GIMA

Geographical Information Management and Applications

Identification, analysis, and mapping of collision risk in the Strait of Istanbul using AIS data

Jesse Dijkstra

3 June 2022

Thesis for the Degree of Master of Science

Summary

The Strait of Istanbul is one of the busiest waterways in the world. It is also one of the smallest corridors in the worlds naval travel network. Connecting the Black Sea with the Mediterranean seas, the Strait of Istanbul is a very important connection between the countries located around the Black Sea and the international naval trading network. The narrow nature and the dense traffic flow results in the Strait of Istanbul to be one of the riskiest naval corridors of the world.

Naval travel is more digital than ever. For information and safety reasons, a lot of information on ship's characteristics, dynamic information and voyage related information is logged and stored. One of the ways in which this is trans ponded for all naval vessels is with AIS messages. These messages trans pond the above-mentioned information for all vessels in short time intervals together with their coordinates. This allows for the data to be analysed thoroughly making AIS messages and important factor in naval risk related research.

This research focusses on the Strait of Istanbul and the mapping of naval risk in this waterway. The type of naval risk used in this research is the risk of collision. The goal of the research is to find a method to identify risk of collision, categorise this risk of collision, analyse it, and map the results of this analysis in the best possible way.

This explorative research starts with drawing an empirical ship domain around the three different encounter types: head on, overtaking, and crossing encounters to be able to identify which ships AIS messages indicate a ship domain violation. The selection of this ship domain violation is analysed with different AIS related parameters which potentially indicate risk as found in the literature study. The ship domain violation encounters are given an indexed score for each of the parameters. These scores are combined to a final risk score using a matrix-based calculation. These values further analysed in-depth with a high-risk analysis and a hotspot analysis.

The findings of this research are that there is potential in the identification, analysis, and mapping of collision risk in the Strait of Istanbul. Mapping the movement between two ships and displaying underlying differences proves to be a complex task and the use of more relevant parameters and the inclusion of more statistical analyses could improve potential of this method further.

Table of contents

1. Introduction

1.1 Ship traffic and collision risk

Human mobility revolves greatly around transportation networks. It is important for the exchange of goods and the spread of species. 90 percent of this transportation total is done overseas. Merchant ships are one of the most important modes of transportation (Kaluza et al., 2010). This large scale of transportation also causes a complex network of vessel transportation. In the past decade, the amount of ship accidents has increased. The general safety of ships has remained stable (Eliopoulou, Papanikolaou & Voulgarellis, 2016), but it has not decreased over the past decades which might raise the question as to what causes these accidents. Though, the seas allow for a large navigational terrain for ships to travel through, some narrow passages can cause a higher traffic density in these areas potentially causing problems and danger. This is also the case with the Strait of Istanbul. The past years have seen several collisions of ships within the strait, sometimes even resulting is casualties (APnews, 2021). The Strait of Istanbul has a combined annual traffic of 300.000 vessels, being both local and transit ships. It is known to be one of the congested and risky waterways in the world (Altan & Otay, 2017).

Figure 1: Strait of Istanbul (Google Imagery, 2021)

The high-density traffic has led to several measures in the Istanbul Strait such as traffic rules and the placement of real-time monitoring stations to provide an overview of the traffic in the Turkish waterways. This monitoring, now globally adopted as AIS (Automatic Identification System), is now a standard piece of equipment on each ship. It is developed with the goal improving maritime safety by means of collision avoidance which can be used by Vessel Traffic Services (VTS). AIS equipment transmits continuous data and information on the vessels' identity, position, speed, and course to the VTS locations (Tetreault, 2005). An advantage of AIS is that equipped vessels can easily identify other AIS transmitting vessels to avoid collision.

