section 9.2.2) are missing the AIS timestamp column,[11] therefore calculation of time deltas is not possible. Moreover, window operations require data to be sorted. Therefore, we first add an artificial column and then perform calculations as in the previously presented approach. The applied operations are presented in the listing below.

```
from pyspark.sql.window import Window

sdf_cusum_waypoint_idx = cusum_waypoints_sdf \
.select("mmsi", "segment", "lon", "lat", "dist_to_wp_km",
F.col("waypoint").alias("to_waypoint")) \
.withColumn("order", F.monotonically_increasing_id())
w = Window.partitionBy('mmsi', 'segment').orderBy("order")
sdf_cusum_waypoint_changed = sdf_cusum_waypoint_idx \
.withColumn("from_waypoint", F.lag('to_waypoint', 1,
↪  0).over(w))
sdf_cusum_waypoint_changed = sdf_cusum_waypoint_changed \
.withColumn("changed",
↪  (sdf_cusum_waypoint_changed['from_waypoint'] !=
sdf_cusum_waypoint_changed['to_waypoint']).cast('int'))
```

In the presented example, the input data (results of the CUSUM algorithm) consisted of 414,824 rows. It is a significant reduction from around 3 million rows (as it is usually 10%–15% of all data). After detecting changes and filtering, according to the operations presented in the above listing, we got 270,528 rows. It is a much smaller number of points than in the border point approach, where we had 1,494,227 messages after filtering. Consequently, the number of the edges is also smaller—133,032. A visualization of the mesh on the map, generated according to the CUSUM approach, is presented in Figure 9.42 (p. 298).

11. This is a design decision, and it can be changed if needed.

9.2.7. Maritime traffic network evaluation

There is no strict methodology that would enable an evaluation of the proposed approaches for mesh generation. On the one hands, the results depend on the waypoints provided (the quality of the results of the previous steps, i.e., CUSUM and genetic algorithm). On the other hand, the output is used to construct a recommended corridor and some issues can be alleviated by an appropriate design of the method. What can be done is to make sure that each separate step of the method provides meaningful and good-looking results. Therefore, in the next paragraphs various visualizations are presented to illustrate and compare the results of different approaches.

Distance between waypoints. First, we want to analyse the distance between waypoints and its distribution. In general, it assesses how good the input data is. Waypoints which are too close mean that many probably unnecessary edges will be generated. Waypoints too far from each other can result in a loss of precision. This is contextual, as we need a different distribution in straits and on the high sea.

We conducted an analysis of the waypoints that were generated for weeks 36 to 43 in 2019 for the German Bight area, for all three types of ships: passenger, cargo, and tanker. In total 8,809 waypoints were analysed.

If we took all pairs of waypoints, we would need to calculate $8809 \times 8808/2 = 38,794,836$ distances. This would take too long. Therefore, we decided to calculate distances to the 4 nearest neighbours, which requires calculating of $4 \times 8809 = 35,236$ distances.

The distribution of the distances between the waypoints is presented in Figure 9.34. It must be noted that the y-axis is logarithmic. There are a lot of close waypoints, i.e., for over 10,000 pairs (out of 35,236) the distance is close to zero.

In order to study the distances in detail, we also provide a 'zoomed' figure where the distance was limited to 1 km—see Figure 9.35. Still, there is a huge number of almost overlapping waypoints (the distance close to zero). It creates a space for further improvement of the waypoint generation steps (i.e., CUSUM and the genetic algorithm).

Delaunay triangulation. The task of finding a waypoint that is closest to a given AIS position can be interpreted as a problem of finding all the points annotated with a given waypoint. This problem is thus closely related to the Voronoi tessellation (Voronoi, 1908). It is about identifying regions, called the Voronoi cells, so that for each seed there is a corresponding region consisting of all points of the plane closer to that seed than to any other. Here, waypoints play the role of seeds.

The Voronoi tessellation has a dual task. If we connect seeds, here waypoints, of the adjacent regions we obtain a mesh that shows how to move directly from one region to the other without crossing any other region. In fact, the mesh consists

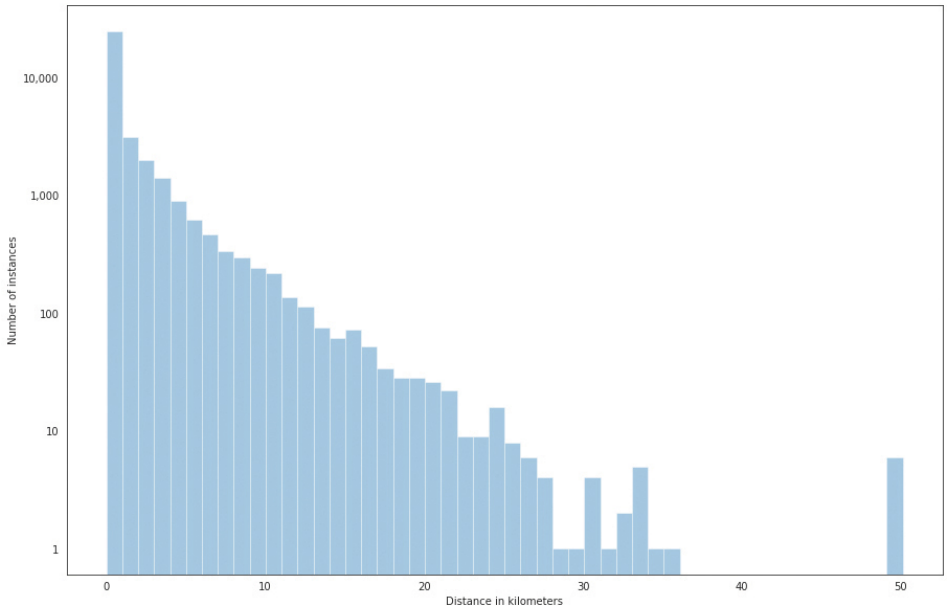


Figure 9.34. Distribution of distances between waypoints for the 4 nearest neighbours

Source: Own work.

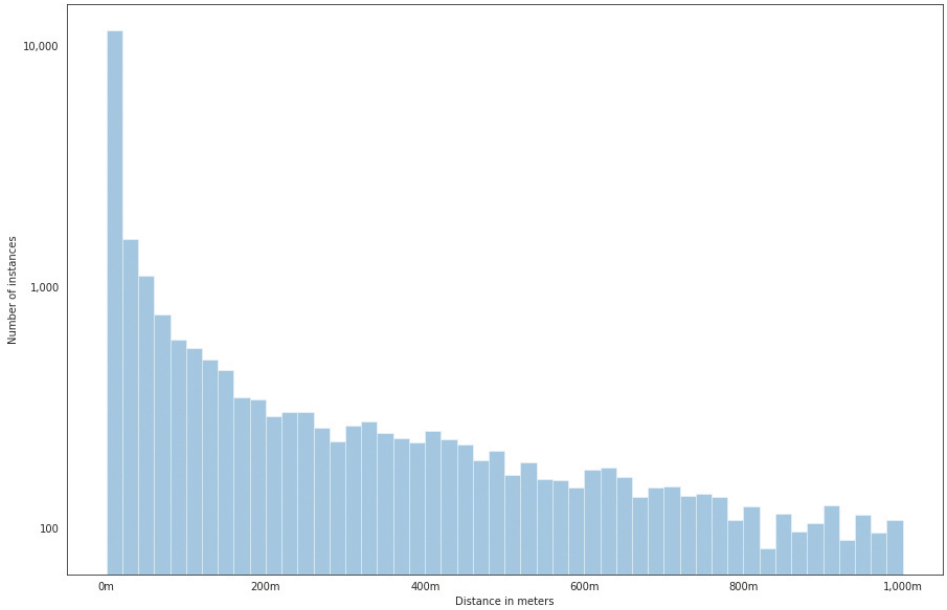


Figure 9.35. Distribution of distances under 1 km between waypoints for the 4 nearest neighbours

Source: Own work.

of triangles and the process is called the Delaunay triangulation (Aurenhammer, Klein, & Lee, 2013). The Delaunay triangulation is a particular way of joining a set of points to make a triangular mesh. The circumcentres of the Delaunay triangles are the vertices of the Voronoi tessellation. Looking at the geometric properties, it is such a triangulation that no point in a set of input points is inside the circumcircle of any triangle. The Delaunay triangulations maximize the minimum angle of all the angles of triangles, thus it tends to avoid sliver triangles.

The Delaunay triangulation for a set of the above mentioned 8,809 waypoints is presented in Figure 9.36. It is interesting to observe the traffic separation zones. If we were to construct a recommended route, it should follow the edges from the triangulation.

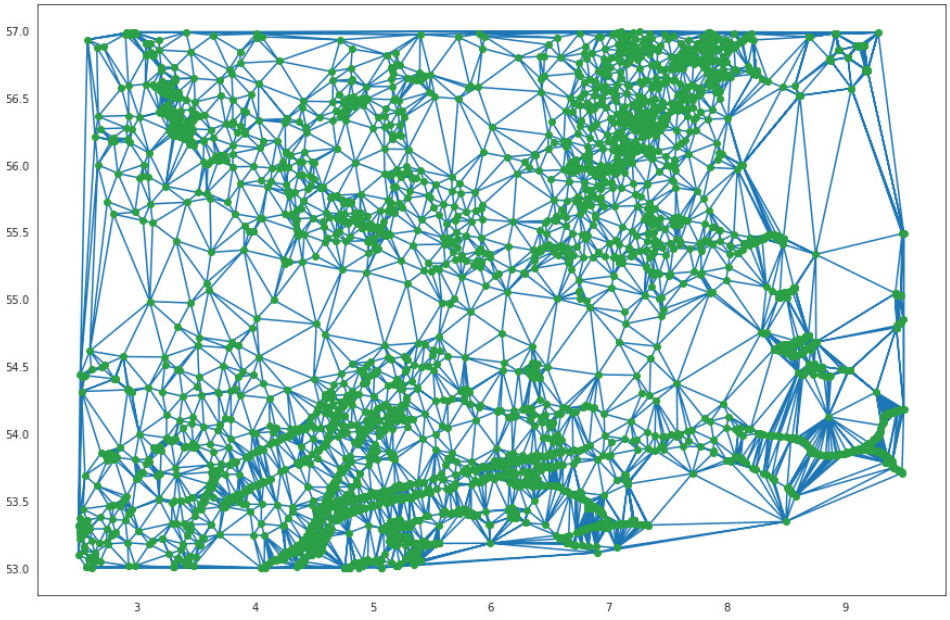


Figure 9.36. The Delaunay triangulation for waypoints in the German Bight

Source: Own work.

The Voronoi seeds are connected via edges that can be derived from adjacency relationships of the Delaunay triangles, i.e., if two triangles share an edge in the Delaunay triangulation, their circumcentres are to be connected with an edge in the Voronoi tessellation. The Voronoi tessellation, although less impressive, is presented in Figure 9.37.

Mesh visualization based on full AIS data. The sample visualizations of meshes presented here are prepared with the use of the methods for edges calculation using

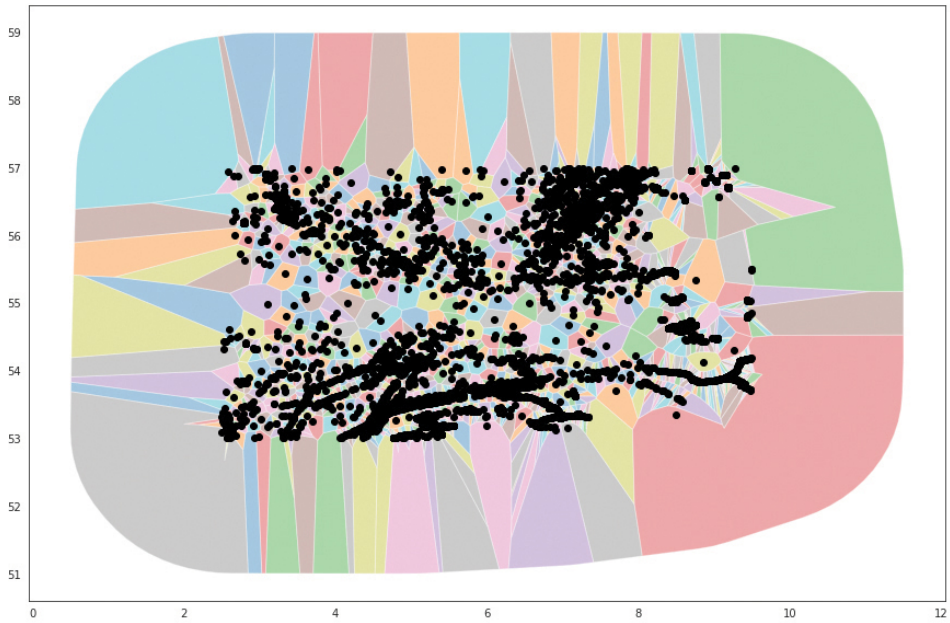


Figure 9.37. The Voronoi tessellation for waypoints in the German Bight

Source: Own work.

border points. The meshes are based on the full AIS data (without any filtering). Further on, filtering is applied to the edges, as described in specific figures below.

Generation of edges for the full AIS data brings a lot of unwanted edges. These are spurious edges, usually spanning distances which are too long. It is a consequence of distance distribution as presented in Figure 9.27. The problems in question can be seen in Figure 9.38[12]. Although such a mesh can be a good option for planning, it does not actually bring any generalization for the maritime picture.

In order to improve the mesh, we conducted a filtering of edges. First, spurious edges were removed, i.e., the edges generated based on just a single trajectory. It means that it is now required that an edge is created by more than one ship; this parameter will by further referred to as a count (cnt). The mesh generated based on the filtered AIS data, where count > 1 and additionally the distance is shorter than 350 km, resulted in 105,094 edges. The maximum count, i.e., the number of trajectories that go through the most popular edge, was 1,393. This maximum value determines the scale used in a visualization, which for clarity is logarithmic. The visualization of this mesh can be found in Figure 9.39.[13]

12. It is interesting to note that such a huge number of edges (around 180,000) required a lot of time for visualization. The presented map was generated in 3 h 44 min 45 s.
13. Generation time: 4 h 6 min 48 s

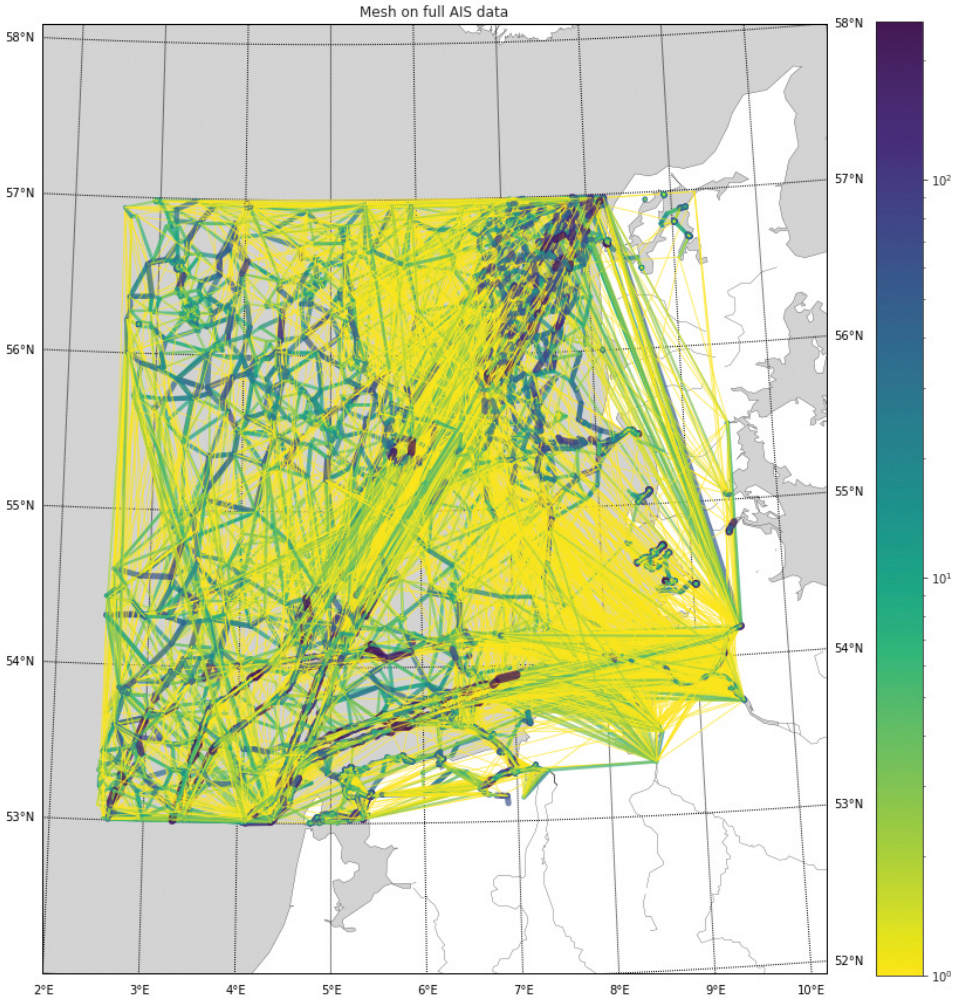


Figure 9.38. Mesh for the German Bight generated on full AIS data for the 8-week period

Source: Own work.

By restricting the set of the edges even further, where count > 10 and distance is < 180 km, we obtain a less appealing mesh, as shown in Figure 9.40. Again, the problem is the quality of the AIS data—only areas with a good AIS coverage look good in the figure. There is an empty space in the middle of the area, where (as it seems obvious) the ships should pass through, but due to the lack of AIS data in this area, they are not contributing to the edges presented in the figure. This issue was one of the motivations for further analysis of the balance flow (see one of the next paragraphs).

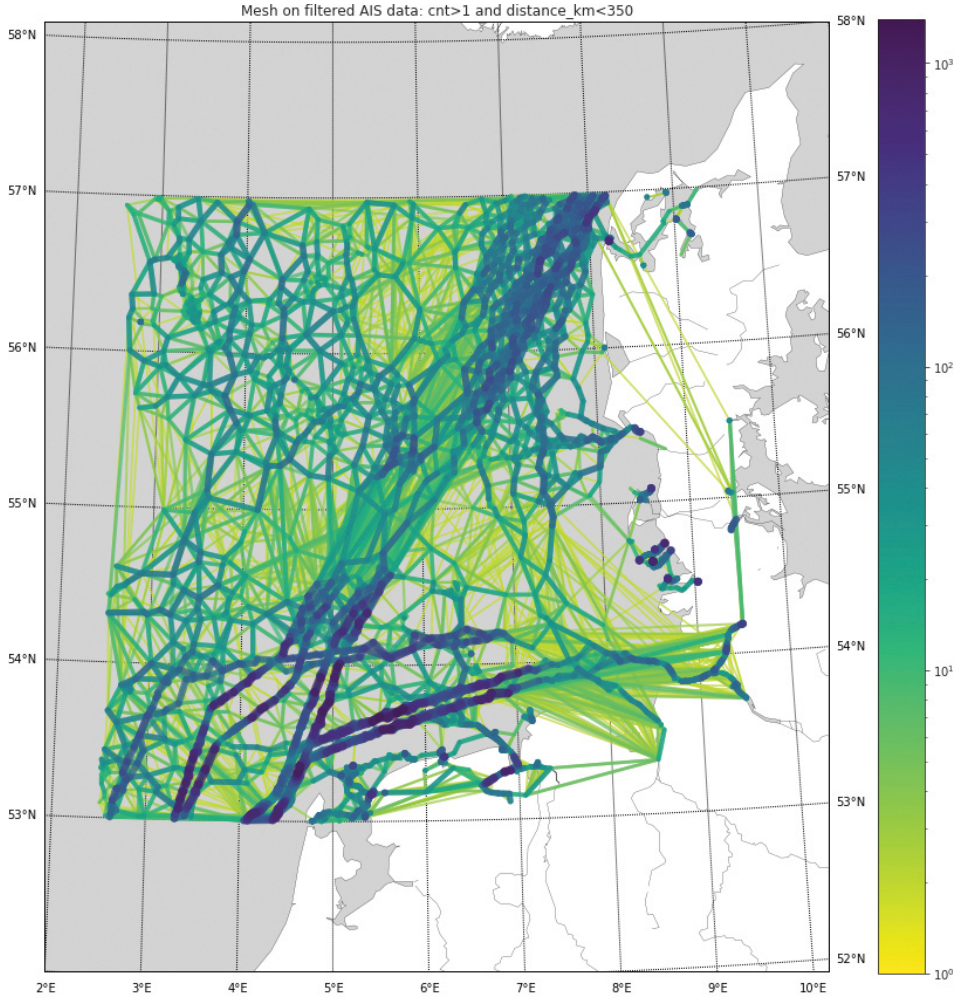


Figure 9.39. Mesh for the German Bight generated on AIS data for 8-week period, filtered by count of contributing edges > 1 and distance shorter than 350 km

Source: Own work.

Mesh based on tbound AIS data. The sample visualization of a mesh presented here is prepared using the methods for edges calculation based on time-bound messages (tbound). The edges connect only consecutive messages captured within 15 minutes from each other. Therefore, we do not observe long edges, even without additional filtering by distance.

Figure 9.41 shows an example of such a mesh. It seems that the mesh looks much better than the ones prepared on the full AIS data, even with filtering. Long connections were eliminated. At the same time, it was possible to keep the edges

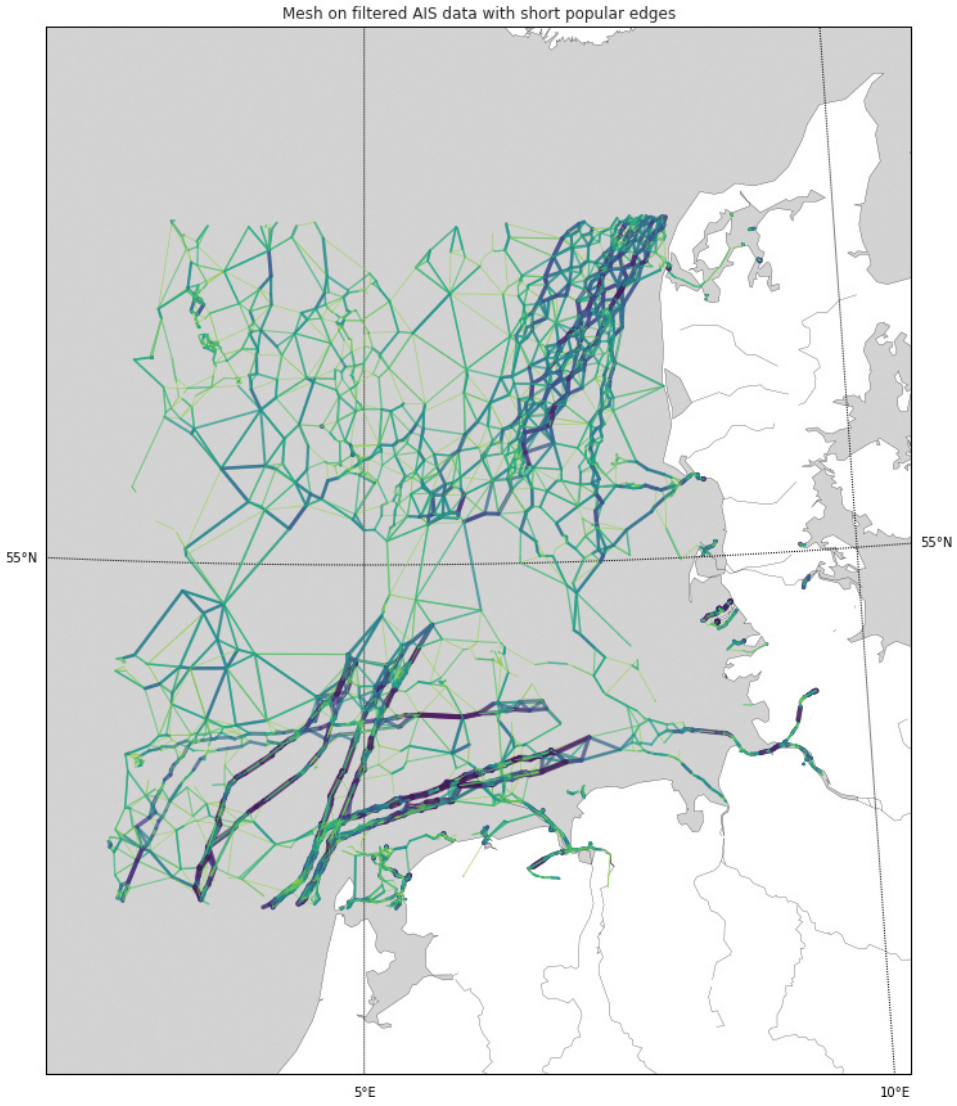


Figure 9.40. Mesh for the German Bight generated on AIS data for 8-week period, filtered by count of contributing edges > 10 and distance shorter than 180 km

Source: Own work.

that were passed by just a single ship. The south-east area, that is close to important ports in Germany, was finally filled in with edges.

Mesh based on CUSUM data. The sample visualizations of a mesh presented here is prepared using the methods for edges calculations based on the CUSUM results.

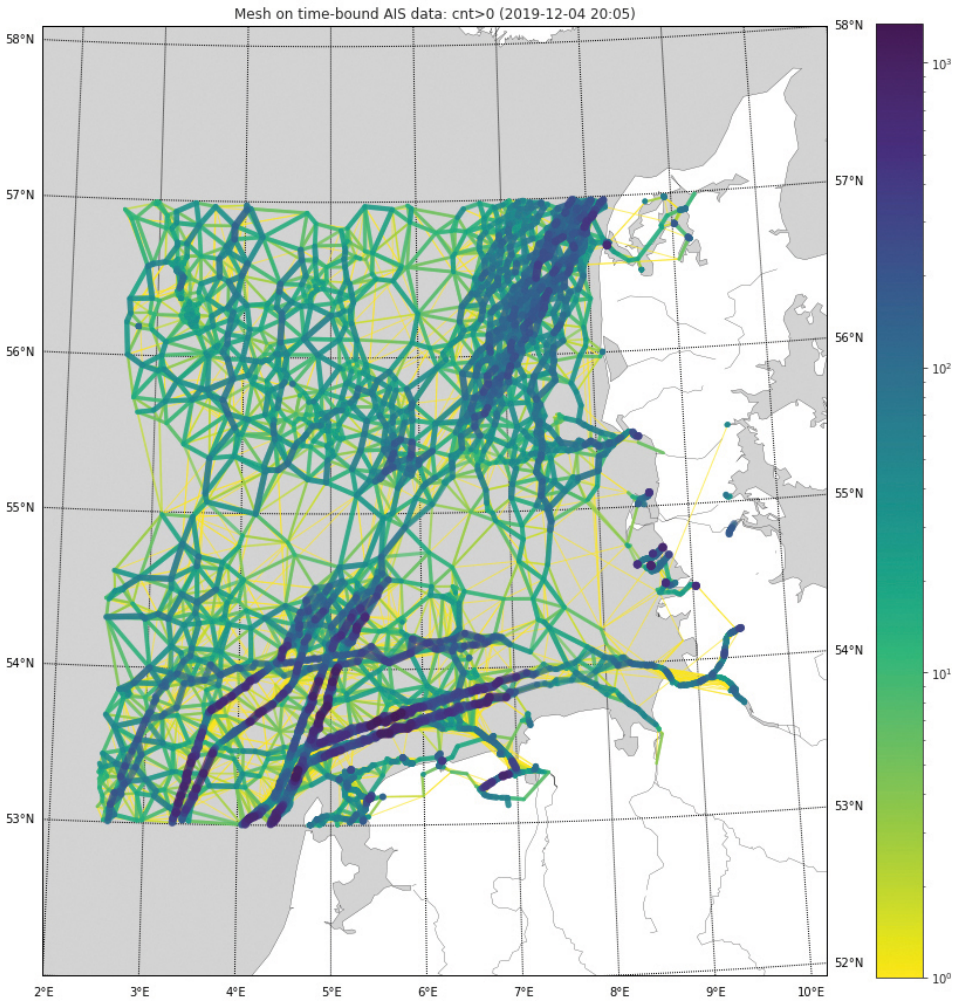


Figure 9.41. Mesh for the German Bight generated on AIS data for 8-week period, filtered by time delta between messages no longer than 15 minutes

Source: Own work.

There is a preference for important manoeuvre points, so the edges should reflect how ships are actually navigated.

The significantly smaller number of input messages results in a lower number of connections. The most popular edge has only 96 counts, which means that we were able to identify only 96 trajectories that contained the edge of interest, compared to around 1300 in the case of the full AIS data.

Figure 9.42 reveals another problem—there are several dead-ends, especially in the area of weak AIS coverage. Such a mesh is problematic for route planning as

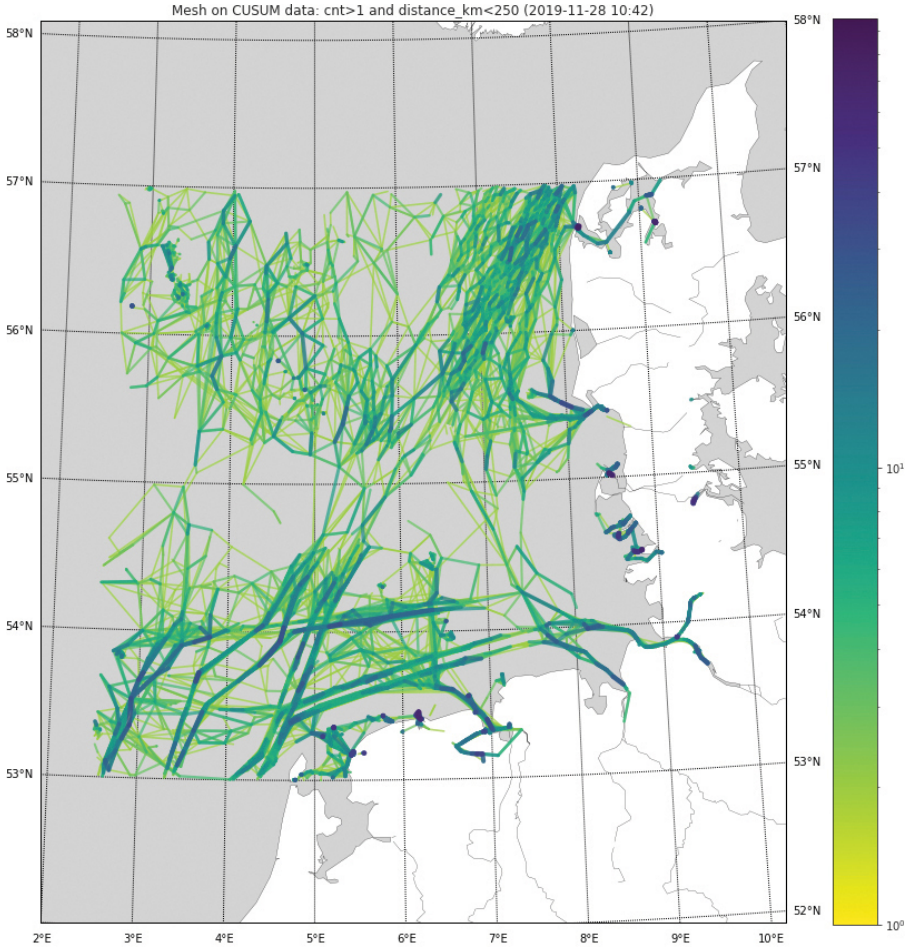


Figure 9.42. Mesh for the German Bight generated on CUSUM data for 8-week period

Source: Own work.

the shortest route, according to the algorithm, will not be the shortest if 'straight lines' are allowed on the high sea.

Mesh with directions. Looking at the south part of the visualizations of the meshes shown in the previous paragraphs, we observe a specific pattern—it resembles a dual carriageway. Therefore we analysed if any of the directions dominates in the traffic.

In Figure 9.43 we use the following colour coding:

- red: vessels moving north,
- blue: vessels moving south.

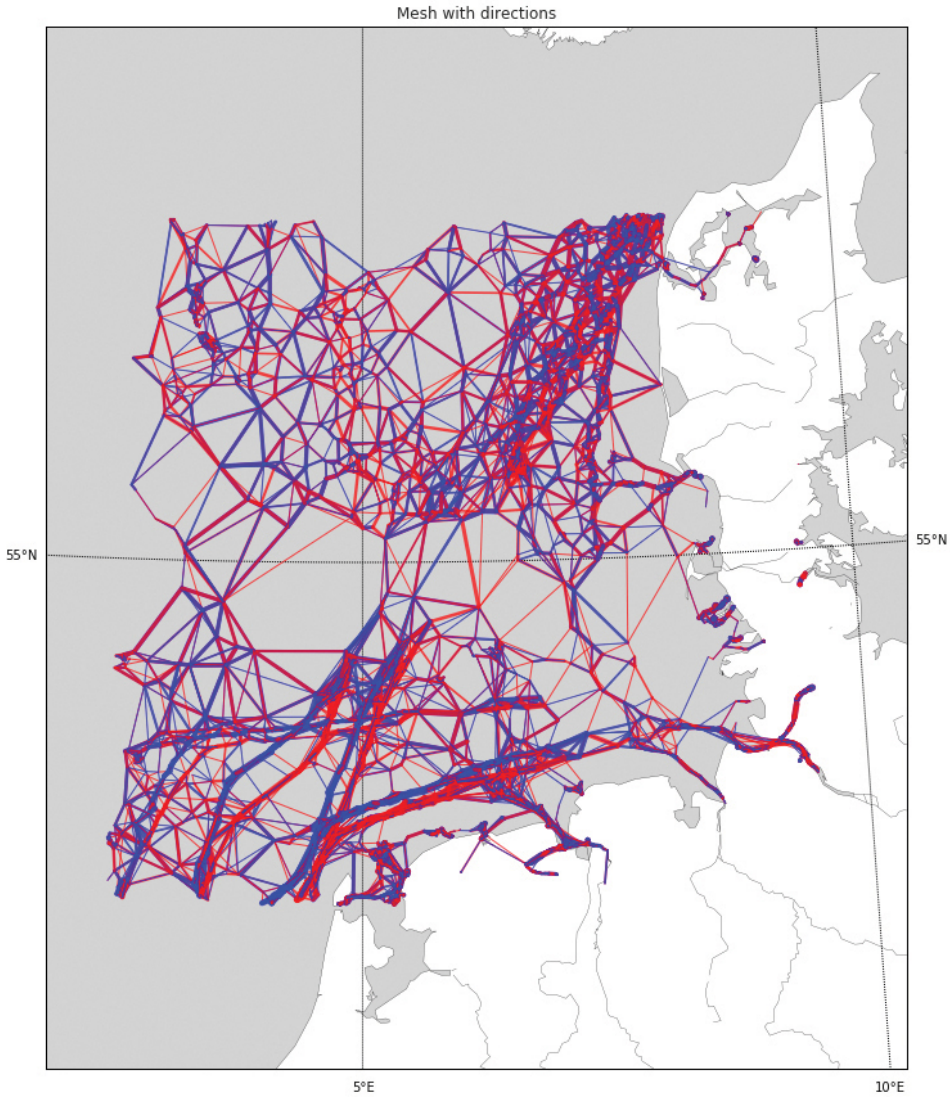


Figure 9.43. Directed mesh for the German Bight generated on AIS data for 8-week period

Source: Own work.

Traffic separation schemes are clearly visible here. When we look at the mentioned pattern, we can observe that the right carriageway is moving north, and the left one is moving south.

As for the north area, apparently no regulations are in force therefore we observe a rather random interweaving of routes.

Analysis of a waypoint cell. The goal is to evaluate the *mindist* approach as compared with the *border points* approach.

First, please note that a waypoint cell is equivalent to a Voronoi cell, but it shows only real AIS points. If there are enough AIS points, the 'dotted' cell will show the real boundaries between the waypoints.

For demonstration purposes, we selected one of the waypoints lying in an empty area (in the middle of the German Bight with a weak AIS coverage)—waypoint 3648. We needed a cell big enough to show many AIS messages in a single trajectory that fit in a cell (compare Figure 9.33). The first variant, presented in Figure 9.44, shows the border points. The green dots are marked as the ones between which time is measured. The blue dots are other AIS messages; they occur if a vessel sent more messages within a segment.

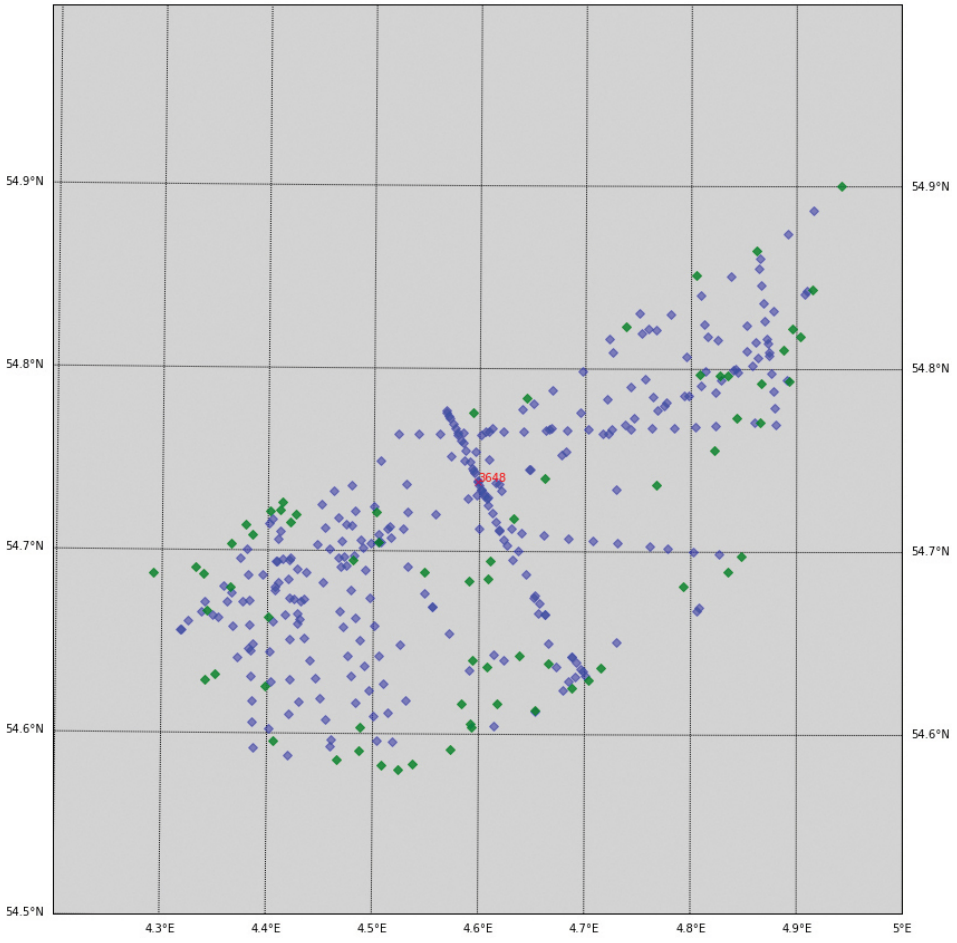


Figure 9.44. Waypoint cell with marked border points

Source: Own work.

The second variant emphasizes the points that were the closest to the waypoint in the given trajectory—see Figure 9.45. We can observe here a movement of green points towards the centre, which makes them better representatives of the waypoint. We need to remember that although the edges are created from the waypoints, calculation of the time necessary to get from one waypoint to another relies on the time deltas between the real points (the green ones). The closer are the points to the waypoint, the smaller is the error of the time calculation.

Although this approach seems to indicate data quality issues, in fact this is the only way to cope with time calculation. We just do not have direct connections between waypoints. We can speculate for example about an average speed of a vessel in an area. Nevertheless, if we take the average of 'green points' then the average time delta should be close to the real value.