As mariners in the Istanbul Strait do not only have to navigate a static terrain, but also run into dynamic obstacles which are other vessels. Mariners have to predetermine manoeuvres to avoid dangerous situations and possible collisions (Im & Luong, 2019). Real time data like AIS gives certain information on the surrounding vessel around a ship. However, it does not exactly give information on where a ship is safe to navigate. This leads to the fact that this determining would have to lie in the hands of the mariner, possibly leading to human errors. Several researches state that 80% to 85% of the at that time recorded maritime accidents were due to human errors (Pietrzykowski & Uriasz, 2009) (Harati-Mokhtari, Wall, Brooks & Wang, 2007). Together with this, ship related accidents can have a detrimental effect on the maritime environment, can cause human casualties, and have large economic impacts. This further emphasised the need for the mitigation of these accidents (Goerlandt & Montewka, 2015)

1.2 Problem definition

This leads to the problem of this research that the way of determining risk of collision in ship traffic has not yet been perfected to the extent of reducing overall accidents. Defining the risk of collision can help to identify risky situations. In general, the literature contains many different types of methodologies for maritime risk analyses such as: simulations, probability, and statistic-based research where many researches end with results either based on single ship encounters, or on a single specific waterway. However, to improve safety on waterways, findings like this should be able to be visualised so potential chokepoints or dangerous, often occurring, situations can be prevented. These results should also be able to flexible so it can be applied to different waterways. Categorising and mapping risk of collision between ships could lead to more insight in potential risks in the maritime environment. This leads to the main question of this research:

To what extent can situations of maritime risk of collision be mapped in the Strait of Istanbul using AIS data?

This main research question can be broken down into sub-questions that can act as a supporting factor to further define and answer the research question and define the goal of the research:

- To what extent can risk be defined in maritime traffic?
- To what extent can a risky situation be identified using AIS data?
- Can the risks be measured using categories or scores?
- To what extent can the risky situation data be analysed to find potential reoccurring problems?

The goal of this research is to ultimately improve the maritime traffic of the Strait of Istanbul. Reducing potential collisions and finding possible ways for safer travel through a high-density waterway as this traffic density might increase more over time.

Another goal of this research is to build on the scientific literature on AIS based maritime safety knowledge. Introducing an accurate method of identifying risky situations in the Strait of Istanbul would the first step into making the method or model replicable for other waterways. If the model is accurate enough, it could only require certain changes to parameters in order to function in different waterway. If this could be applied elsewhere, it will be able to analyse risky situations in many different waterways and ultimately improve the safety not only in the Strait of Istanbul, but also in other areas.

1.3 Research Limitations

This research seeks to create a model which uses the most recent findings in ship risk collision identifications. The goal of the model is to properly identify and categorise collision risks in a way in which these risks can be mapped for interpretational purposes. The conclusion of the research should include a different approach to ship collision risk identification together with a clear visualised assessment of risks and potential relations between ship collision risk and other factors which come to play in the Strait of Istanbul. Though, this research could prove to give valuable information to improve the safety of the Strait of Istanbul, the aim of this research is not to directly improve the

maritime safety of the Strait of Istanbul, this would require more research into the environmental possibilities and legal aspect of potential changes. This research uses the Strait of Istanbul as a research case and is not directly involved in solving problems in the Strait. The goal of this research is to expand upon the overall knowledge on ship collision risk detection and mapping and aims to let this cover more ground than just the Strait of Istanbul.

The research will be performed on a selection of AIS data from the Strait of Istanbul, the timeframe of the used data is one month of AIS data. One month of data should provide enough results to be able to properly analyse the data whilst not requiring too much processing power. The research limits itself to the Strait of Istanbul to scale down the amount of data and get a comprehensive case for one important area in the global maritime transport network.

2. Theoretical framework

The theoretical framework of this research will start with a comparison between the collision diameter and the ship domain approach in maritime risk assessment. Following this, a review on relevant ship risk identification parameters is performed to identify possible collision risk identifiers. Finally, a schematic overview of the findings in this study will be given together with the knowledge gap that this research aims to fill.

2.1 Maritime risk assessment

Maritime collision risk has been analysed and assessed many times throughout the last decades with further enhancement each iteration. Most of these researches base their collision risk studies around distance-based research. The amount of risk is based around relative distance between two ships. The distance which is seen as risk is based on different ship and travel related factors. Two main methods are presented to determine distance-based risk: The collision diameter method and the ship domain method.