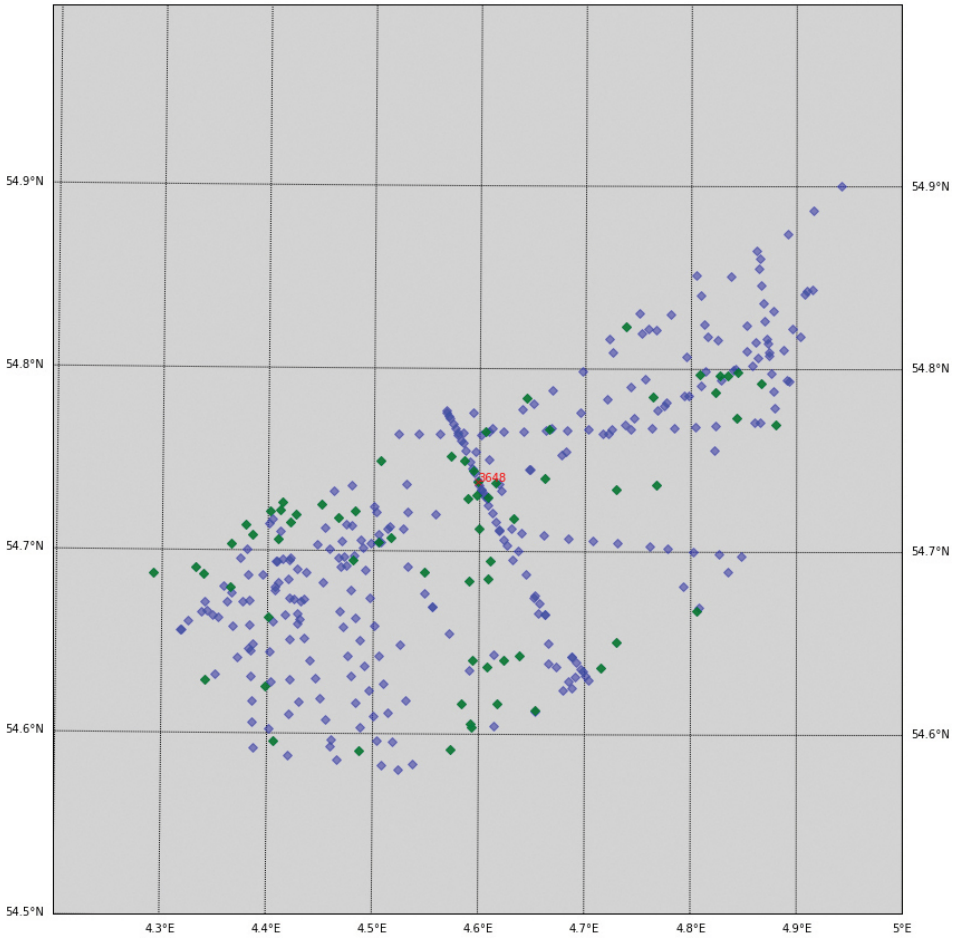


Figure 9.45. Waypoint cell with marked minimum distance

Source: Own work.

Analysis of the flow balance in waypoints. The last analysis was inspired by the empty central area, visible on all visualization (in the middle of the German Bight with a weak AIS coverage). We identified some dead ends in Figure 9.42. The vessels do not disappear and do not turn back. Therefore, we wanted to check if we have 'sources' or 'sinks' of vessels.

By analogy to the first Kirchhoff law,[14] we can define a law for vessels: the number of vessels flowing into a waypoint should equal the number of vessels that leave that waypoint. Any significant difference would mean that there is something

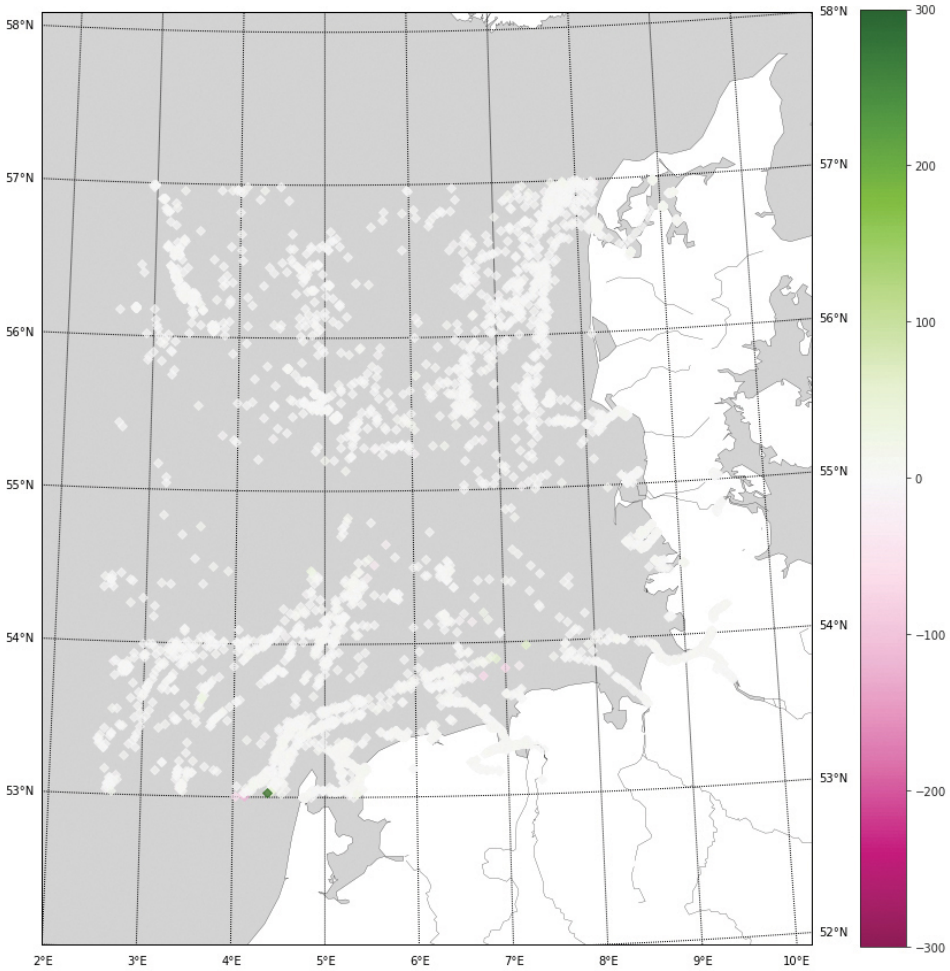


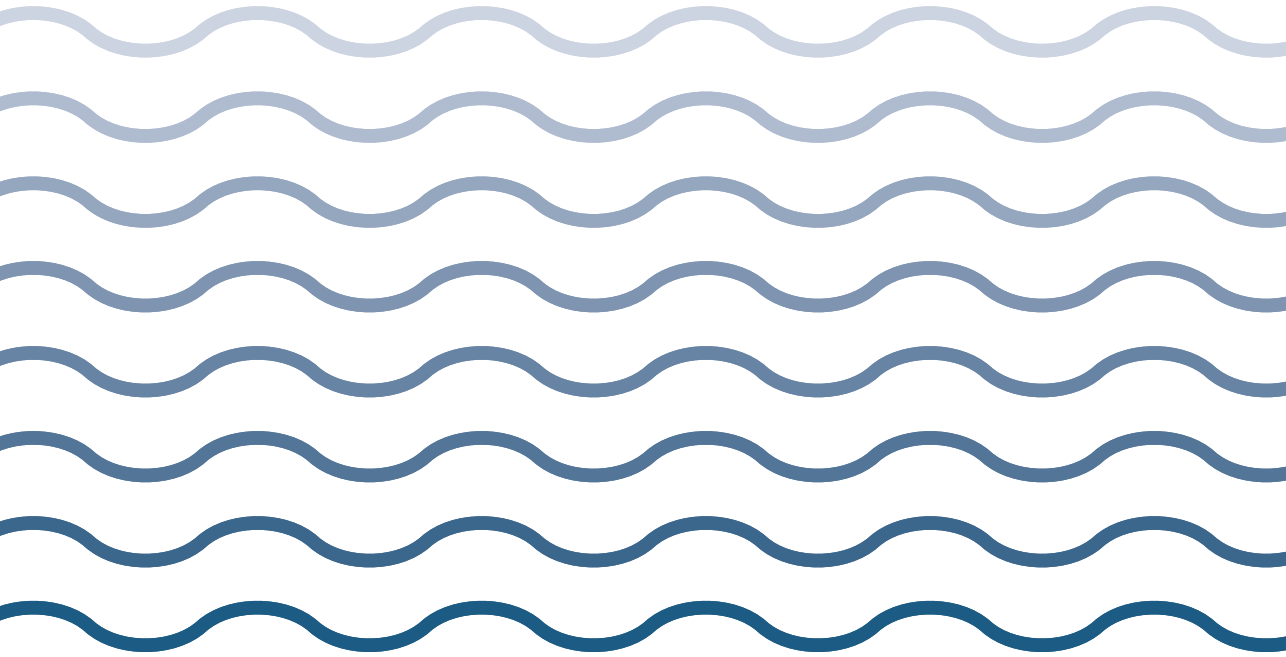
Figure 9.46. Flow balance of waypoints in the German Bight

Source: Own work.

14. https://isaacphysics.org/concepts/cp_kirchhoffs_laws

wrong with our method, i.e., vessels are not counted properly. As can be seen in Figure 9.46, such a phenomenon was not observed. The majority of the waypoints are well balanced. There are just a few imbalanced points in the south of the German Bight, but they result from the traffic separation rules. A green dot means that ships appear (there is flow-in and flow-out) and a pink one—disappear (there is flow-in but no flow-out).

Chapter 10



10. SUMMARY

Nowadays, technological developments which spur the emergence of new technologies, such as sensors and satellite and terrestrial systems, generate huge amounts of data that need to be efficiently and effectively retrieved, stored and, above all, processed and analysed to extract relevant and valuable information. This process is happening also in the maritime domain, which requires the capabilities to track ships and monitor what is happening on the seas in order to support maritime actors in decision making and provide them with a real time assessment of the situation. It holds especially true if we consider the growing seaborne trade, the expanding usage of maritime areas and the rising number of maritime threats and anomalous behaviours that are observed on the sea. Therefore, this book provides a comprehensive approach to dealing with the maritime data that encompasses identification and selection of appropriate data sources, data retrieval and fusion, quality enhancement, and particularly data analysis based on state-of-the-art data science methods.

First, the book presented a theoretical background to the approaches and methods developed by the authors in the course of many years of research. This theoretical elaboration introduced the problem of assessing the reliability and risk for maritime transport services and showed why it is crucial for various entities from the maritime domain to effectively assure a high quality of these services. It also presented the role of information in the process of reliability and risk assessment. Moreover, it provided an overview of various approaches and methods that have been developed by various researchers so far as well as some systems that are currently used in the maritime domain. The overview encompasses the areas of risk assessment, detection of maritime threats and anomalies, ship routes prediction, and maritime traffic analysis. Then, various data types and data sources available and used in the maritime domain were discussed with a special focus on data quality and appropriate selection of data sources to be used in the research. This theoretical analysis allowed us to explore the existing research gap in that area and created a foundation for the concepts of the original methods presented in this book.

Second, the book proposed novel approaches and methods for maritime data retrieval, fusion, enhancement, and analysis that try to address the identified gap along with suggestions how they might be applied in different maritime scenarios. These approaches included above all:

- A framework for the selection of open data sources which provide maritime-related data that can be fused with maritime data coming from other types of sources (e.g. sensors). The framework deals with internet sources and focuses mainly on the quality of data sources.
- Methods for retrieval of maritime data from various sources and its fusion.
- Methods for detection of static, dynamic, and loitering-related anomalies.
- An approach to generate maritime traffic networks based on historical AIS data, consisting of methods for waypoints generation, spatial data partitioning and traffic patterns exploration.
- A method for reliability and risk assessment of the maritime transport service.
- A method for punctuality prediction of ships.

Third, it provided a verification of the proposed methods using real maritime data. To this end, real data from various sources was used, such as worldwide satellite and terrestrial AIS data covering a period of a few years, weather data from the Copernicus platform, and data from open internet sources providing information on ships and their characteristics, detentions and inspections of ships, ship classifications, risk indexes, ship accidents and reported piracy attacks, as well as GIS data. All of those created a huge set of real maritime data. As a result, the research presented in this book was very data intensive.

In the course of the research, the problem of appropriate data quality emerged on many occasions. It concerned especially AIS data, both its static and dynamic attributes. The conducted assessment of AIS data quality revealed that problems can be found for each analysed attribute, which in turn negatively influences the quality of the analyses conducted based on AIS data. Therefore, along with the development of the analytics methods presented in the book, various other approaches for the improvement of maritime data quality had to be provided too.

Finally, due to the growing amount of maritime data that is available, the book tried to show the advantages stemming from the application of big data technologies for processing maritime data. Since huge amounts of AIS data were analysed in the presented research, appropriate infrastructure had to be used. To this end, either services and resources offered by the Microsoft Azure platform (thanks to the research grant in the program Microsoft Azure for Research) or the internal infrastructure of the research projects the authors had previously participated in were used. The book presents application of the big data approach to the problem of maritime anomaly detection and the generation of maritime traffic networks based on AIS data fused with data retrieved from open internet sources. As an example, we conducted, inter alia, a large-scale spatial analysis of the behaviour of tankers in 2015 and compared our big data approach with a traditional SQL-based solution. Not only novel concepts were introduced, but they were also implemented and tested with a parallel and distributed computational environment on Apache Spark. The evaluation of the developed methods showed

that the proposed implementation is scalable and can work with real-world AIS data streams. Thus, the results of the experiments provided preliminary evidence that incorporation of big data techniques and the Lambda architecture in AIS data processing increases the speed and efficiency of analyses, and as such should be the preferred solution used in the maritime domain.

We believe that the results presented in this book may be significant for various stakeholders and the proposed methods might be used in many different business scenarios. Due to the growing importance of maritime trade and the rising ship traffic on the local and international waters, a demand for the tools and methods developed here seems obvious. Such solutions may support compiling a maritime picture by integrating various sources of information and provide a near real time support in detection of maritime risks and threats. The methods can be helpful in the process of planning a ship's voyage when information about its ETA has to be provided to plan other activities in advance. It can also be used when a transport service is already under realization in order to track punctuality and update the ETA. Moreover, information about potential hazards may be used to plan and monitor a ship's voyage from the point of view of potential maritime threats. Last but not least, ships that will probably be delayed may be quickly identified.

Thereby, the developed methods have a great potential to be exploited in the real environment by various maritime stakeholders, such as European agencies (European Maritime Safety Agency), authorities and entities interested in monitoring maritime traffic and ensuring the security and safety of maritime transport as well as logistic companies, senders and recipients of goods. Moreover, they could be incorporated into the existing maritime and logistic systems to monitor fleets and maritime traffic as well as in intelligent navigational systems to support users in decision making.

While conducting the research presented in the book, possible directions for future research in this area were also identified. The first direction concerns further development of the methods for maritime traffic analysis, risk and reliability assessment. It may take into account new variables that can influence the reliability and punctuality of a transport service. Here, we see a great potential in inclusion of historical data and forecasts about the weather and the sea from the Copernicus data source and fusing them with AIS data. Thanks to this, a more detailed analysis of different ship routes under different weather conditions could be conducted in order to detect some unobvious interactions, and thus better predict ship routes for the forecasted weather. Moreover, an option for automatic re-planning (update) of a proposed route in situations of difficult weather to avoid potential danger could be developed.

With regard to the methods for maritime risk assessment, future work may focus on further improvement of the developed classifiers to increase the accuracy of the method. In this area, application of a larger training set may be foreseen, to see how it would influence the probability distribution of BN and, as a result, the

accuracy of predictions. Also, an analysis of the relationships between different variables (e.g. ship characteristics, attributes of their operational environments) and the attributes of the reliability of a transport service seem to be a good plan for future research.

In relation to the application of big data technologies, further evaluation of these technologies in the process of analysis of AIS and other maritime-related data might be considered in future studies, especially in terms of meticulous measuring of analysis time using different frameworks and storage formats. Future studies might ascertain the veracity of the yielded results in the light of different settings, investigate to what extent they are bound to the big data architecture, and meet the challenge of testing different solutions in this paradigm (e.g. Apache Storm or Cassandra).

APPENDIX A. EVALUATION OF THE MRRAM METHOD—RESULTS

The Appendix presents the results of analysis performed with evaluation of the MRRAM method, described in Section 7.3.

A1. Statistics of accidents for ship types and classification societies

Table A1. Statistic of accidents for ship types

Ship type	Number of accidents	Proportion (%)	Ship type	Number of accidents	Proportion (%)
GENERAL CARGO	66	14.38	Ro-Ro Cargo Ship	2	0.44
BULK CARRIER	52	11.33	Stern Trawler	2	0.44
CONTAINER SHIP	33	7.19	Tug	2	0.44
RO-RO/PASSENGER SHIP	29	6.32	WOOD CHIPS CARRIER	2	0.44
TRAWLER	21	4.58	Bulk Dry / Oil Carrier	1	0.22
Fish Catching Vessel	20	4.36	CARGO	1	0.22
OIL/CHEMICAL TANKER	19	4.14	Cargo ship	1	0.22
Passenger Ship	18	3.92	CARGO/PASSEN-GER SHIP	1	0.22
FISHING VESSEL	16	3.49	Chemical Tanker	1	0.22
General Cargo Ship	16	3.49	Crude Oil Tanker	1	0.22
OIL PRODUCTS TANKER	11	2.40	DECK CARGO SHIP	1	0.22
TUG	11	2.40	Domestic passenger boat	1	0.22
CRUDE OIL TANKER	10	2.18	Domestic Passenger Ship	1	0.22
PASSENGERS SHIP	9	1.96	DRILL SHIP	1	0.22
Towing / Pushing Tug	9	1.96	DRILLING JACK UP	1	0.22
VEHICLES CARRIER	8	1.74	DRILLING RIG	1	0.22

Ship type	Number of accidents	Proportion (%)	Ship type	Number of accidents	Proportion (%)
Other Ships Structures	7	1.53	FACTORY TRAWLER	1	0.22
REEFER	7	1.53	Fully decked Malahide Workboat	1	0.22
RO-RO CARGO	7	1.53	GRAB DREDGER	1	0.22
CHEMICAL TANKER	6	1.31	HEAVY LOAD CARRIER	1	0.22
CEMENT CARRIER	5	1.09	Liquefied Gas Tanker	1	0.22
Container Ship	4	0.87	LNG TANKER	1	0.22
OFFSHORE SUPPLY SHIP	4	0.87	LPG TANKER	1	0.22
ANCHOR HANDLING VESSEL	3	0.65	Non-Propelled Ships	1	0.22
Bulk Dry (general, ore) Carrier	3	0.65	Not Specified	1	0.22
CREW BOAT	3	0.65	O shore Supply Ship	1	0.22
Fish Factory Ship / Fish Carrier	3	0.65	PALLET CARRIER	1	0.22
Other Activities Ships	3	0.65	Passenger / General Cargo Ship	1	0.22
UTILITY VESSEL	3	0.65	Passenger / Ro-Ro Cargo Ship	1	0.22
ASPHALT/BITUMEN TANKER	2	0.44	PATROL VESSEL	1	0.22
Container Ship (Fully Cellular)	2	0.44	PIPELAY CRANE VESSEL	1	0.22
LIVESTOCK CARRIER	2	0.44	Platform Supply Ship	1	0.22
Oil Tanker	2	0.44	POLLUTION CONTROL VESSEL	1	0.22
NA	2	0.44	Research Ship	1	0.22
RESEARCH/SURVEY VESSEL	2	0.44	Trawler	1	0.22

Source: Own work.

Table A2. Statistic of accidents for classification societies

Classification society	Number of accidents	Proportion (%)
Det Norske Veritas	42	13.12
Germanischer Lloyd	40	12.5
Nippon Kaiji Kyokai	37	11.56
American Bureau of Shipping	32	10.00
Lloyds Register	30	9.38
Bureau Veritas	28	8.75
Biro Klasiękasi Indonesia	16	5.00
Korean Register of Shipping	16	5.00
Registro Italiano Navale	11	3.44
China Classięcation Society	10	3.12
Russian Maritime Register of Shipping	9	2.81
Bureau Veritas (BV)	7	2.19
Vietnam Register of Shipping	6	1.88
Turk Loydu	3	0.94
ASIA Classięcation Society (ACS)	2	0.62
Hellenic Register of Shipping	2	0.62
Indian Register of Shipping	2	0.62
International Naval Surveys Bureau	2	0.62
International Register of Shipping	2	0.62
Nippon Kaiji Kyokai (NKK)	2	0.62
No class	2	0.62
Union Marine Classięcation Society	2	0.62
American Bureau of Shipping (ABS)	1	0.31
Det Norske Veritas (DNV)	1	0.31
DNV GL AS (DNVGL)	1	0.31
Dromon Bureau Of Shipping	1	0.31
Flag Administration	1	0.31
Intermaritimecertięcationservices .S.A	1	0.31
InternationalShipClassięcation	1	0.31
Lloyd's Register (LR)	1	0.31
Mongolia Ship Registry	1	0.31
OTHER (PANAMA SHIPPING REGSTRAR INC.)	1	0.31
Overseas Marine Certięcation Services	1	0.31
Panama Maritime Documentation Services	1	0.31
Phoenix Register of Shipping	1	0.31
Polish Register of Shipping	1	0.31
Polish Register of Shipping (PRS)	1	0.31
SING LLOYD	1	0.31
Universal Maritime Bureau Ltd	1	0.31

Source: Own work.

A2. Bayesian Network parameters for the risk classifiers

The listings A1, A2, A3 present the estimations of a posteriori probabilities for all variables included in a given BN.

Listing A1. A posteriori conditional probabilities for the factor of the ship-related classifier Source: Printout from R based on own work

```
Bayesian network parameters

Parameters of node age (multinomial distribution)

Conditional probability table:

delay
age                    0           1
        middle age 0.779874214 0.728571429
        new        0.213836478 0.242857143
        old        0.006289308 0.028571429

Parameters of node delay (multinomial distribution)

Conditional probability table:
0         1
0.6943231 0.3056769

Parameters of node flag (multinomial distribution)

Conditional probability table:

delay
flag                   0           1
        black          0.00000000 0.01428571
        grey           0.07547170 0.02857143
        white          0.92452830 0.95714286

Parameters of node size (multinomial distribution)

Conditional probability table:

delay
size                  0         1
        large      0.2893082 0.3428571
        medium     0.5723270 0.5428571
        very large 0.1383648 0.1142857

Parameters of node soc (multinomial distribution)
```

```
Conditional probability table:

delay
soc                  0         1
        reliable   0.5471698 0.5571429
        unreliable 0.4528302 0.4428571

Parameters of node status (multinomial distribution)

Conditional probability table:

delay
status               0          1
        delivered  0.93710692 0.90000000
        reassigned 0.00000000 0.01428571
        reinstated 0.02515723 0.05714286
        withdrawn  0.03773585 0.02857143

Parameters of node type (multinomial distribution)

Conditional probability table:

delay
type                 0          1
        dangerous 0.03144654 0.02857143
        safe      0.96855346 0.97142857
```

Listing A2. A posteriori conditional probabilities for the factor of the voyage-related classifier Source: Printout from R based on own work

```
Bayesian network parameters

Parameters of node cargotype (multinomial distribution)

Conditional probability table:

rel
cargotype          0                    1
        0                0.6428571                0.6540881
        1                0.3571429                0.3459119

Parameters of node congestion (multinomial distribution)

Conditional probability table:

rel
congestion          0                   1
```

```
        0                0.8714286                0.8238994
        1                0.1285714                0.1761006

Parameters of node delay (multinomial distribution)

Conditional probability table:

rel
delay         0                         1
        0                0.3142857                0.4779874
        1                0.6857143                0.5220126

Parameters of node hazard (multinomial distribution)

Conditional probability table:

rel
hazard          0                         1
        0                0.5571429                0.4591195
        1                0.4428571                0.5408805

Parameters of node rel (multinomial distribution)

Conditional probability table:
0         1
0.3056769 0.6943231

Parameters of node traveltime (multinomial distribution)

Conditional probability table:

rel
traveltime          0                          1
        0                0.1714286                0.3396226
        1                0.8285714                0.6603774
```

Listing A3. A posteriori conditional probabilities for the factor of the history-related classifier Source: Printout from R based on own work

```
Bayesian network parameters

Parameters of node Accidents (multinomial distribution)

Conditional probability table:

delay
Accidents           0                        1
        0                0.94502618               0.92857143
```

```
        1                0.05497382              0.07142857

Parameters of node BlackPorts (multinomial distribution)

Conditional probability table:

delay
BlackPorts           0                   1
        0               0.955497382      0.994505495
        1               0.044502618      0.005494505

Parameters of node CargoLoss (multinomial distribution)

Conditional probability table:

delay
CargoLoss          0                    1
        0               0.97643979              0.97252747
        1               0.02356021              0.02747253

Parameters of node Casualties (multinomial distribution)

Conditional probability table:

delay
Casualties          0                   1
        0               0.98691099              0.97252747
        1               0.01308901              0.02747253

Parameters of node delay (multinomial distribution)

Conditional probability table:
0        1
0.677305 0.322695

Parameters of node Detentions (multinomial distribution)

Conditional probability table:

delay
Detentions         0                    1
        0               0.8821990               0.8406593
        1               0.1178010               0.1593407

Parameters of node Identification (multinomial distribution)

Conditional probability table:

delay
```

```
Identification        0                   1
        0                0.986910995      0.994505495
        1                0.013089005      0.005494505

Parameters of node Incomplete (multinomial distribution)

Conditional probability table:

delay
Incomplete          0                        1
        0                0.05497382                0.07142857
        1                0.94502618                0.92857143

Parameters of node Loitering (multinomial distribution)

Conditional probability table:

delay
Loitering         0                       1
        0                0.8193717                0.7527473
        1                0.1806283                0.2472527

Parameters of node PastCS (multinomial distribution)

Conditional probability table:

delay
PastCS                  0           1
        delivered       0.891361257 0.881868132
        reassigned      0.001308901 0.024725275
        reinstated      0.032722513 0.068681319
        withdrawn       0.074607330 0.024725275

Parameters of node PastDelays (multinomial distribution)

Conditional probability table:

delay
PastDelays        0                       1
        0                0.5052356                0.4230769
        1                0.4947644                0.5769231

Parameters of node PastStatus (multinomial distribution)

Conditional probability table:

delay
PastStatus          0         1
        reliable        0.7460733 0.7087912
```

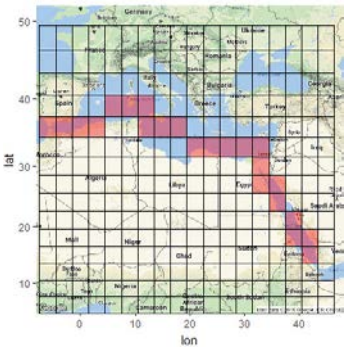
```
        unreliable       0.2539267 0.2912088

Parameters of node Pollution (multinomial distribution)

Conditional probability table:

delay
Pollution          0                         1
        0                0.98691099                0.97252747
        1                0.01308901                0.02747253

Parameters of node ProtectedAreas (multinomial distribution)

Conditional probability table:

delay
ProtectedAreas     0                   1
        0                0.6308901                 0.5329670
        1                0.3691099                 0.4670330

Parameters of node StaticChanges (multinomial distribution)

Conditional probability table:

delay
StaticChanges      0                     1
        0                0.4005236                 0.3571429
        1                0.5994764                 0.6428571
```

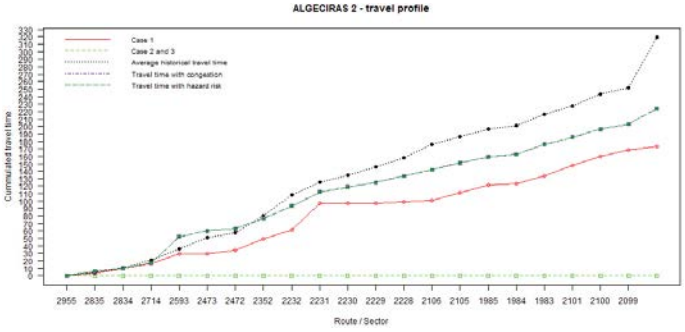
APPENDIX B. EVALUATION OF THE SPP METHOD—RESULTS

The Appendix presents the results of the routes prediction method (see Section 8.6). For each voyage, starting point, destination, sequence of sectors and visualization of the route is provided.

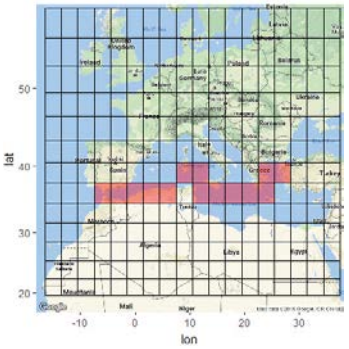
B1. Results of route prediction method

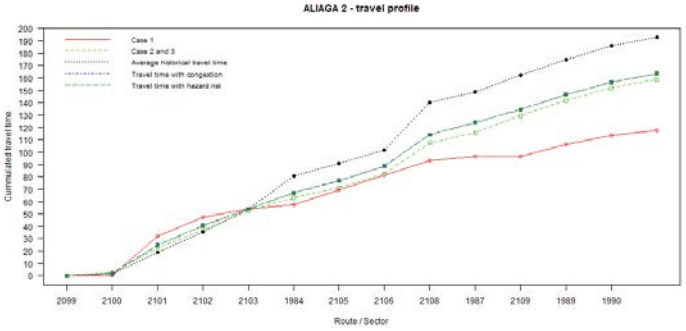
Destination: ALGECIRAS
Starting point: 16.15, 41.3
Predicted route: (23 sectors)
2955, 2835, 2834,
2714,2593,2473,
2472,2352,2232,
2231,2230,2229,
2228,2107,2106,
2105,1985,1984,
1983,2102,2101,
2100,2099



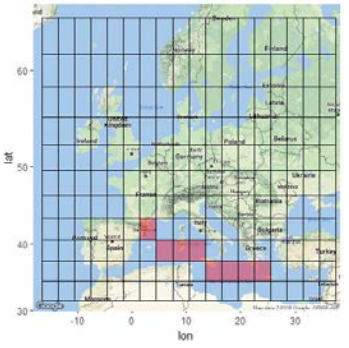


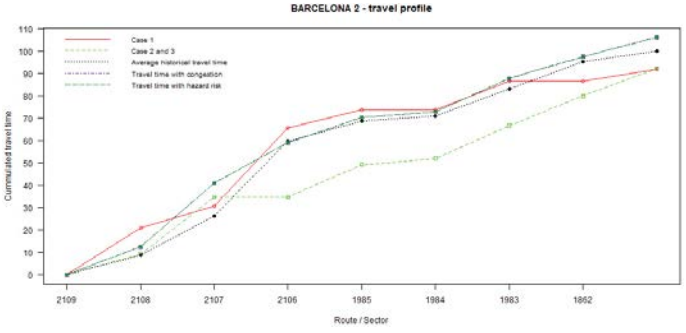
Destination: ALIAGA
Starting point: 36.15, -4.7
Predicted route: (13 sectors)
2099, 2100, 2101,
2102, 2103, 1984,
2105, 2106, 2108,
1987, 2109, 1989,
1990



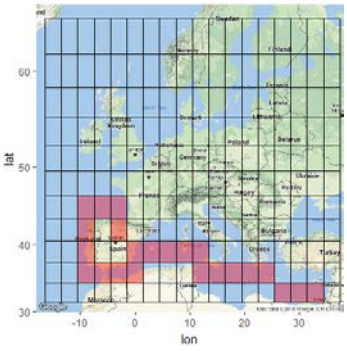


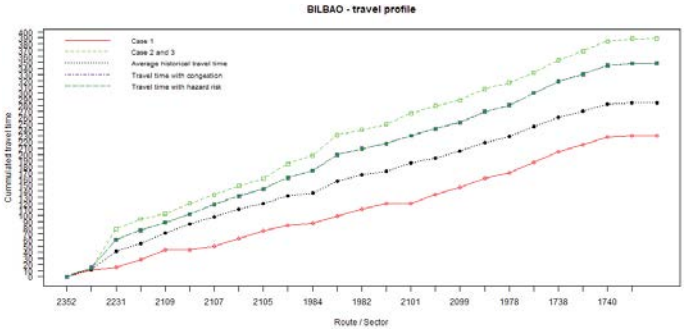
Destination: BARCELONA
Starting point: 35.5, 25.5
Predicted route: (8 sectors)
2109, 2108, 2107,
2106, 1985, 1984,
1983, 1862



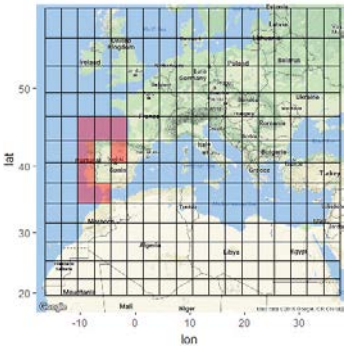


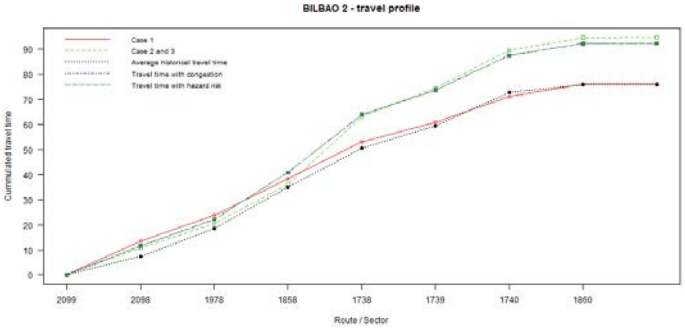
Destination: BILBAO 1
Starting point: 30.00, 32.5
Predicted route: (23 sectors)
2352, 2232, 2231,
2230, 2109, 2108,
2107, 2106, 2105,
1985, 1984, 1983,
1982, 1981, 2101,
2100, 2099, 2098,
1978, 1858, 1738,
1739, 1740, 1860



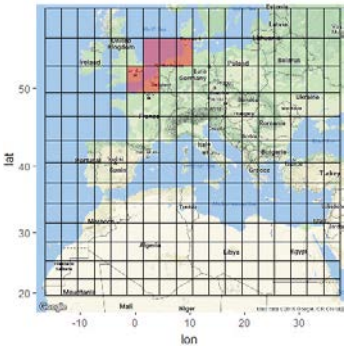


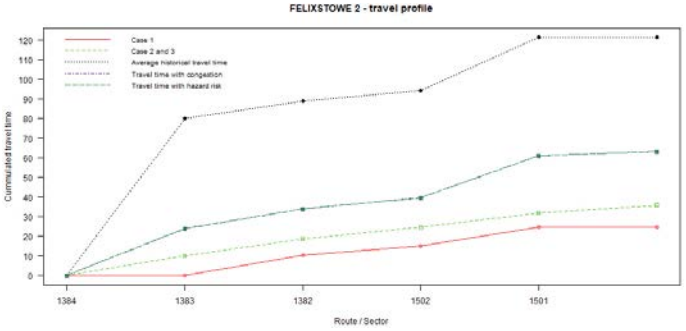
Destination: BILBAO 2
Starting point: 35.9, -5.6
Predicted route: (8 sectors)
2099, 2098, 1978,
1858, 1738, 1739,
1740, 1860



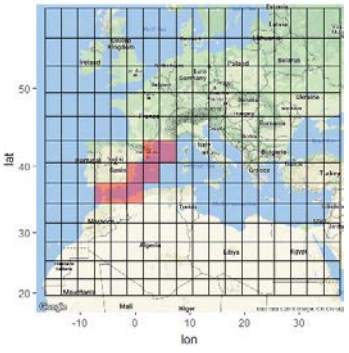


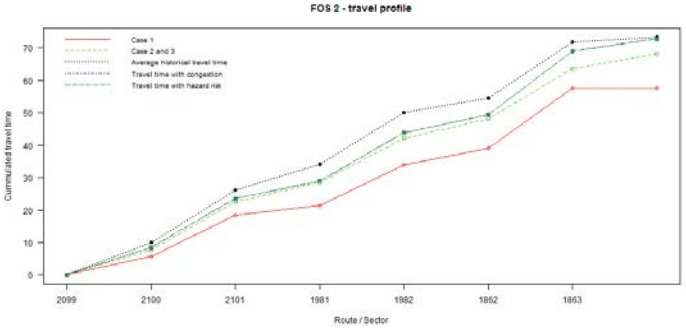
Destination: FELIXSTOWE
Starting point: 54.0, 7.75
Predicted route: (5 sectors)
1384, 1383, 1382,
1502, 1501



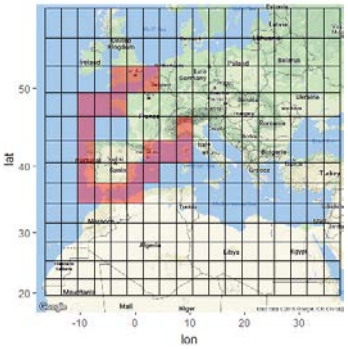


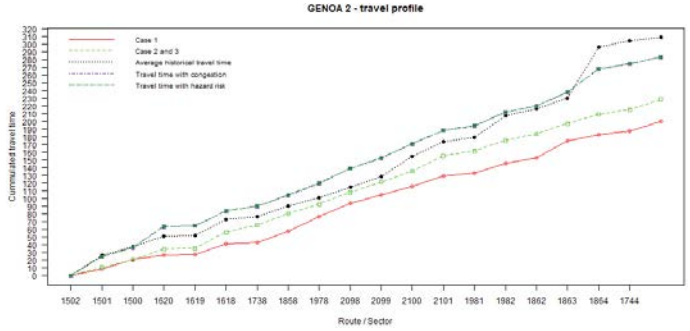
Destination: FOS
Starting point: 35.92, -6.25
Predicted route: (7 sectors)
2099, 2100, 2101,
1981, 1982, 1862,
1863



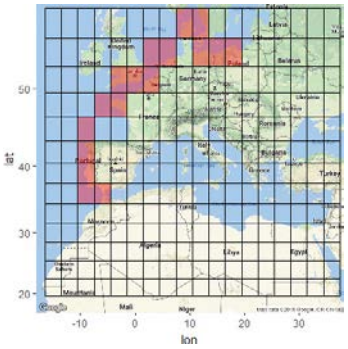


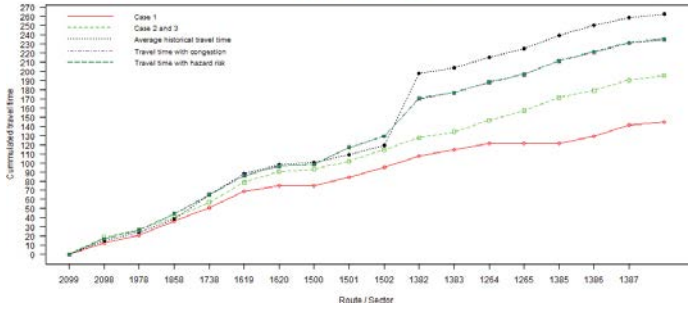
Destination: GENOA
Starting point: 52.4. 4.0
Predicted route: (19 sectors)
1502, 1501, 1500,
1620, 1619, 1618,
1738, 1858, 1978,
2098, 2099, 2100,
2101, 1981, 1982,
1862, 1863, 1864,
1744