2.1.1 Collision diameter

The most direct way of identifying collision risk is to calculate the probability of ships physically colliding. First introduced by Fujii and Tanaka (1971), and Macduff (1974), this approach uses the molecular collision theory to calculate the probability of collision. The molecular collision theory states that if two randomly moving particle centres are within a certain distance of each other, these particles will collide. This specific distance is the collision diameter. If the conditions for the collision are known, the collision diameter can be used together with the trajectory, density, and the velocity of the particles to determine the number of collisions. If this method is applied to ships, three assumptions can be made which are based on the assumptions made in the molecular collision theory (Altan, 2019):

- All ship paths are straight lines through the observed area
- All ships have rectangular shapes
- Only two ships are involved in the collision

These assumptions could help to define possible collisions. Later the definition of maritime collision using the collision diameter was further expanded on by Pedersen (1995) by introducing a general formula that can be used being a function of ship length overall (LOA), ship beam, the velocities of the ships and the meeting angles of the ships. Friis-Hansen et al. (2007) defined the ship collision diameter as: “The maximum possible length of the projection of ship dimensions on to a line which is perpendicular to relative velocity of the ships at the contact condition”. Altan (2019) further developed the collision diameter approach using the case of the Strait of Istanbul to include the difference between a ship heading (HDG) and its course over ground (COG) where there were differences in determined potential collisions in high cross-current sections of the Strait of Istanbul.

2.1.2 Ship domain

The other method often used to calculate ship encounter risk and identifying near misses, is the ship domain method. Ship domain is a generally much used method in maritime risk research. It is known as a generalisation of a safe distance around a ship. This safe area is not the same in all directions and therefore asymmetrical (Szlapczynski & Szlapczynska, 2016). According to Szlapczynski and Szlapczynska (2016), the term “ship domain” is a widely used term which is not properly defined causing different definitions of the ship domain over time. The method, originated from Fuji and Tanaka (1971) and Goodwin (1975) and was first introduced using statistically processed radar data. The ship domain was an ellipsoid generated from data with an empirical approach. Later, AIS data was used for more advanced statistical ship domain methods.

Most ship domain methods do share a similar type of rule in their approach, which states that if the ship domain is crossed in some way, we cannot speak of a safe situation anymore. Fuji and Tanaka (1971) state that the ship domain: “…may be considered as the area of evasion”. If this area would be

crossed in some way, we can speak of a ship domain violation. An important note to make is that these ship domain violations are not always interpreted the same and can be defined in different ways:

- The own ship domain should not be violated by a target ship
- A target ship domain should not be violated by the own ship
- Neither of the domains should be violated
- The ship domains of the own ship and target ship should not overlap.

AIS records and ship locations will be analysed from the own ship view for every record, meaning that every record will be used as the own ship once, looking for certain violations with target ships. As mentioned above, these violations will be based on the definition of the violation. The size of the domains and therefore the minimum distance kept will differ based on the definition chosen (Szlapczynski & Szlapczynska, 2016). A ship domain is not the same for every ship, ships that travel faster manoeuvre less easy should have a larger safety area than small, more manoeuvrable ships. Therefore, a ship domain is often generated using various factors that influence the safety area of a certain ship. The usage of certain factors differs per performed research, but much used factors are the length of the ship, the speed of the ship and the manoeuvrability of the ships. Aside from ship-related factors, other factors like the type of encounter, traffic conditions weather conditions and even a human factor can be used to define the domain of a ship in a certain situation (Szlapczynski & Szlapczynska, 2016).

Over time, three different types of ship domain models have been formulated and have been the most used types of models. These different types are: Empirical models, analytical models, and knowledge-based models. Empirical models are based on analyses of a large set of data and data-based models. The models base the domain on locations of surrounding ships paired with other additional parameters. Analytical models are based on different parameters and not on past collected data of ships. They are models that differ per situation and are made to fit a certain scenario such as a channel crossing. They mostly have ellipses or off-centred circles and many times; researchers evolve past models with new findings. Finally, there is the knowledge-based model type. This model type uses navigators' knowledge often paired with neural networks to base their model around. Often also off-centred circles like analytical models, the domain created by this type of modelling is unique to the fact that it takes knowledge of navigators into account and thus being able to deviate from the limitations of empirical and analytical models (Szlapczynski & Szlapczynska, 2016).