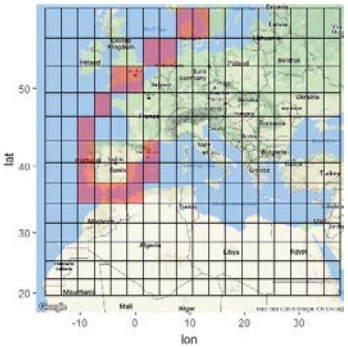


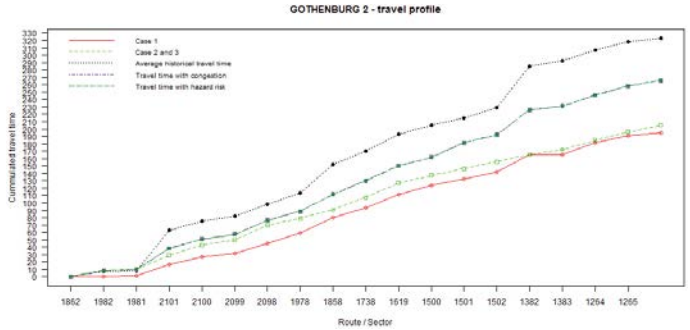
Destination: GDANSK
Starting point: 36.00, -4.55
Predicted route: (17 sectors)
2099, 2098, 1978,
1858, 1738, 1619,
1620, 1500, 1501,
1502, 1382, 1383,
1264, 1265, 1385,
1386, 1387



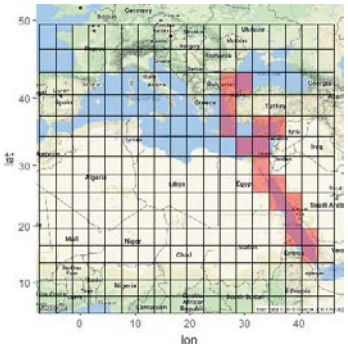


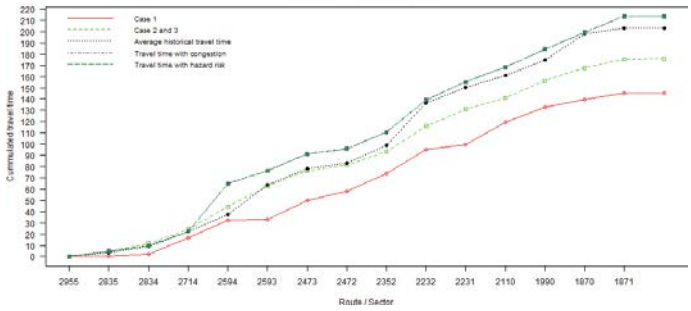
Destination: GOTHENBURG
Starting point: 40.5, 1.75
Predicted route: (18 sectors)
1862, 1982, 1981,
2101, 2100, 2099,
2098, 1978, 1858,
1738, 1619, 1500,
1501, 1502, 1382,
1383, 1264, 1265



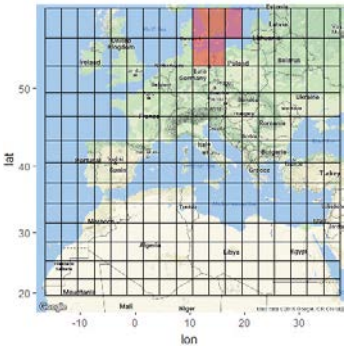


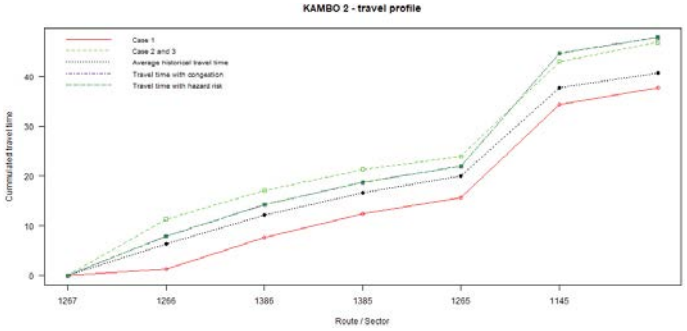
Destination: ISTANBUL
Starting point: 16.15, 41.3
Predicted route: (15 sectors)
2955, 2835, 2834,
2714, 2594, 2593,
2473, 2472, 2352,
2232, 2231, 2110,
1990, 1870, 1871



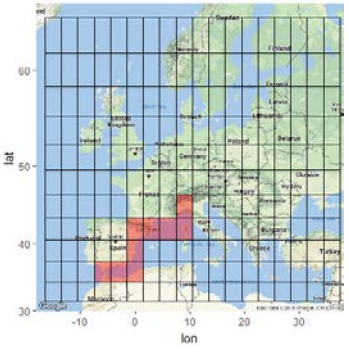


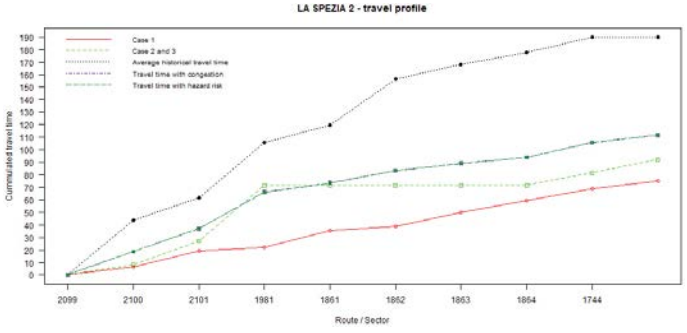
Destination: KAMBO
Starting point: 55.9, 17.3
Predicted route: (6 sectors)
1267, 1266, 1386,
1385, 1265, 1145



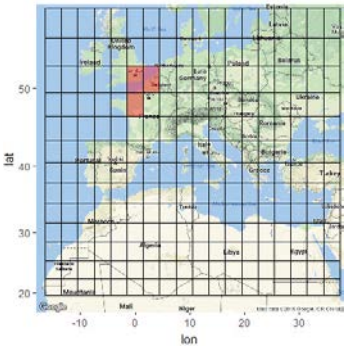


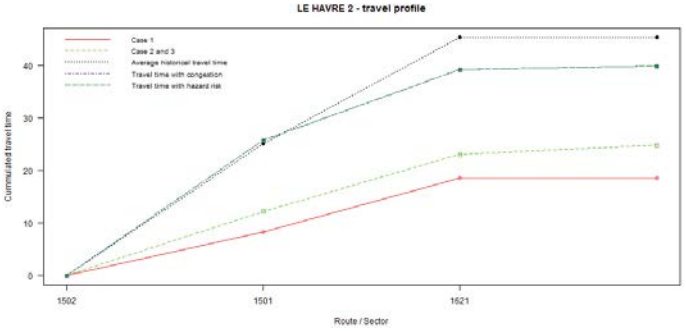
Destination: LA SPEZIA
Starting point: 36.1, -5.4
Predicted route: (9 sectors)
2099, 2100, 2101,
1981, 1861, 1862,
1863, 1864, 1744



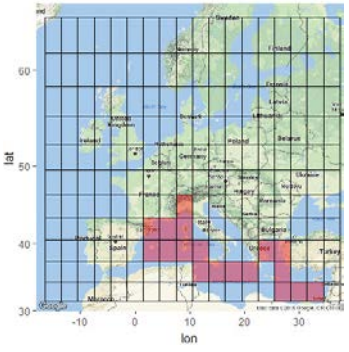


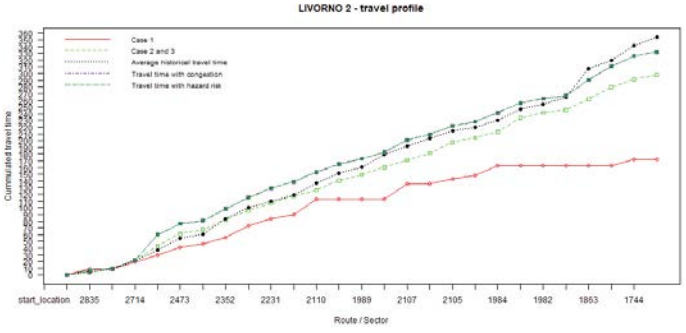
Destination: LE HAVRE
Starting point: 51.9, 3.7
Predicted route: (3 sectors)
1502, 1501, 1621



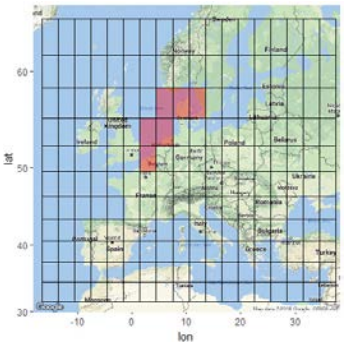


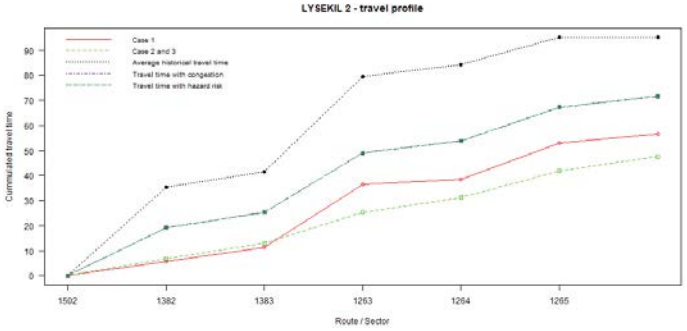
Destination: LIVORNO
Starting point: 16.15, 41.3
Predicted route: (26 sectors)
2955, 2835, 2834, 2714,
2593, 2473, 2472, 2352,
2232, 2231, 2230, 2110,
1990, 1989, 2108, 2107,
2106, 2105, 1985, 1984,
1983, 1982, 1862, 1863,
1864, 1744



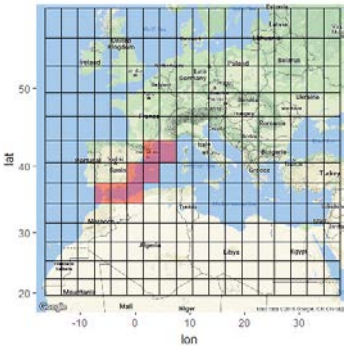


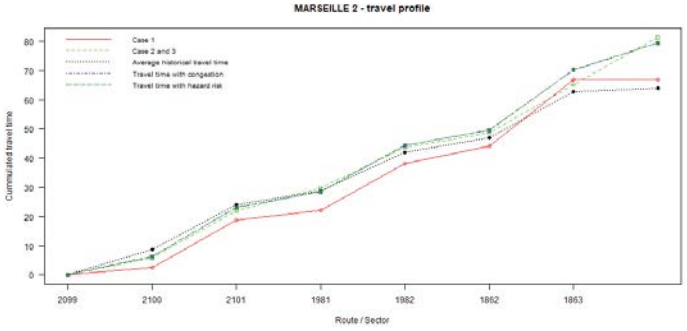
Destination: LYSEKIL
Starting point: 52.4, 3.2
Predicted route: (6 sectors)
1502, 1382, 1383,
1263, 1264, 1265



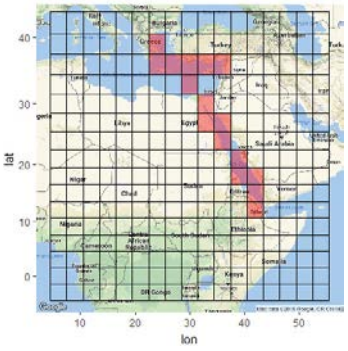


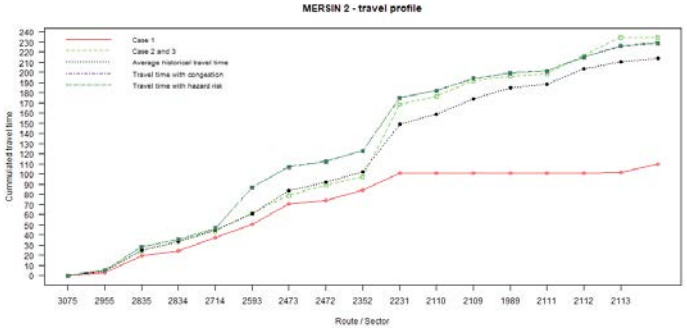
Destination: MARSEILLE
Starting point: 52.4, 3.2
Predicted route: (6 sectors)
2099, 2100, 2101,
1981, 1982, 1862,
1863



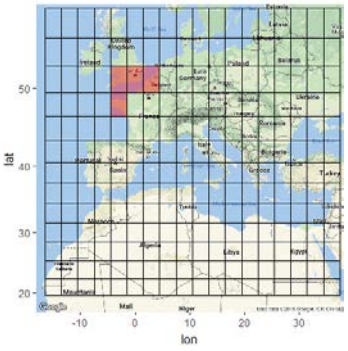


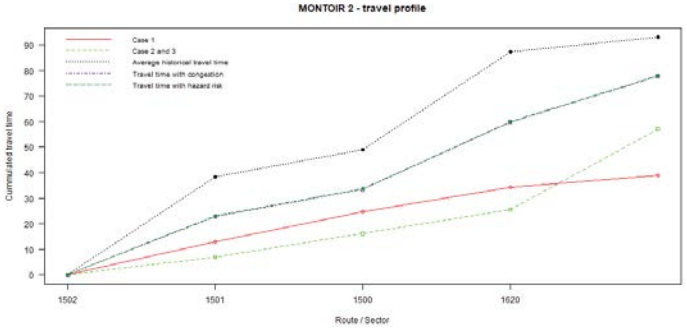
Destination: MERSIN
Starting point: 12.6, 43.1
Predicted route: (16 sectors)
3075, 2955, 2835,
2834, 2714, 2593,
2473, 2472, 2352,
2231, 2110, 2109,
1989, 2111, 2112,
2113



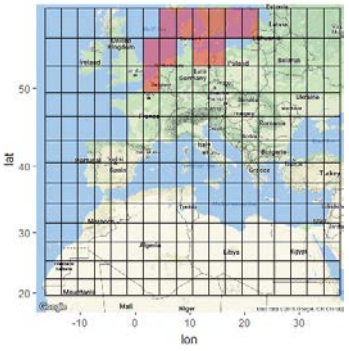


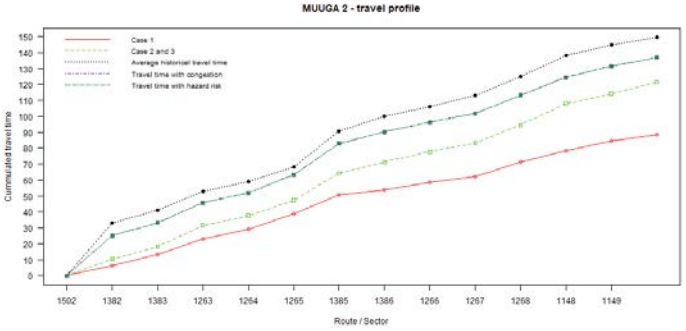
Destination: MONTOIR
Starting point: 52.37, 3.4
Predicted route: (4 sectors)
1502, 1501, 1500,
1620



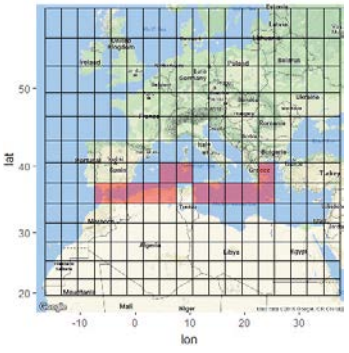


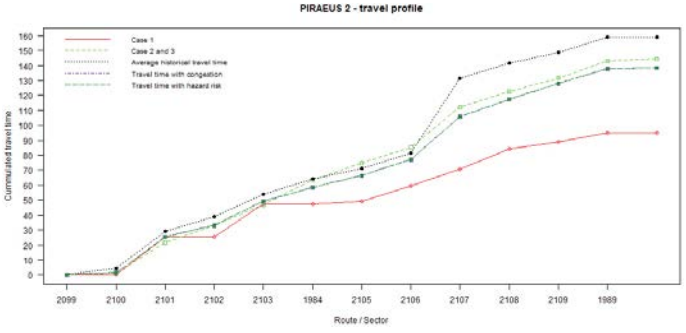
Destination: MUUGA
Starting point: 51.4, 2.0
Predicted route: (13 sectors)
1502, 1382, 1383,
1263, 1264, 1265,
1385, 1386, 1266,
1267, 1268, 1148,
1149



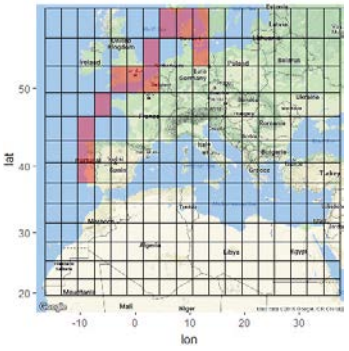


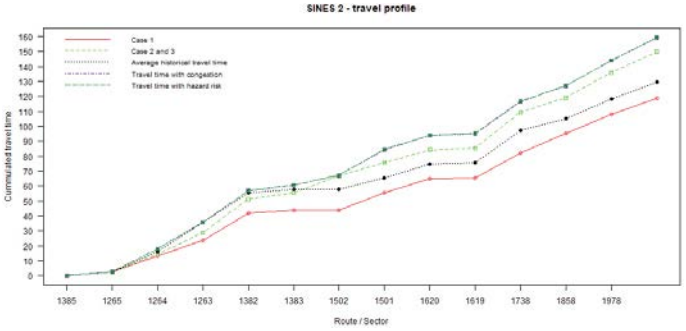
Destination: PIRAEUS
Starting point: 36.0, -4.7
Predicted route: (12 sectors)
2099, 2100, 2101,
2102, 2103, 1984,
2105, 2106, 2107,
2108, 2109, 1989



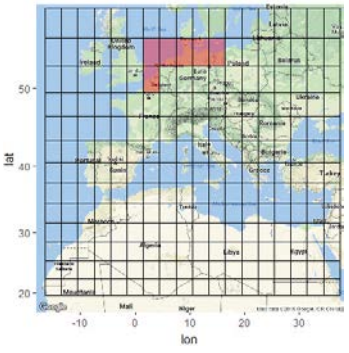


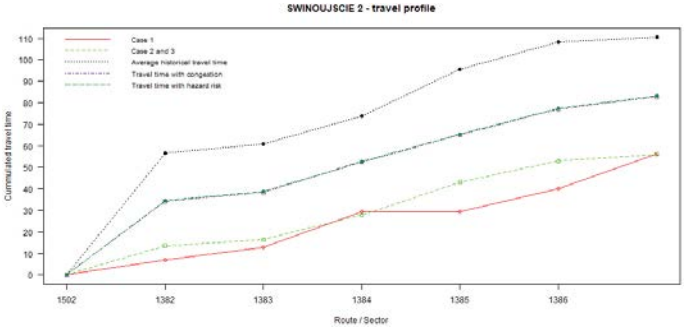
Destination: SINES
Starting point: 54.5, 10.5
Predicted route: (13 sectors)
1385, 1265, 1264,
1263, 1382, 1383,
1502, 1501, 1620,
1619, 1738, 1858,
1978



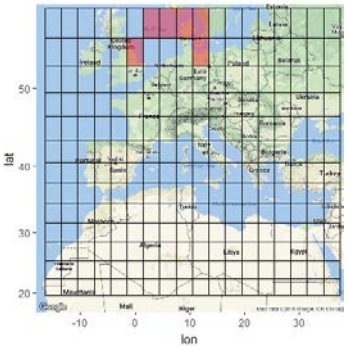


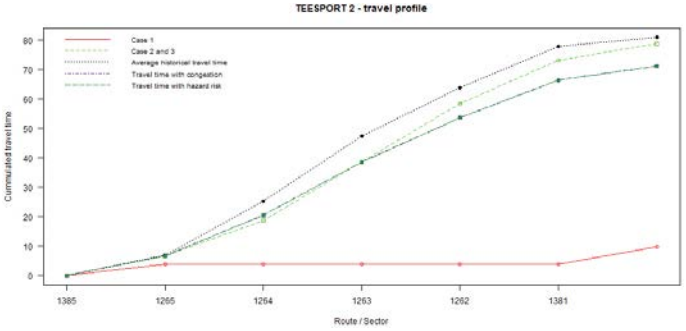
Destination: SWINOUJSCIE
Starting point: 51.5, 2.14
Predicted route: (6 sectors)
1502, 1382, 1383,
1384, 1385, 1386