The research by Szlapczynski and Szlapczynska (2016) gave an in-depth analysis on past research and the paradigm of the ship domain literature. This research will follow the empirical modelling type path as it will try to expand on past research performed by Altan and Meijers (2019). Altan and Meijers (2019) created an empirical ship domain in the Strait of Istanbul using a statistical calculation. The ship domain method created in this research will form the basis of the further analyses in this research and is therefore an important factor in this research. The reason for the choice to build on the research of Altan and Meijers (2019) in this way is that their empirical way of computing ship domain allows for computation of different ship domain contours for different situations. Figure 2 gives an example of two different ship domain contours based on Length Over-All (LOA)

This flexibility of ship domain generation makes it possible to model different situations which can then be modelled and analysed further. As the goal of this research is to identify ship collision risk, model it and analyse these results in the Strait of Istanbul, using a flexible way of determining important distance-based parameters is vital to a good quantitative way of finding as much valuable information during the analysis stage of the research.

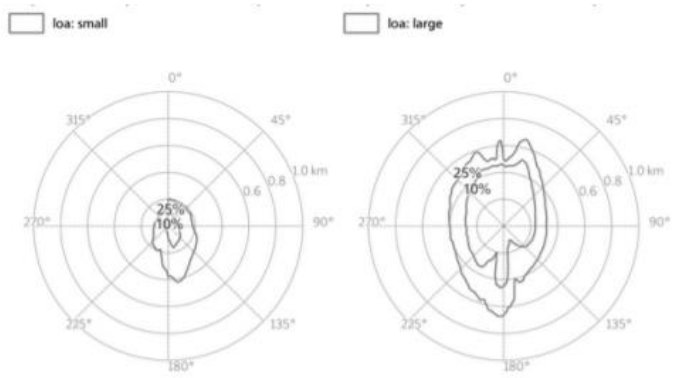

Figure 2: different ship domain contours based on different situations in an empirical-based ship domain model (Altan & Meijers, 2019)

2.2 Indicators of collision risk

Ship domain is considered an important player in ship risk detection. Ship domain violations are, however, not entirely reliable as an indicator for risk. Empirical ship domain models are created based on the locations of the target ships around the own ship, creating a domain based on the average distance kept from this own ship. In the instance of a wide waterway, ships have more space to keep distance from each other and as these ships are keeping more distance than the minimal safe distance, the ship domain could be larger than the actual safety area. Because of this, the empirical ship domain it cannot be used to directly derive ship collision risk (Szlapczynski & Szlapczynska, 2017), there is need for additional parameters.

So, as the risk cannot be directly derived from ship domain violations, it is important to do in-depth on possible other factors which could indicate risk of collision in a maritime environment. This section will give an overview of risk indicators previously found in maritime collision risk research and will finally contain a review on these used methods and the possible applications on this research. This section will not only contain non-ship domain-based parameters but also ship domain-related parameters which can enhance the ship risk identification.

Balmat et al. (2009) attempted to make a unified ‘fuzzy’ ship risk factor containing different static and dynamic parameters that help steer into a unified risk identification factor. Balmat et al. (2009) includes factors like, the history of the ship, gross tonnage, ship age and ship type for static parameters and sea state, wind speed, visibility, and time of day (night or day) as dynamic parameters. Enhanced versions of this make up the risk factors, starting with separate dynamic and static risk factors, and combining these to make the ‘fuzzy’ risk factor. This method presented by Balmat et al. (2009) is however, more focussed on individual ship risk than ship-to-ship collision risk. Still this research is of interest as it presents some potentially relevant dynamic and static parameters which could also be of use in the development of a collision risk model. Moreover, the combination of static and dynamic factors using the ‘fuzzy’ approach is also a thing that should be considered.