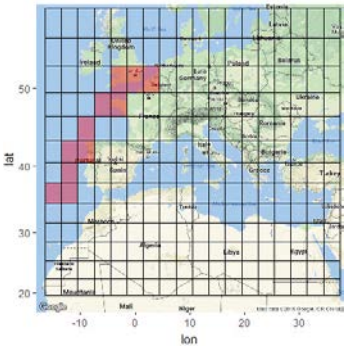


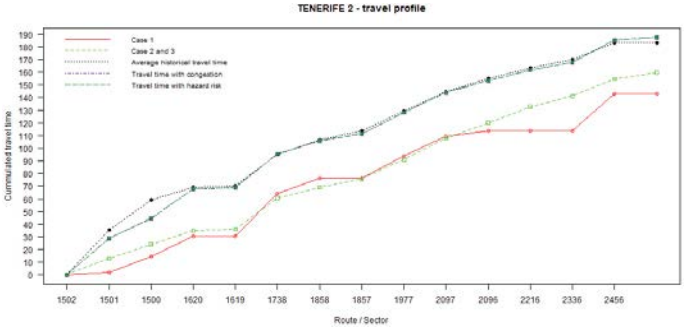
Destination: TEESPORT
Starting point: 54.59, 12.28
Predicted route: (6 sectors)
1385, 1265, 1264,
1263, 1262, 1381



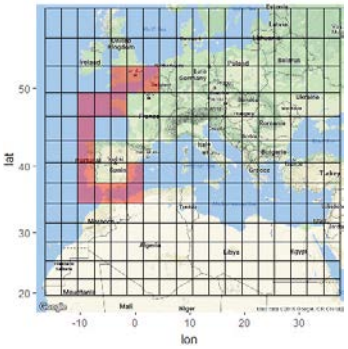


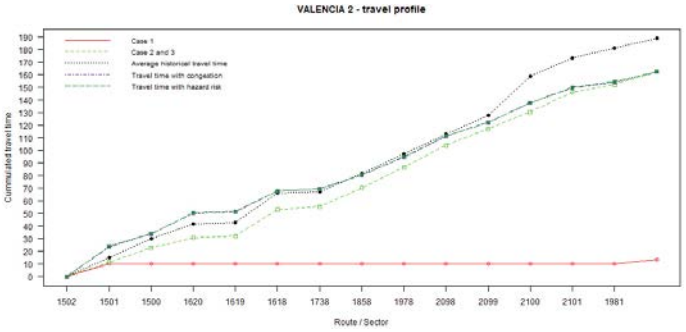
Destination: TENERIFE
Starting point: 51.5, 3.01
Predicted route: (14 sectors)
1502, 1501, 1500,
1620, 1619, 1738,
1858, 1857, 1977,
2097, 2096, 2216,
2336, 2456





Destination: VALENCIA
Starting point: 51.9, 4.0
Predicted route: (14 sectors)
1502, 1501, 1500,
1620, 1619, 1618,
1738, 1858, 1978,
2098, 2099, 2100,
2101, 1981





B2. Hazard index—results

Table B1. Accident, Piracy and Country risk values for selected maritime sectors (areas)

Sector	$A(S)_1$	$A(S)_2$	$A(S)_3$	$A(S)_4$	$A(S)_5$	$A(S)_6$	$A(S)_7$	$A(S)_8$	$A(S)_9$	$A(S)_{10}$	$A(S)_{11}$	$A(S)_{12}$	P	CR
1502	0.0833	0.1666	0.0833	0.2500	0	0.0833	0.1666	0	0	0.0833	0.0833	0	0	0.2297
2731	0.0833	0.0833	0.0833	0	0.3333	0.0833	0	0.0833	0.0833	0	0	0	0.0439	0.5798
3101	0	0.0833	0	0	0	0.0833	0.0833	0.0833	0.1666	0	0.2500	0	0.0140	0.6089
3195	0	0	0.0833	0	0	0	0.0833	0	0	0	0	0	0.0012	1
3575	0.2500	0.1666	0	0	0	0.3333	0.2500	0.2500	0.0833	0.0833	0.1666	0.1666	0.1747	0.3736
4317	0	0	0	0.0833	0	0	0.0833	0	0	0	0	0	0	0.7080

Legend: $A(S)_i$: Accident rate in sector S in month i; P: piracy rate; CR: country risk.

Source: Own work.

Table B2. Hazard index for selected maritime sectors (areas)

Sector	$H(S)_1$	$H(S)_2$	$H(S)_3$	$H(S)_4$	$H(S)_5$	$H(S)_6$	$H(S)_7$	$H(S)_8$	$H(S)_9$	$H(S)_{10}$	$H(S)_{11}$	$H(S)_{12}$
1502	0.070958	0.095958	0.070958	0.120958	0.045958	0.070958	0.095958	0.045958	0.045958	0.070958	0.070958	0.045958
2731	0.162924	0.162924	0.162924	0.137924	0.237924	0.162924	0.137924	0.162924	0.162924	0.137924	0.137924	0.137924
3101	0.128821	0.153821	0.128821	0.128821	0.128821	0.153821	0.153821	0.153821	0.178821	0.128821	0.203821	0.128821
3195	0.200639	0.200639	0.225639	0.200639	0.200639	0.200639	0.225639	0.200639	0.200639	0.200639	0.200639	0.200639
3575	0.237112	0.212112	0.162112	0.162112	0.162112	0.262112	0.237112	0.237112	0.187112	0.187112	0.212112	0.212112
4317	0.141766	0.141766	0.141766	0.166766	0.141766	0.141766	0.166766	0.141766	0.141766	0.141766	0.141766	0.141766

Legend: $H(S)_i$: hazard index for sector S and month i.

Source: Own work.

REFERENCES

ABP Marine Environmental Research Ltd. (2013). *Spatial trends in shipping activity*. Retrieved from http://webarchive.nationalarchives.gov.uk/20140108121958/http:/www.marinemanagement.org.uk/evidence/documents/1042.pdf

Abramowicz, W., Filipiak, D., Małyszko, J., Stróżyna, M., & Węcel, K. (2016). *Maritime domain awareness system supplied with external information-usecase of the SIMMO system.* (7th International Science and Technology Conference NATCON on "Naval Technologies for Defence and Security" (p. 1-20). Gdynia: Akademia Marynarki Wojennej.

Abramowicz-Gerigk, T., Burciu, Z., & Kamiński, P. (2013). Kryteria akceptowalności ryzyka w żegludze morskiej. *Prace Naukowe Politechniki Warszawskiej*, *96*.

ABS. (2020). Guidance notes on risk assessment applications for the marine and offshore industries. American Bureau of Shipping. Retrieved from https://ww2.eagle.org/content/dam/eagle/rules-and-guides/current/other/97_riskassessapplmarineandoffshoreoandg/risk-assessment-gn-may20.pdf

Alberga, C. N. (1967). String similarity and misspellings. *Communications of the ACM*, *10*(5), 302–313. https://doi.org/10.1145/363282.363326

Alonso, J., Ambur, O., Amutio, M. A., Azañón, O., Bennett, D., Flagg, R., ..., Sheridan, J. (2009*). Improving access to government through better use of the web.* World Wide Web Consortium. Retrieved from http://www.w3.org/TR/2009/NOTE-egov-improving-20090512/

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. *Proc. Spie*, *7305*, 73050G–73050G–9). https://doi.org/10.1117/12.818567

Angerman, W. S. (2004). *Coming full circle with Boyd's OODA loop ideas: An analysis of innovation diffusion and evolution.* DTIC Document. Retrieved from https://apps.dtic.mil/sti/citations/ADA425228

Arguedas, V. F., Pallotta, G., & Vespe, M. (2017). Maritime traffic networks: From historical positioning data to unsupervised maritime traffic monitoring. *IEEE Transactions on Intelligent Transportation Systems*, *19*(3), 722–732.

Arici, S. S., Akyuz, E., & Arslan, O. (2020). Application of fuzzy bow-tie risk analysis to maritime transportation: The case of ship collision during the STS operation. *Ocean Engineering*, *217*, 107960. https://doi.org/https://doi.org/10.1016/j.oceaneng.2020.107960

Asariotis, R., & Benamara, H. (2012). *Review of maritime transport, 2012.* New York, Geneva: United Nations. Retrieved from http://trid.trb.org/view.aspx?id=1238887

Aurenhammer, F., Klein, R., & Lee, D. T. (2013). *Voronoi diagrams and delaunay 302 B Evaluation of the SPP method—results triangulations.* https://doi.org/10.1142/8685

Auslander, B., Gupta, K. M., & Aha, D. W. (2012). *Maritime threat detection using plan recognition.* (IEEE conference on technologies for Homeland Security, pp. 249–254). https://doi.org/10.1109/THS.2012.6459857

Azariadis, P. (2017). On using density maps for the calculation of ship routes. *Evolving Systems*, *8*(2), 135–145.

Bakir, N. O. (2007). A brief analysis of threats and vulnerabilities in the maritime domain. In I. Linkov, R. J. Wenning, & G. A. Kiker (Eds.), *Managing critical infrastructure risks. NATO science for peace and security* (pp. 17–49). Springer.

Balduzzi, M., Wilhoit, K., & Pasta, A. (2014). A security evaluation of AIS. *Trend Micro*, 1–9.

Balmat, J. F., Lafont, F., Maifret, R., & Pessel, N. (2009). MAritime RISk Assessment (MARISA), a fuzzy approach to define an individual ship risk factor. *Ocean Engineering*, *36*(15–16), 1278–1286. https://doi.org/10.1016/j.oceaneng.2009.07.003

Başhan, V., Demirel, H., & Gul, M. (2020). An FMEA-based TOPSIS approach under single valued neutrosophic sets for maritime risk evaluation: The case of ship navigation safety. *Soft Computing*, *24*(24), 18749–18764. https://doi.org/10.1007/s00500-020-05108-y

Basseville, M., & Nikiforov, I. V. (1993). *Detection of abrupt changes: Theory and application*. Englewood Cliffs: Prentice Hall.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, *41*(3), 1–52. https://doi.org/10.1145/1541880.1541883

Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, *38*(8), 716–719.

Berle, Ø., Asbjørnslett, B. E., & Rice, J. B. (2011). Formal vulnerability assessment of a maritime transportation system. *Reliability Engineering and System Safety*, *96*(6), 696–705. https://doi.org/10.1016/j.ress.2010.12.011

Bijlsma, S. J. (2001). A computational method for the solution of optimal control problems in ship routing. *Navigation*, 48(3), 144–154.

Bijlsma, S. J. (2008).Minimal time route computation for ships with pre-specified voyage fuel consumption. *The Journal of Navigation*, *61*(4), 723–733.

Bijlsma, S. J. (2010). Optimal ship routing with ocean current included. *The Journal of Navigation*, *63*(3), 565.

Bilenko, M., & Mooney, R. J. (2003). *Adaptive duplicate detection using learnable string similarity measures.* (Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 39–48). New York: ACM. https://doi.org/10.1145/956750.956759

Blaich, M., Köhler, S., Reuter, J., & Hahn, A. (2015). Probabilistic collision avoidance for vessels. *IFAC-PapersOnLine*, *48*(16), 69–74.

Bomberger, N. A., Rhodes, B. J., Seibert, M., &Waxman, A. M. (2006). *Associative learning of vessel motion patterns for maritime situation awareness.* (9th International Conference on Information Fusion, pp. 1–8).

Bouejla, A., Chaze, X., Guarnieri, F., & Napoli, A. (2014). A Bayesian Network to manage risks of maritime piracy against offshore oil fields. *Safety Science*, *68*, 222–230.

Brax, C. (2011). *Anomaly detection in the surveillance domain.* (Unpublished doctoral dissertation). Örebro University, School of Science and Technology.

Brown, B. B. (1968). *Delphi process: A methodology used for the elicitation of opinions of experts.* Santa Monica: Rand Corp.

Cai, Y., Wen, Y., & Wu, L. (2014). Ship route design for avoiding heavy weather and sea conditions. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, *8*.

Calkoen, C., & Santbergen, P. (2016). *MetOcean services to the marine transport sector. Deliverable of Melodies project.* Retrieved from https://www.melodiesproject.eu

Calvert, S., Deakins, E., & Motte, R. (1991). A dynamic system for fuel optimization trans-ocean. *The Journal of Navigation*, *44*(2), 233–265.

Castillo, C. (2004). *Effective Web crawling* (Unpublished doctoral dissertation). University of Chile.

Cazzanti, L., & Pallotta, G. (2015). *Mining maritime vessel traffic: Promises, challenges, techniques.* Oceans Genova. https://doi.org/10.1109/OCEANS-Genova.2015.7271555

Cem Kuzu, A., Akyuz, E., & Arslan, O. (2019). Application of Fuzzy Fault Tree Analysis (FFTA) to maritime industry: A risk analysing of ship mooring operation. *Ocean Engineering*, *179*, 128–134. https://doi.org/10.1016/j.oceaneng.2019.03.029

Čepinskis, J., & Masteika, I. (2015). Impact of logistics processes on competitiveness of companies. *Management of Organizations: Systematic Research*, (55), 71–89.

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

Chang, K. C. C., He, B., Li, C., Patel, M., & Zhang, Z. (2004). Structured databases on the web: Observations and implications. *ACM SIGMOD Record*, *33*(3), 61–70.

Chang, Y. C., Tseng, R. S., Chen, G. Y., Chu, P. C., & Shen, Y. T. (2013). Ship routing utilizing strong ocean currents. *The Journal of Navigation*, *66*(6), 825–835.

Chatzikokolakis, K., Zissis, D., Vodas, M., Spiliopoulos, G., & Kontopoulos, I. (2019). *A distributed lightning fast maritime anomaly detection service.* https://doi.org/10.1109/OCEANSE.2019.8867269

Chaze, X., Bouejla, A., Napoli, A., Guarnieri, F., Eude, T., & Alhadef, B. (2012). The contribution of Bayesian Networks to manage risks of maritime piracy against oil offshore fields. In H. Yu, G. Yu, W. Hsu, Y.-S. Moon, R. Unland, & J. Yoo (Eds.), *Database systems for advanced applications* (vol. 7240, pp. 81–91). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-29023-7_9

Chen, C.-H., Khoo, L. P., Chong, Y. T., & Yin, X. F. (2014, May). Knowledge discovery using genetic algorithm for maritime situational awareness. *Expert System with Applications*, *41*(6), 2742–2753. https://doi.org/10.1016/j.eswa.2013.09.042

Chen, C. H., Xu, G., & Li, F. (2017). *Local AIS data analytics for efficient operation 304 B Evaluation of the SPP method—results management in Vessel Traffic Service.* (13th IEEE Conference on Automation Science and Engineering, pp. 1668–1672).

Chen, M. M. L., & Chesneau, L. S. (2008). *Heavy weather avoidance and route design: Concepts and applications of 500 MB charts.* Paradise Cay Publications.

Chen, Z., Guo, J., & Liu, Q. (2017). *DBSCAN algorithm clustering for massive AIS data based on the hadoop platform.* 2017 International Conference on Industrial Informatics--computing Technology, Intelligent Technology, Industrial Information Integration, pp. 25–28).

Cho, J., & Garcia-Molina, H. (2003). Effective page refresh policies for web crawlers. *ACM Transactions on Database Systems*, *28*(4), 390–426.

Choi, B. H., Pelinovsky, E., Lee, H. J., & Woo, S. B. (2005). Estimates of tsunami risk zones on the coasts adjacent to the East (Japan) Sea based on the Synthetic Catalogue. *Natural Hazards*, *36*(3), 355–381. https://doi.org/10.1007/s11069-005-1937-3

Christiansen, M., Fagerholt, K., Nygreen, B., & Ronen, D. (2013). Ship routing and scheduling in the new millennium. *European Journal of Operational Research*, *228*(3), 467–483.

Coffman, E. G., Liu, Z., & Weber, R. R. (1998). Optimal robot scheduling for Web search engines. *Journal of Scheduling*, *1*(1), 15–29. https://doi.org/10.1002/(SICI)1099-1425(199806)1:1<15::AID-JOS3>3.0.CO;2-K

Coleman, J., Kandah, F., & Huber, B. (2020). *Behavioral model anomaly detection in Automatic Identification Systems (AIS).* (10th Annual Computing and Communication Workshop and Conference (CCWC), pp. 481–487).

Commission of the European Communities. (2002). eEurope 2002: Quality criteria for health related websites. *Journal of Medical Internet Research*, *4*(3). https://doi.org/10.2196/jmir.4.3.e15

Copernicus Marine Environment Monitoring Service (CMEMS). (2019). Retrieved June 14, 2019 from http://marine.copernicus.eu/

Cross, R., & Ballesio, J. (2003). A quantitative risk assessment model for oil tankers. *Transactions of the Society of Naval Architects and Marine Engineers*, *111*, 608–623.

d'Afflisio, E., Braca, P., Millefiori, L. M., & Willett, P. (2018). *Maritime anomaly detection based on mean-reverting stochastic processes applied to a realworld scenario.* (21st International Conference on Information Fusion, p. 1171–1177). https://doi.org/10.23919/ICIF.2018.8455854

Danielis, R., Marcucci, E., & Rotaris, L. (2005). Logistics managers' stated preferences for freight service attributes. *Transportation Research Part E: Logistics and Transportation Review*, *41*(3), 201–215.

de Vries, G., & van Someren, M. (2013). Recognizing vessel movements from historical data. In P. van de Laar, J. Tretmans & M. Borth (Eds.), *Situation awareness with systems of systems* (pp. 105–118). New York: Springer.

De Wit, C. (1990). Proposal for low cost ocean weather routing. *The Journal of Navigation*, *43*(3), 428–439.

Diligenti, M., Coetzee, F., Lawrence, S., Giles, C. L., & Gori, M. (2000). *Focused crawling using context graphs.* (Proceedings of the 26th International Conference on Very Large Data Bases, pp. 527–534). San Francisco: Morgan Kaufmann Publishers Inc.

Dill, S., Kumar, R., McCurley, K. S., Rajagopalan, S., Sivakumar, D., & Tomkins, A. (2002). Self-similarity in the web. ACM Transactions on Internet Technology, 2(3), 205–223.

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 6th International Conference of Information Fusion, vol. 2, pp. 1324–1333). Australia.

Directorate-General for Maritime Affairs and Fisheries (EC). (2010). *Integrating maritime surveillance.* European Commission. https://doi.org/10.2771/64104

Dobrkovic, A., Iacob, M.-E., & van Hillegersberg, J. (2015). *Using machine learning for unsupervised maritime waypoint discovery from streaming AIS data.* (Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business, p. 16).

Dobrkovic, A., Iacob, M.-E., & van Hillegersberg, J. (2018). Maritime pattern extraction and route reconstruction from incomplete AIS data. *International Journal of Data Science and Analytics*, *5*(2-3), 111–136.

Dorofeyuk, A. A., Pokrovskaya, I. V., & Chernyavkii, A. L. (2004). Expert methods to analyze and perfect management systems. *Automation and Remote Control*, *65*(10), 1675–1688. https://doi.org/10.1023/B:AURC.0000044276.99273.a0

Douglas, D. H., & Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *10*(2), 112–122.

Eiden, G., & Martinsen, T. (2010). *Maritime traffic density-results of PASTA MARE project. Preparatory Action for Assessment of the Capacity of Spaceborne Automatic Identification System Receivers to Support EU Maritime Policy*. Technical Note 4.1 Vessel Density Mapping.

Eleye-Datubo, A. G., Wall, A., & Wang, J. (2008). Marine and offshore safety assessment by incorporative risk modeling in a fuzzy-Bayesian network of an induced mass assignment paradigm. *Risk Analysis*, *28*(1), 95–112. https://doi.org/10.1111/j.1539-6924.2008.01004.x

Ellis, J., Forsman, B., Gehl, S., Langbecker, U., Riedel, K.,& Sames, P. C. (2008). A risk model for the operation of container vessels. *WMU Journal of Maritime Affairs*, *7*(1), 133–149. https://doi.org/10.1007/BF03195128

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge*, *19*(1), 1–16.

el Pozo, F., Dymock, A., Feldt, L., Hebrard, P., & di Monteforte, F. S. (2010). *Maritime surveillance in support of CSDP*. European Defence Agency.