Kim (2020) mentions different risk identification factors in a more knowledge-based approach using the perception of ship operators in ship encounter situations. Right away, distance and angle of approach are two factors that are deemed to be of importance. These two values, however, are not used regularly, the approach angle is based around the maximum collision bearing-risk angle (MCBRA), this is the angle at which the risk of collision is perceived highest by ship operators. The distance used in this research is the distance at which the collision risk begins to increase significantly (DCRBIS). These two factors are similar to a ship domain-based approach in the sense of them having different values based on angle of approach and distance, yet Kim (2020) argues that this method is a more accurate method as psychological pressure of ship operators is also included. This is however not a feasible goal for this research, yet it is of importance to acknowledge the presence distance and approach identification methods.

Qu et al. (2011) assesses ship collision risk in the Singapore Strait. The situation in the Singapore Strait is a very similar situation compared to the Strait of Istanbul. Qu et al. (2011) use the trajectory of entire encounters in analysing the potential risk of ships in the Strait and revolve their findings not around individual ships, but divide the Strait in sectors, which will in turn give information on which risks might be present in the area. The main risk identification factors that were used in this case speed dispersion, acceleration, and ship domain. Speed dispersion is an index of the difference in average speed per sector in the strait. The acceleration and deceleration factor are an index of the change of speed over ground (SOG) per individual ship in each sector and the ship domain is a fuzzy domain created to look for ship domain violations. The findings of Qu et al. (2011) were that out of the three risk identification factors, most sectors which did violate these factors, violated the ship domain and the speed dispersion rules or even all three rules. This indicates that speed dispersion and ship domain violations can be of importance for identifying collision risk of vessels. Qu et al. (2011) divided the Strait into different sectors, analysing these instead of applying a method which could identify single clusters. The analysis of clusters of risk and their influencing factors is an important aim for this research so as this information is of importance, the feasibility should be highly considered in the creation of the model.

Altan and Meijers (2019) have analysed their previously mentioned dynamic ship domain contours in their research and found some factors that change the ship domain contour. Although this is the ship domain that will be drawn later in the methodology of this research, the findings of these contours can give some information on what factors change the size of the ship domain and thus the size of the 'safety area'. A ship domain is not an exact way of measuring ship collision risk, but an analysis can give insight into possible risk identifiers. Altan and Meijers (2019) have found different factors which cause the ship domain size to increase in their model. The time of day (day/night) was an influencing factor where the domain size grew in nighttime, indicating that ships kept more distance of each other at night thus getting closer can cause more risk. The encounter angle between the own ship and the target ship were also of importance for the domain size. In an overtaking encounter, the ships tend to keep less distance whereas head on and crossing scenarios cause a larger domain which can indicate the risk of certain approach angles. Altan and Meijers (2019) also found velocity to cause a higher relative distance kept between vessels in both the own velocity and the relative velocity analyses. Lastly, as shown in figure 2 the length of a ship can also indicate a larger risk of collision.

In earlier research of Szlapczynski and Szlapczynska (2016), ship collision risk parameters were analysed. They found that previous research often used distance and time as important factors to indicate collision risk. These parameters were defined as the distance at closest point of approach (DCPA) and the time to the closest point of approach (TCPA). The minimum distance between the two ships and time it takes to reach that minimum distance form a firm base to identify risk. However, Szlapczynski and Szlapczynska (2016) argue that these are not sufficient for proper risk estimation and present two newly formed parameters: degree of domain violation (DDV) and time to domain violation (TDV). The DDV expands on the DCPA by adding a ship domain in determining the risk factor. The DCPA only takes the distance into account and the DDV indicates the domain violation and the degree that the domain has been violated. The TDV is a similar addition to what the DDV is to the DCPA in the sense that it does not measure the time to the closest point, but the time to the domain violation instead. These parameters might be somewhat complicated to directly implement in the risk identification model, but it does indicate that the way in which the ship domain is violated plays an important role in the risk identification.

The section above is an in-depth overview of different ship risk identification methods that could prove to be of use in determining the ship risk identification model in this research. Table 1 presents a schematic overview of the different categories of risk identification factors that have been mentioned in the previously mentioned literature.

Table 1: Potential risk identifiers

Risk parameter	Source
Ship domain violations	Qu et al. (2011), Altan & Meijers (2019), Szlapczynski & Szlapczynska (2016)
Speed	Qu et al. (2011), Altan & Meijers (2019)
Ship dimensions	Balmat et al. (2009)
Weather conditions	Balmat et al. (2009)
Time of day	Balmat et al. (2009), Altan & Meijers (2019)
Distance	Kim (2020), Szlapczynski & Szlapczynska (2016)
Time of violation	Szlapczynski & Szlapczynska (2016)
Angle of approach	Kim (2020), Altan & Meijers (2019)

The goal of this section was to provide an overview of past research with different angles of approach in the literature to allow for a broad view in possible influencing factors. Whether all the mentioned factors will be applied in the risk identification model will later be discussed in the methodology section.

2.3 Knowledge gap

From the finding in this literature review and the past knowledge on ship risk identification and analysis in the Strait of Istanbul, the following scientific gaps have been found which will aim to be filled in this research:

- Combined research using the findings from different forms of literature to create a unified ship domain-based risk identification model. Szlapczynski & Szlapczynska (2016 & 2019) have worked on this in a way but have not included external factors like Balmat et al. (2009) and Altan and Meijers (2019) have mentioned in their research. Combining these could lead to more insight in possible risky scenarios.
- Altan and Meijers (2019) created ship domain contours based on different parameters like time of day or velocity. They have however not looked at a statistical and geographical aspect of the potential risks. Performing an analysis on the potential risk and testing what types of parameters cause clusters of risk could lead to additional insight in the risk of collisions in the Strait of Istanbul.
- Szlapczynski & Szlapczynska (2016) mention that the degree of a violation does matter in the determination of collision risk. However, as there is a degree of violation considered, it would seem logical to also be able to identify a degree of risk. Categorising and analysing collision risk could give insight into what types of risks occur in what circumstances.

3. Methodology:

This methodology chapter will give insight in the steps taken to output the results needed to answer the research questions presented in section 1.2. The chapter starts with an overview of the used data and how this data is prepared and stored. Following this, the method for the ship domain generation is explained. The next section will go into the risk identification model and how the variables are determined and standardised into a risk factor. Finally, the chapter will go into the forms of analysis that will be performed with the gathered collision risk factors.

3.1 Data

3.1.1 AIS

The data that will be used in this research will be data from the Automatic Identification System (AIS). AIS transmits information from ships to other ships and coastal stations. This data includes information on their position, heading, size and cargo (Harati-Mokhtari et al., 2007). Currently, this data is also gathered using satellite-based AIS, using satellites to monitor the ships movement on a global scale (Fournier et al., 2018). The system is based around ship-born transmitters and land-based receivers. These transmitters automatically send regular information to these transceivers. This transmitted information consists of three different types (IMO, 2015):

- Static information: basic, non-dynamic information, mostly about the ship's characteristics such as name and size, which is implemented on the installation of the AIS on the ship. Will only have to change if certain changes happen to the vessel.
- Dynamic information: automatically updated information from the ship sensors, mostly navigational information such as location, timestamps, heading and course over ground (COG).
- Voyage-related information: Information related to the specific voyage that has to be entered manually such as, draught, destination, and the presence of hazardous cargo.

Figure 3 presents an overview of the AIS system and which types of information is transmitted through the different channels.