Elsayed, T. (2009). Fuzzy inference system for the risk assessment of liquefied natural gas carriers during loading/offloading at terminals. *Applied Ocean Research*, *31*(3), 179–185. https://doi.org/10.1016/j.apor.2009.08.004

Endrina, N., Rasero, J. C., & Konovessis, D. (2018). Risk analysis for RoPax vessels: A case of study for the Strait of Gibraltar. *Ocean Engineering*, *151*, 141–151. https://doi.org/10.1016/j.oceaneng.2018.01.038

Eppler, M. J. (2006). *Managing information quality: Increasing the value of information in knowledge-intensive products and processes* (2nd ed.). Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/3-540-32225-6

Equasis. (2013). *The world merchant fleet in 2013. Statistics from Equasis*. Retrieved from https://www.equasis.org/Fichiers/Statistique/MOA/AnnualStatistics/Equasis Statistics-Theworldfleet2013.pdf

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*. (Kdd'96: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231).

Ester, M., & Wittmann, R. (1998). *Incremental generalization for mining in a data warehousing environment*. (International Conference on Extending Database Technology, pp. 135–149).

European Parliament. (2009). Regulation (EC) No. 223/2009 of the European Parliament and the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom). *Official Journal of the European Union*, *52*.

European Statistical System. (2014). *ESS handbook for quality reports.* Eurostat.

European Union. (2003). Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. *Official Journal of the European Union*, *46*(L 345), 90–96.

Fagerholt, K. (2004). A computer-based decision support system for vessel fleet scheduling—Experience and future research. *Decision Support Systems*, *37*(1), 35–47. https://doi.org/10.1016/S0167-9236(02)00193-8

Fagerholt, K., & Lindstad, H. (2007). TurboRouter: An interactive optimisation-based decision support system for ship routing and scheduling. *Maritime Economics & Logistics*, (9), 214–233. https://doi.org/10.1057/palgrave.mel.9100180

Faithfull,W. (2017). *Change detection for software engineers part I: Introduction and CUSUM.* Retrieved December 29, 2019 from https://faithfull.me/change-detection-for-software-engineers-part-i-introduction-and-cusum/

Fang, M. C., & Lin, Y. H. (2015). The optimization of ship weather-routing algorithm based on the composite influence of multi-dynamic elements (ii): Optimized routings. *Applied Ocean Research*, *50*, 130–140.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37.

Ferrer, J., Karlberg, J., & Hintlian, J. (2007). Integration: The key to global success. *Supply Chain Management Review*, *11*(2).

Ficoń, K. (2010). *Logistyka morska. Statki, porty, spedycja.* Warszawa: BEL Studio.

Ficoń, K., & Sokołowski, W. (2012). Środki transportu morskiego w zapewnieniu bezpieczeństwa dostaw gazu ziemnego. *Logistyka*, *401.*

Filipiak, D., Stróżyna, M., Węcel, K., & Abramowicz, W. (2018). Big data for anomaly detection in maritime surveillance: spatial ais data analysis for tankers. *Zeszyty Naukowe Akademii Marynarki Wojennej*, *59*.

Filipiak, D., Stróżyna, M., Węcel, K., Abramowicz, W., & Steidel, M. (2021). *Application of AI and in-memory computing for extracting vessel movement patterns from historical data.* (14th NATO Operations Research and Analysis Conference: Emerging and Disruptive Technology: Meeting Proceedings).

Filipiak, D., Węcel, K., Stróżyna, M., Michalak, M., & Abramowicz, W. (2020). Extracting maritime traffic networks from AIS data using evolutionary algorithm. *Business & Information Systems Engineering*, *62*(5), 435–450.

Fischer, Y., & Bauer, A. (2010, November). *Object-oriented sensor data fusion for wide maritime surveillance.* (2010 International Waterside Security Conference, p. 1–6). https://doi.org/10.1109/WSSC.2010.5730244

Fooladvandi, F., Brax, C., Gustavsson, P., & Fredin, M. (2009). *Signature-based activity detection based on Bayesian networks acquired from expert knowledge.* (12th International Conference on Information Fusion, pp. 436–443).

Fossen, T. I. (2011). *Handbook of marine craft hydrodynamics and motion control.* Hoboken: John Wiley & Sons.

Franceschini, F., & Rafele, C. (2000). Quality evaluation in logistic services. *International Journal of Agile Management Systems*, *2*(1), 49–54.

Gaonkar, R. S. P., Xie, M., Ng, K. M., Habibullah, M. S., Prabhu Gaonkar, R. S., Xie, M., ..., Habibullah, M. S. (2011). Subjective operational reliability assessment of maritime

transportation system. *Expert Systems with Applications*, *38*(11), 13835–13846. https://doi.org/10.1016/j.eswa.2011.04.187

Gerigk, M. (2012). Ocena ryzyka i zarządzanie bezpieczeństwem w czasie katastrofy obiektu oceanotechnicznego lub statku na morzu. *Prace Naukowe Politechniki Warszawskiej*, 25-40.

Gilberg, A., Kleiven, E., & Bye, R. J. (2016). *Marine navigation accidents and influencing conditions: An exploratory statistical analysis using AIS data and accident databases*. (Risk, Reliability and Safety: Innovating Theory and Practice: Proceedings of ESREL 2016, Glasgow, Scotland, 25–29 September 2016).

Giraud, M.-A., Alhadef, B., Guarnieri, F., Napoli, A., Bottala-Gambetta, M., Chaumartin, D., ..., Michel, P. (2011). *SARGOS: Securing offshore infrastructures through a global alert and graded response system.* (Mast Europe 2011—Maritime System and Technology, 7th Global Conference & Eexhibition, Global Maritime Cooperation). Marseille, France. Retrieved from https://hal-mines-paristech.archives-ouvertes.fr/hal-00660219

Goerlandt, F., & Montewka, J. (2015). Maritime transportation risk analysis: Review and analysis in light of some foundational issues. *Reliability Engineering & System Safety*, *138*, 115–134. https://doi.org/10.1016/j.ress.2015.01.025

Greidanus, H., Alvarez, M., Eriksen, T. K., Argentieri, P., Çokacar, T., Pesaresi, A., ..., Alessandrini, A. (2013). Basin-wide maritime awareness from multisource ship reporting data. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, *7*(2), 185–192.

Grêt-Regamey, A., & Straub, D. (2006). Spatially explicit avalanche risk assessment linking Bayesian Networks to a GIS. *Natural Hazards and Earth System Science*, *6*(6), 911–926.

Grzelakowski, A. (2009). Maritime transport development in the global scale–the main chances, threats and challenges. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, *3*(2), 17–18.

Grzelakowski, A. (2012). Globalizacja i jej wpływ na rozwój transportu morskiego i globalnych łańcuchów dostaw. Wyzwania gospodarki globalnej. *Prace i Materiały Instytutu Handlu Zagranicznego Uniwersytetu Gdańskiego*, (31). Retrieved from https://ekonom.ug.edu.pl/web/download.php?OpenFile=951

Gyftakis, S., Koromila, I., Giannakopoulos, T., Nivolianitou, Z., Charou, E., & Perantonis, S. (2018). Decision support tool employing Bayesian risk framework for environmentally safe shipping. In C. Konstantopoulos, & G. Pantziou (Eds.), *Modeling, computing and data handling methodologies for maritime transportation* (pp. 117–143). Cham: Springer.

Hagiwara, H., & Spaans, J. (1987). Practical weather routing of sail-assisted motor vessels. *The Journal of Navigation*, *40*(1), 96–119.

Hahn, A. (2014). Test bed for safety assessment of new e-navigation systems. *International Journal of e-Navigation and Maritime Economy*, *1*, 14–28.

Hajduk, J. (2009). *Gospodarka morska – stan obecny, oczekiwania, potrzeby*. Retrieved from http://www.am.szczecin.pl/userfiles/File/aktualnosci/news{_}2010{_}03{_}02/referat{_}Senat{_}RP{_}J{_}Hajduk{_}3{_}2.pdf

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

Hall, M. J., Hall, S. A., & Tate, T. (2000). *Removing the HCI bottleneck: How the Human Computer Interface (HCI) affects the performance of data fusion systems.* (Proceedings of the MSS National Symposium on Sensor and Data Fusion, pp. 89–104).

Haltiner, G., Hamilton, H., & Arnason, G. (1962). Minimal-time ship routing. *Journal of Applied Meteorology*, *1*(1), 1–7.

Harati-Mokhtari, A., Wall, A., Brooks, P., & Wang, J. (2007). Automatic Identification System (AIS): Data reliability and human error implications. *Journal of Navigation*, *60*(03), 373–389.

He, B., Patel, M., Zhang, Z., & Chang, K. C. C. (2007). Accessing the Deep Web. *Communication of the ACM*, *50*(5), 94–101. https://doi.org/10.1145/1230819.1241670

Health and Safety Executive. (2015). *ALARP 'at a glance'.* Retrieved July 29, 2015 from https://www.hse.gov.uk/managing/theory/index.htm

Heinrich, B., & Klier, M. (2015). Metric-based data quality assessment—Developing and evaluating a probability-based currency metric. *Decision Support Systems*, *72*, 82–96. https://doi.org/10.1016/j.dss.2015.02.009

Helldin, T., & Riveiro, M. (2009). *Explanation methods for Bayesian Networks: Review and application to a maritime scenario.* (Proceedings of the 3rd Annual Skövde Workshop on Information Fusion Topics, pp. 11–16).

Heydon, A., & Najork, M. (1999). Mercator: A scalable, extensible web crawler. *World Wide Web*, *2*, 219–229.

Heywood, C., Connor, C., Browning, D., Smith, M. C., & Wang, J. (2009). *GPS tracking of intermodal transportation: System integration with delivery order system.* (Systems and Information Engineering Design Symposium, pp. 191–196).

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

Hoque, X., & Sharma, S. K. (2020). *Ensembled deep learning approach for maritime anomaly detection system.* (Proceedings of ICETIT 2019, pp. 862–869). Springer.

Hornauer, S., & Hahn, A. (2013). Towards marine collision avoidance based on automatic route exchange. *IFAC Proceedings*, *46*(33), 103–107.

International Maritime Organisation. (1974). *International Convention for the Safety of Life at Sea (SOLAS) with amendments.* London: IMO/ICAO.

International Maritime Ogranization. (2011). *International Maritime Dangerous Goods Code (IMDG) with amedments.* Retrieved from http://www.imo.org/en/Publications/IMDGCode/Pages/Default.aspx

International Maritime Organisation. (2013). *The international aeronautical and maritime search and rescue (IAMSAR) Manual.* London: IMO/ICAO.

International Maritime Organization. (2021). *Global Integrated Shipping Information System. Regional analysis of reports on acts of piracy and armed robbery.*

International Telecommunication Union. (2010). *Recommendation ITU-R M.1371-4. Technical characteristics for an automatic identification system using timedivision multiple access in the VHF maritime mobile band.* Retrieved from https://www.itu.int/dms%7B/_%7Dpubrec/itu-r/rec/m/R-REC-M.1371-4-201004-S!!PDF-E.pdf

Iphar, C., Napoli, A., & Ray, C. (2015a). Data quality assessment for maritime situation awareness. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *II-3*/W5, 291–296. https://doi.org/10.5194/isprsannals-II-3-W5-291-2015

Iphar, C., Napoli, A., & Ray, C. (2015b). *Detection of false AIS messages for the improvement of maritime situational awareness.* (Oceans'15 MTS/IEEE, pp. 1–7). Washington.

Jackson, P., & Moulinier, I. (2002). *Natural language processing for online applications. Text retrieval, extraction and categorization.* Amsterdam: John Beniamins Publishing.

James, R. W. (1957). *Application of wave forecasts to marine navigation.* Washington: U.S. Naval Oceanographic Office.

Janvier-James, A. M. (2012). A new introduction to supply chains and supply chain management: Definitions and theories perspective. *International Business Research*, *5*(1), 194–207.

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.

Jarysz-Kamińska, E. (2013). Ocena ryzyka w transporcie morskim. *Logistyka*, *6*, 238–246.

Johansson, F., & Falkman, G. (2007). *Detection of vessel anomalies—a Bayesian network approach.* (3rd International Conference on Intelligent Sensors, Sensor Networks and Information, pp. 395–400).

Kaczmarek, T., & Węckowski, D. (2013). Harvesting Deep Web data through Produser involvement. In M. Pankowska (Ed.), *Frameworks of IT prosumption for business development* (pp. 200–221). Hershey: IGI Global.

Kaczmarek, T. T. (2010). *Zarządzanie ryzykiem. Ujęcie interdyscyplinarne.* Warszawa: Difin.

Kaczmarek, T. T. (2012). *Zarządzanie ryzykiem w handlu międzynarodowym.* Warszawa: Difin.

Kangsvik, T., Størkersen, K. V., & Antonsen, S. (2017). One size fits all? Safety management regulation of ship accidents and personal injuries. *Journal of Risk Research*, *20*(9), 1154–1172.

Karataş, G. B., Karagoz, P., & Ayran, O. (2021). *Trajectory pattern extraction and anomaly detection for maritime vessels.* Internet of Things, 100436. https://doi.org/10.1016/j.iot.2021.100436

Kazemi, S., Abghari, S., Lavesson, N., Johnson, H., & Ryman, P. (2013). Open data for anomaly detection in maritime surveillance. *Expert Systems with Applications*, *40*(14), 5719–5729.

Kim, K. H., & Lee, H. (2015). Container terminal operation: Current trends and future challenges. In C. Y. Lee, & Q. Meng (Eds.), *Handbook of ocean container transport logistics* (pp. 43–73). International Series in Operations Research & Management Science, vol 220. Cham: Springer.

Kobayashi, M., & Takeda, K. (2000). Information retrieval on the Web. *ACM Computer Surveys*, *32*(2), 144–173. https://doi.org/10.1145/358923.358934

Kosmas, O., & Vlachos, D. (2012). Simulated annealing for optimal ship routing. *Computers & Operations Research*, *39*(3), 576–581.

Koster, R. (1996). *A standard for robot exclusion.* Retrieved from http://www.robotstxt.org/orig.html

Kowalkiewicz, M., Safrudin, N., & Schulze, B. (2017). The business consequences of a digitally transformed economy. In G. Oswald, & M. Kleinemeier (Eds.), *Shaping the digital enterprise* (pp. 29–67). Cham: Springer.

Kraiman, J. B., Arouh, S. L., & Webb, M. L. (2002). Automated anomaly detection processor. In A. F. Sisti & D. A. Trevisani (Eds.), *Proceedings of spie: Enabling technologies for simulation science vi* (pp. 128–137). Orlando.

Lam, J. (2012). *Rough set approach to marine cargo risk analysis.* (International Forum on Shipping, Ports and Airports (IFSPA): Transport Logistics for Sustainable Growth at a New Level, pp. 1–12).

Lamm, A., & Hahn, A. (2017). *Detecting maneuvers in maritime observation data with CUSUM.* (2017 IEEE International Symposium on Signal Processing and Information Technology, pp. 122–127).

Lamm, A., & Hahn, A. (2019). *Statistical maneuver net generation for anomaly detection in navigational waterways.* (2019 6th International Conference on Control, Decision and Information Technologies, pp. 1438–1443).

Lane, R. O., Nevell, D. A., Hayward, S. D., & Beaney, T. W. (2010). *Maritime anomaly detection and threat assessment.* (2010 13th Conference on Information Fusion, pp. 1–8).

Läsche, C., Pinkowski, J., Gerwinn, S., Droste, R., & Hahn, A. (2014). *Model-based risk assessment of offshore operations.* (ASME 33rd International Conference on Ocean, Offshore and Arctic Engineering, 45387).

Laxhammar, R. (2008). *Anomaly detection for sea surveillance.* (2008 11th International Conference on Information Fusion, pp. 1–8).

Laxhammar, R., & Falkman, G. (2010). *Conformal prediction for distribution-independent anomaly detection in streaming vessel data.* Proceedings of the first international workshop on novel data stream pattern mining techniques (pp. 47–55).

Laxhammar, R., Falkman, G., & Sviestins, E. (2009). *Anomaly detection in sea traffic—a comparison of the gaussian mixture model and the kernel density estimator.* (2009 12th International Conference on Information Fusion, pp. 756–763).

Lee, C. J., & Lee, K. J. (2006). Application of Bayesian network to the probabilistic risk assessment of nuclear waste disposal. *Reliability Engineering & System Safety*, *91*(5), 515–532.

Li, S., Meng, Q., & Qu, X. (2012). An overview of maritime waterway quantitative risk assessment models. *Risk Analysis*, *32*(3), 496–512.

Liu, J., Yang, J. B., Wang, J., & Sii, H. S. (2005). Engineering system safety analysis and synthesis using the fuzzy rule-based evidential reasoning approach. *Quality and Reliability Engineering International*, *21*(4), 387–411. https://doi.org/10.1002/qre.668

Liu, K. F. R., Lu, C. F., Chen, C. W., & Shen, Y. S. (2012). Applying Bayesian belief networks to health risk assessment. *Stochastic Environmental Research and Risk Assessment*, *26*(3), 451–465.

Lloyd. (2013). *Lloyd's Risk Index 2013.* Retrieved from https://www.ipsos.com/sites/default/files/publication/1970-01/loyalty-lloyds-risk-index-2013-report.pdf

Lu, C.-S. (2000). Logistics services in Taiwanese maritime firms. *Transportation Research Part E: Logistics and Transportation Review*, *36*(2), 79–96.

Maki, A., Akimoto, Y., Nagata, Y., Kobayashi, S., Kobayashi, E., Shiotani, S., ..., Umeda, N. (2011). A new weather-routing system that accounts for ship stability based on a real-coded genetic algorithm. J*ournal of Marine Science and Technology*, *16*(3), 311.

Małyszko, J., Abramowicz, W., & Stróżyna, M. (2016). Named entity disambiguation for maritime-related data retrieved from heterogenous sources. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, *10*(3), 465–477.

Mano, J. P., Georgé, J. P., & Gleizes, M. P. (2010). Adaptive multi-agent system for multi-sensor maritime surveillance. In T. Ditzinger (Ed.), *Advances in practical applications of agents and multiagent systems* (pp. 285–290). Cham: Springer.

Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Doshi, E. A. (2013). *Open data: Unlocking innovation and performance with liquid information.* McKinsey Global Institute. Retrieved from https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/open%20data%20unlocking%20innovation%20and%20performance%20with%20liquid%20information/mgi_open_data_fullreport_oct2013.ashx

Marie, S., Courteille, E. (2009). Multi-objective optimization of motor vessel route. *Proceedings of the International Symposium TransNav*, *9*, 411–418.

Marine Management Organisation. (2014). *Mapping UK shipping density and routes from AIS; MMO Project No: 1066.* Newcastle: Marine Management Organisation.

Markowski, A. S., Mannan, M. S., & Bigoszewska, A. (2009). Fuzzy logic for process safety analysis. *Journal of Loss Prevention in the Process Industries*, *22*(6), 695–702.

Martineau, E., & Roy, J. (2011). *Maritime anomaly detection: domain introduction and review of selected literature.* Defence Research and Development Canada Valcartier. Retrieved from https://apps.dtic.mil/sti/pdfs/ADA554310.pdf

Mascaro, S., *Nicholson, A. E., & Korb, K. B. (2011). Anomaly detection in vessel tracks using Bayesian Networks.* (Proceedings of the 8th Bayesian modeling applications workshop, *818*, 99–107). https://doi.org/10.1016/j.ijar.2013.03.012

Mascaro, S., Nicholson, A. E., & Korb, K. B. (2014). Anomaly detection in vessel tracks using Bayesian Networks. *International Journal of Approximate Reasoning*, *55*(1), 84–98.

Matear, S., & Gray, R. (1993). Factors influencing freight service choice for shippers and freight suppliers. *International Journal of Physical Distribution & Logistics Management*, *23*(2), 25–35.

Matthews, M., Martin, L. B., Tario, C. D., & Brown, A. L. (2009). *A non-intrusive alert system for maritime anomalies: Literature review and the development and assessment of interface design concepts.* Canada: DTIC Document.