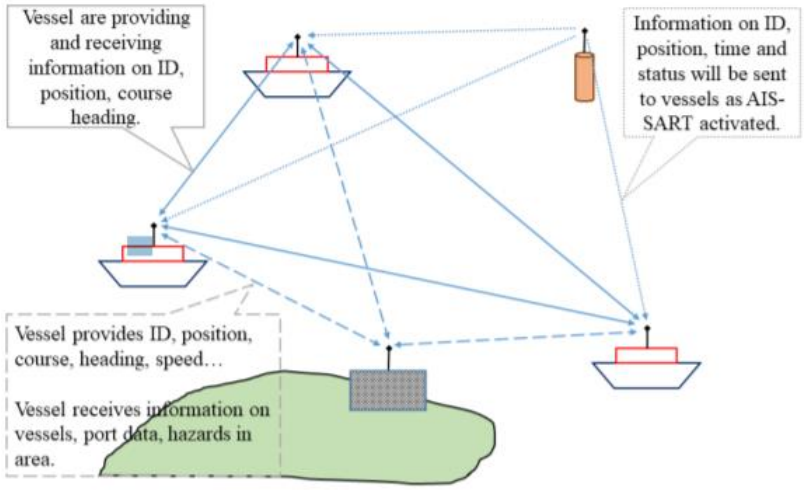

Figure 3**:** Overview of AIS (IMO, 2015)

AIS transponders are separated into two different classes, class A and class B transponders. Class A transponders are mandated under regulations for vessels over 300 tonnes and on all passenger ships regardless of size. They transmit information continuously and are equipped to automatically adjust its transmission to avoid interfering with other transponders. The system also shrinks the area of coverage to ensure the system is not overloaded in high density areas. Class B transponders can give smaller vessels the access to AIS systems with a transponder that transmits information every 30 seconds. They are also equipped with a system that checks for Class A transponders to ensure the AIS channels

are not overloaded. As most of the ships are equipped with a Class A transponder and Class B transponders also have a relatively short interval between timesteps, this data can be interpolated to create relatively accurate space-time trajectories.

3.1.2 Database

The AIS data is retrieved using a format from the National Marine Electronics Association (NMEA). The data are respectfully known as NMEA sentences. NMEA sentences are a specific data specification for communication between marine electronics. The data is transmitted from a data bus transmitting the data to the receivers and is printable in ASCII format (NMEA, n.d.). The Istanbul AIS data is delivered in this these NMEA sentences in ASCII format. This data provides little insight and is not yet workable. The NMEA sentences can be parsed into AIS messages which contain navigation related information. The ship data is stored in a database using PostgreSQL with the PostGIS extension by running a python parsing script into a newly generated SQL database. This stores the AIS messages into two tables. A table with information on the ship itself, known as static information, and a table with the navigational information, known as the dynamic information, as previously read in section 3.3.1.

The static information contains descriptive information on the ship, connected to a unique Maritime Mobile Ship Identity (MMSI) number (U.S. Coast Guard, n.d.) with which the ships information can be connected to an actual vessel. This static table also contains information on ships dimensions being the distance from the AIS system to the four-dimensional edges of the ships (bow, stern, port and starboard). It also contains voyage-related information (see 3.3.1). However, as this information has to be entered manually each voyage, which is often not done properly, this data is not very accurate.

The dynamic table contains location information on each of the ships in a certain timeframe, together with additional information. The main components are the MMSI number to identify the ship, a timestamp indicating the time of recording the message, and a latitude and longitude indicating the location of the message. Using PostGIS, this location information is converted to a point on the location in a coordinate system. The dynamic table also includes information on the course and heading of the ship which help to identify the movements of the ship.

3.1.3 Data enhancement

Once the tables have been made, there are some measures taken to remove errors from the dataset which could negatively manipulate the results of the analysis. The first measure is performed on the static table data. A script is run which orders and counts the ship types of each MMSI number ship, and only keeps the records which fall in the largest ship type group per MMSI number. In this way, for each of the ships, errors or changes in ship type are accounted for. For example, if a ship has 100 records for the ship type ‘cargo’ and 5 records for the ship type ‘passenger’, the passenger records will all be left out as it can skewer the data.

Another measure taken before analysis of the ship’s encounters is by altering the dynamic table data to exclude errors in that data. This mainly revolves around removing wrong locations, numbers, and illogical movements of ships. The latitude and longitude values are first checked whether the coordinates are located logically, if not, they are removed in the newly created table. This also done for the ship’s speed being a logical speed for a ship.