Mazaheri, A. (2017). *A framework for evidence-based risk modeling of ship grounding.* (Doctoral dissertation). Aalto: Aalto University. Retrieved from http://urn.fi/URN:ISBN:978-952-60-7478-8

Mazaheri, A., Montewka, J., & Kujala, P. (2013). Modeling the risk of ship grounding—a literature review from a risk management perspective. *WMU Journal of Maritime Affairs*, *13*, 269–297. https://doi.org/10.1007/s13437-013-0056-3

Mazzarella, F., Arguedas, V. F., & Vespe, M. (2015). *Knowledge-based vessel position prediction using historical AIS data.* (2015 Sensor Data Fusion: Trends, Solutions, Applications, pp. 1–6).

Mazzarella, F., Vespe, M., Damalas, D., & Osio, G. (2014). *Discovering vessel activities at sea using AIS data: Mapping of fishing footprints.* (17th International Conference on Information Fusion, pp. 1–7).

Mestl, T., Tallakstad, K. T., & Castberg, R. (2016). Identifying and analyzing safety critical maneuvers from high resolution AIS data. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, *10*.

Miler, R. K. (2015). *Bezpieczeństwo transportu morskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Mutke, S., Augenstein, C., Roth, M., Ludwig, A., & Franczyk, B. (2015). Real-time information acquisition in a model-based integrated planning environment for logistics contracts. *Journal of Object Technology*, *14*(1), 1–2.

Nahari, M. K., Ghadiri, N., Jafarifard, Z., Dastjerdi, A. B., & Sack, J. R. (2017). *A framework for linked data fusion and quality assessment.* (2017 3th International Conference on Web Research, pp. 67–72).

Nguyen, D., Vadaine, R., Hajduch, G., Garello, R., & Fablet, R. (2021). GeoTrackNet—a maritime anomaly detector using probabilistic neural network representation of AIS tracks and a contrario detection. *IEEE Transactions on Intelligent Transportation Systems*, 1–13. https://doi.org/10.1109/TITS.2021.3055614

Nivolianitou, Z. S., Koromila, I. A., & Giannakopoulos, T. (2016). Bayesian Network to predict environmental risk of a possible ship accident. *International Journal of Risk Assessment and Management*, *19*(3), 228–239.

Nowakowski, T. (2011). *Niezawodność systemów logistycznych.* Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Oltmann, J. H. (2015). *ACCSEAS North Sea Region Route Topology Model (NSRRTM). ACCSEAS Project.* Retrieved from http://archive.northsearegion.eu/files/repository/20150519163329_ACCSEASRouteTopologyModelReport.PDF

Opengeospatial.org. (2018). *OGC standard netCDF Classic and 64-bit Offset*. Retrieved December 27, 2018 from http://www.opengeospatial.org/standards/netcdf

Paixão Casaca, A. C., & Marlow, P. B. (2005). The competitiveness of short sea shipping in multimodal logistics supply chains: Service attributes. *Maritime Policy & Management*, *32*(4), 363–382.

Pallotta, G., Horn, S., Braca, P., & Bryan, K. (2014). *Context-enhanced vessel prediction based on ornstein-uhlenbeck processes using historical ais traffic patterns: Real-world experimental results.* (17th International Conference on Information Fusion, pp. 1–7).

Pallotta, G., Vespe, M., & Bryan, K. (2013). Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction. *Entropy*, *15*(6), 2218–2245.

Patino, L., & Ferryman, J. (2017). *Loitering behaviour detection of boats at sea.* (Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 44–50).

Patrick, G., Davies, H., Baldacci, A., & den Breejen, E. (2015). *The addition of near real time data and forecast data. Deliverable of ProMerc project. Deliverable of ProMerc project.* Retrieved from http://www.promerc.eu

Pedersen, P. T. (1995). Collision and grounding mechanics. *Proceedings of WEMT*, *95*(1995), 125–157.

Pietrzykowski, Z. (2011). Navigational decision support system as an element of intelligent transport systems. *Scientific Journals Maritime University of Szczecin*, *25*(97), 41–47.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002, April). Data quality assessment. *Commununication of the ACM*, *45*(4), 211–218. https://doi.org/10.1145/505248.506010

Pollino, C. A., Woodberry, O., Nicholson, A., Korb, K., & Hart, B. T. (2007). Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment. *Environmental Modelling & Software*, *22*(8), 1140–1152.

Qi, K. (2016). *A scalable framework for AIS data analysis and visualisation.* (Unpublished master's thesis). Dalhousie University.

Qi, Y. G., Martinelli, D. R., Teng, H. H., & Jiang, P. (2010, April). An application of the CUSUM algorithm to freeway incident detection based on two contiguous detectors. *Journal of Advanced Transportation*, *39*(2), 221–241. https://doi.org/10.1002/atr.5670390206

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bullettin*, *23*(4), 3–13.

Rhodes, B. J., Bomberger, N. A., Seibert, M., & Waxman, A. M. (2005). *Maritime situation monitoring and awareness using learning mechanisms.* (Military Communications Conference, 2005, pp. 646–652).

Ristic, B. (2014). Detecting anomalies from a multitarget tracking output. *IEEE Transactions on Aerospace and Electronic Systems*, *50*(1), 798–803.

Riveiro, M. (2011). *Visual analytics for maritime anomaly detection.* (Unpublished doctoral dissertation). Orebro University.

Riveiro, M., Falkman, G., & Ziemke, T. (2008). *Improving maritime anomaly detection and situation awareness through interactive visualization.* (2008 11th International Conference on Information Fusion, pp. 1–8).

Riveiro, M., Pallotta, G., & Vespe, M. (2018). Maritime anomaly detection: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(5), e1266.

Robey, D., & Markus, M. L. (1984). Rituals in information system design. *MIS Quarterly*, 5–15.

Robinson, J. T. (1981). *The KDB-tree: A search structure for large multidimensional dynamic indexes.* (Proceedings of the 1981 ACM Sigmod International Conference on Management of Data, pp. 10–18).

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1–2), 1–39.

Rong, H., Teixeira, A., & Guedes Soares, C. (2020). Data mining approach to shipping route characterization and anomaly detection based on AIS data. *Ocean Engineering*, *198*, 106936. https://doi.org/10.1016/j.oceaneng.2020.106936

Roy, J. (2008). *Anomaly detection in the maritime domain.* (Proceedings of Spie. Optics and Photonics in Global Homeland Security iv, p. 69450W).

Roy, J., & Davenport, M. (2009). *Categorization of maritime anomalies for notification and alerting purpose.* Defence Research & Development Canada Valcartier.

Royal Society Study Group. (1983). *Risk assessment: Report of a Royal Society Study Group.* London: Royal Society.

Rydzkowski, W., & Wojewódzka-Król, K. (2007). *Transport.* Warszawa: Wydawnictwo Naukowe PWN.

Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, *48*(1), 9–26.

Sahay, B. S., Seth, N., Deshmukh, S. G., & Vrat, P. (2006). A conceptual model for quality of service in the supply chain. *International Journal of Physical Distribution & Logistics Management*, *36*(7), 547–575.

Samet, H. (1984). The quadtree and related hierarchical data structures. *ACM Computing Surveys*, *16*(2), 187–260.

Samson, N., & Ibitoru, D. F. (2015). Modeling cargo ship behavior in extreme rough weather condition. *International Journal of Scientific & Engineering Research*, *6*(7), 782–792.

Sauer, O., & Norkus, J. (2015). *A taxonomy for describing BI cloud services.* (The International Conference on Semantic Web Business and Innovation, p. 1).

Schøyen, H., & Bråthen, S. (2015). Measuring and improving operational energy efficiency in short sea container shipping. *Research in Transportation Business & Management*, *17*, 26–35.

Shao, W., Zhou, P., & Thong, S. K. (2012). Development of a novel forward dynamic programming method for weather routing. *Journal of Marine Science and Technology*, *17*(2), 239–251.

Shelmerdine, R. L. (2015). Teasing out the detail: How our understanding of marine AIS data can better inform industries, developments, and planning. Marine Policy, 54, 17–25.

Shestakov, D., Bhowmick, S. S., & Lim, E. P. (2005). DEQUE: Querying the deep web. *Data & Knowledge Engineering*, *52*(3), 273–311.

Sidibé, A., & Shu, G. (2017). Study of automatic anomalous behaviour detection techniques for maritime vessels. *The Journal of Navigation*, *70*(4), 847–858.

Singh, S. K., & Heymann, F. (2020). *Machine learning-assisted anomaly detection in maritime navigation using AIS data.* (2020 IEEE/ION Position, Location and Navigation Symposium, p. 832–838). https://doi.org/10.1109/PLANS46316.2020.9109806

Sivanandam, S., & Deepa, S. (2008). Genetic algorithms. In S. Sivanandam, & S. Deepa (Eds.), *Introduction to genetic algorithms* (pp. 15–37). Cham: Springer.

Smith, M., Reece, S., Roberts, S. J., & Rezek, I. (2012). *Online maritime abnormality detection using Gaussian processes and extreme value theory.* (2012 IEEE 12th International Conference on Data Mining, pp. 645–654).

Soares, C. G., & Teixeira, A. P. (2001). Risk assessment in maritime transportation. *Reliability Engineering and System Safety*, *74*(3), 299–309. https://doi.org/10.1016/S0951-8320(01)00104-1

Stemmler, L. (2007). Risk in the supply chain. In D. Waters (Ed.), *Global logistics: New directions in supply chain management* (pp. 210–222). London: Kogan Page Limited.

Straub, D. (2005). Natural hazards risk assessment using Bayesian Networks. *Safety and Reliability of Engineering Systems and Structures*, 2535–2542.

Stróżyna, M. (2017a). A Bayesian Network approach to assessing the risk and reliability of maritime transport. In W. Abramowicz, R. Alt, & B. Franczyk (Eds.), *Business information systems workshops* (pp. 367–378). Cham: Springer International Publishing.

Stróżyna, M. (2017b). Hazard index for assessment of reliability of supply and risk in maritime domain. In W. Abramowicz (Ed.), *Business information systems workshops* (pp. 228–241). Cham: Springer.

Stróżyna, M., & Abramowicz, W. (2015). A dynamic risk assessment for decision support systems in the maritime domain. *Studia Ekonomiczne*, *243*, 295–307.

Stróżyna, M., Eiden, G., Abramowicz, W., Filipiak, D., Małyszko, J., & Węcel, K. (2018). A framework for the quality-based selection and retrieval of open data—a use case from the maritime domain. *Electronic Markets*, *28*(2), 219–233.

Stróżyna, M., Eiden, G., Filipiak, D., Małyszko, J., & Węcel, K. (2016). *A methodology for quality-based selection of Internet data sources in maritime domain.* (19th International Conference on Business Information System (BIS), Poznań, pp. 15–27). https://doi.org/10.1007/978-3-319-39426-8_2

Stróżyna, M., Filipiak, D., & Węcel, K. (2020). Data quality assessment—a use case from the maritime domain. In W. Abramowicz & G. Klein (Eds.), *Business information systems workshops* (pp. 5–20). Cham: Springer.

Stróżyna, M., Małyszko, J., Węcel, K., Filipiak, D., & Abramowicz,W. (2016). Architecture of maritime awareness system supplied with external information. *Annual of Navigation*, *23*(1), 135–149.

Szelangiewicz, T., Wiśniewski, B., & Żelazny, K. (2014). The influence of wind, wave and loading condition on total resistance and speed of the vessel. *Polish Maritime Research*, *21*(3), 61–67.

Szlapczynska, J., & Smierzchalski, R. (2009). Multicriteria optimisation in weather routing. *Marine Navigation and Safety of Sea Transportation*, *423*.

Szymanek, A. (2008). Risk acceptation principles in transport. *Journal of KONBiN*, *5*(2), 271–281.

Tan, W. C., Weng, C. Y., Zhou, Y., Chua, K. H., & Chen, I. M. (2018). *Historical data is useful for navigation planning: Data driven route generation for autonomous ship.* (2018 IEEE International Conference on Robotics and Automation, pp. 7478–7483).

Trucco, P., Cagno, E., Ruggeri, F., & Grande, O. (2008). A Bayesian Belief Network modelling of organisational factors in risk analysis: A case study in maritime transportation. *Reliability Engineering and System Safety*, *93*(6), 845–856. https://doi.org/10.1016/j.ress.2007.03.035

Trujillo, G., Kim, C., Jones, S., Garcia, R., & Murray, J. (2015). *Virtualizing hadoop: How to install, deploy, and optimize hadoop in a virtualized architecture.* VMware Press.

Tsou, M. C., & Cheng, H. C. (2013). An ant colony algorithm for efficient ship routing. *Polish Maritime Research*, *20*(3), 28–38.

Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., & Huang, G. B. (2017). Exploiting AIS data for intelligent maritime navigation: A comprehensive survey from data to methodology. *IEEE Transactions on Intelligent Transportation Systems*, *19*(5), 1559–1582.

Tun, M. H., Chambers, G. S., Tan, T., & Ly, T. (2007). Maritime port intelligence using AIS data. In P. Mendis, J. Lai, E. Dawson, & H. Abbass (Eds.), *Recent advances in security technology* (p. 33). Melbourne: Proceedings of the 2007 RNSA Security Technology Conference.

UNCTAD. (2013). *Review of maritime transport 2013, Chapter 1.* United Nations Conference on Trade and Development.

UNCTAD. (2017). *Review of maritime transport 2017 (technical report).* United Nations Conference on Trade and Development. Retrieved from https://unctad.org/system/files/official-document/rmt2017_en.pdf

United Nations. (1982). *United Nations convention on the law of the sea.*

United Nations. (2015). *UNCTAD Stat Data Center*. Retrieved from http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx

Urbański, J., Morgaś, W., & Specht, C. (2008). Bezpieczeństwo morskie—ocena i kontrola ryzyka. *Zeszyty Naukowe Akademii Marynarki Wojennej*, *2*(173), 53–68.

van Laere, J., & Nilsson, M. (2009). *Evaluation of a workshop to capture knowledge from subject matter experts in maritime surveillance.* (2009 12th International Conference on Information Fusion, pp. 171–178).

Varlamis, I., Tserpes, K., Etemad, M., Júnior, A. S., & Matwin, S. (2019, March 26). *A network abstraction of multi-vessel trajectory data for detecting anomalies.* (Proceedings of the Workshops of the EDBT/ICDT 2019 Joint Conference), Lisbon.

Vassiliadis, P. (2009). A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining*, *5*(3), 1–27.

Veldhuis, H. D. (2015). *Developing an automated solution for ETA definition concerning long distance shipping.* (Unpublished doctoral dissertation). University of Twente.

Venskus, J., Treigys, P., Bernatavičienė, J., Tamulevičius, G., & Medvedev, V. (2019). Real-time maritime traffic anomaly detection based on sensors and history data embedding. *Sensors*, *19*(17). https://doi.org/10.3390/s19173782

Vernimmen, B., Dullaert, W., & Engelen, S. (2007). Schedule unreliability in liner shipping: Origins and consequences for the hinterland supply chain. *Maritime Economics & Logistics*, *9*(3), 193–213.

Vespe, M., Sciotti, M., & Battistello, G. (2008). *Multi-sensor autonomous tracking for maritime surveillance.* (2008 International Conference on Radar, pp. 525–530).

Vettor, R., & Soares, C. G. (2016). Development of a ship weather routing system. *Ocean Engineering*, *123*, 1–14.

Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die Reine und Angewandte Mathematik*, (133), 97–178. https://doi.org/10.1515/crll.1908.133.97

Wan, C., Yan, X., Zhang, D., Qu, Z., & Yang, Z. (2019). An advanced fuzzy Bayesian-based FMEA approach for assessing maritime supply chain risks. *Transportation Research Part E: Logistics and Transportation Review*, *125*, 222–240.

Wan, C., Yan, X., Zhang, D., & Yang, Z. (2019). Analysis of risk factors influencing the safety of maritime container supply chains. *International Journal of Shipping and Transport Logistics*, *11*(6), 476–507.

Wang, H.-B., Li, X.-G., Li, P.-F., Veremey, E. I., & Sotnikova, M. V. (2018). Application of real-coded genetic algorithm in ship weather routing. *Journal of Navigation*, *71*(4), 989–1010. https://doi.org/10.1017/S0373463318000048

Wang, R. Y., Reddy, M. P., & Kon, H. B. (1995). Toward quality data: An attribute based approach. *Decision Support Systems*, *13*(3), 349–372.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Wang, X., Liu, X., Liu, B., de Souza, E. N., & Matwin, S. (2014). *Vessel route anomaly detection with Hadoop MapReduce.* (2014 IEEE International Conference on Big Data, pp. 25–30).

Waters, D. (2011). *Supply chain risk management. Vulnerability and resilience in logistics* (2nd ed.). London: Kogan Page Limited.

Weber, P., Medina-Oliva, G., Simon, C., & Iung, B. (2012). Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. *Engineering Applications of Artificial Intelligence*, *25*(4), 671–682.

Węcel, K., & Lewoniewski, W. (2015). Modelling the quality of attributes in Wikipedia Infoboxes. In W. Abramowicz (Ed.), *Business information systems workshops* (vol. 228, pp. 308–320). Springer. https://doi.org/10.1007/978-3-319-26762-3_27

Węcel, K., Nuevo, M., Filipiak, D., Małyszko, J., Stróżyna, M., & Abramowicz, W. (2016). *SIMMO Project. Deliverable 3.3: Intelligence analysis module.* Poznań University of Economics.

Wentland, W., Knopp, J., Silberer, C., & Hartung, M. (2008). *Building a multilingual lexical resource for named entity disambiguation, translation and transliteration.* (Proceedings of the 6th International Conference on Language Resources and Evaluation).

Wielgosz, M., Wiśniewski, B., & Korwin-Piotrowski, T. (2012). Navigational aspects of ship's voyage planning taking into account calculations of eta (estimated time of arrival). *Scientific Journals of the Maritime University of Szczecin*, (29), 182–187.

Wieteska, G. (2011). *Zarządzanie ryzykiem w łańcuchu dostaw na rynku B2B.* Warszawa: Difin.

Wu, L., Xu, Y., Wang, Q., Wang, F., & Xu, Z. (2017). Mapping global shipping density from AIS data. *The Journal of Navigation*, *70*(1), 67–81.

Xiao, Z., Ponnambalam, L., Fu, X., & Zhang, W. (2017). Maritime traffic probabilistic forecasting based on vessels' waterway patterns and motion behaviors. *IEEE Transactions on Intelligent Transportation Systems*, *18*(11), 3122–3134.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., ..., Stoica, I. (2012). *Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing.* (Proceedings of the 9th Usenix Conference on Networked Systems Design and Implementation, p. 2).

Zhao, L., & Shi, G. (2019). Maritime anomaly detection using density-based clustering and recurrent neural network. *Journal of Navigation*, *72*(4), 894–916. https://doi.org/10.1017/S0373463319000031

Zhen, R., Jin, Y., Hu, Q., Shao, Z., & Nikitakos, N. (2017). Maritime anomaly detection within coastal waters based on vessel trajectory clustering and Naïve Bayes classifier. *Journal of Navigation*, *70*(3), 648–670. https://doi.org/10.1017/S0373463316000850

Zissis, D., Xidias, E. K., & Lekkas, D. (2016). Real-time vessel behavior prediction. *Evolving Systems*, *7*(1), 29–40.

LIST OF TABLES

LIST OF FIGURES

The book delivers an important contribution both to theory and practice at the intersection of maritime transport in global economies, reliability and risk assessment and information systems and data processing. Besides embedding the research into current state-of-the-art the book delivers novel methods for relevant research problems, which were rigorously evaluated with real-world data and use cases (...). In particular, the enormous amount of data that exists in the maritime context and the professional way the authors use the data for evaluation with regards to accuracy, real-world example compliance, efficiency, and usefulness needs to be highlighted.

These new methods can be used by different stakeholders such as shippers, port terminals, carriers, freight forwarders or customs. Furthermore, the presented methods can also be applied to other modes of transportation, thus can be generalized and applied to other contextual fields and advance the overall topic as such.

André Ludwig, Kühne Logistics University, Hamburg

eISBN 978-83-8211-137-8