A final step that is taken before the encounter analysis is the definition of ship dimensions for each of the vessels in the analysis. With this step, a new table is created from the static table, with the ship dimensions together with some geometry aspects of the ship. First, the unique MMSI records are checked for different dimensions from the AIS transceiver to the bow, stern, port and starboard. If this differs for a set of records with the same MMSI number, only the ones with the majority of the same records are kept. Following this, the dimensions of the ships can be defined. First, the receiver point, and the centre of the ship are defined followed by the full dimensions including a three-point bow to represent the front of the ship. These geometries can then be visualised in a GIS.

3.2 Ship domain creation

3.2.1 Ship encounters: space-time frame

Before elaborating on the creation of the space-time frame, it is first important to define the way in which the ship encounters are identified. The calculations of ship encounters are performed by looking at the ships surrounding a single ship. This is performed for each ship and each record per ship. The ship which you look from for potential encounters, is known as the own ship, and the ships which surround that own ship, are known as the target ships. This means that in the calculation of encounters, each timestep for each ship, can have multiple target ships at that timestep. Figure 4 presents an example of own ship and target ship perspectives.

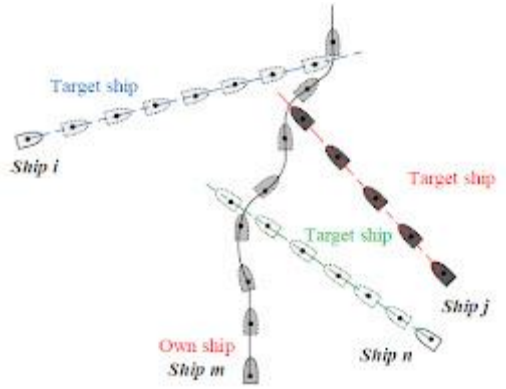

Figure 4: overview of own ship and target ships (Li et al., 2019)

To identify situations with risk of collision in the Strait of Istanbul, the data must first be manipulated in a way in which encounters of ships can be identified. The data in the current database consists of dynamic point data, which are now 2-dimensional points plotted with latitude and longitude information. The timesteps, which are included in each of the dynamic table records, need to be converted into a continuous space-time frame where encounters of ships within 2.5 km of each other can be identified.

To create this continuous space-time frame within the database, 2-dimensional movement of the ship should be defined as a continuous line. The AIS data initially consists of point records from which the own ship records are derived. However, it is possible for a target ship to not have a record at the exact timestep as the own ship. Therefore, the data needs to be plotted into a line so the timesteps can be derived from that line. As there is no exact information on a ship's location between two timesteps, the line will have to be interpolated. As most of the timesteps have an average length of 6 seconds and the script will only consider the steps with a maximum length of 44 seconds, there will not likely be a larger deviation in movement between two steps. Therefore, straight lines will be drawn between the timesteps as this is relatively accurate and a simple step to take. These single lines between two timesteps are called segments.

After these segments have been drawn, each of the records will be used as an own ship value once. The script will look for target segments at the same within a 5x5 km square around the own ship. This leads to a table with pairs of own ships and target ship segments. This data is then enriched with the relative velocity between the ships, the type of encounter based on the angle of approach, and the interpolated position of the target ship based on the time of recording of the own ship.

3.2.2 Ship domain

Many studies, including Altan and Meijers (2019) use the ship domain as a method of defining the space around a ship which can be seen as safe. This method based on Altan and Meijers (2019) uses a different approach based on the own ship and target ship data. Where many previous studies like Fuji and Tanaka (1971) draw an ellipsoid around the ship based on certain parameters. This study uses the direct location of target ships to define a ship domain in different scenarios.

The table of own ships and target ships, explained in the previous paragraph, forms the basis of this ship domain creation. The geometries of the ships made based on the static table are rotated to a local coordinate system where the own ships are the centre facing north with their course over ground. In this way, the relative positions of the own ships and target ships remain intact, and a web of target ships will surround the set of own ships around the centre of the local coordinate system. Altan and Meijers (2019) base ship domain on a percentage of total ship encounters in the 5x5 km area. The total amount of ships surrounding the own ship points would account for 100% of the total distribution. Thresholding this distribution will result in a certain ship domain contour based on the density of ships in a certain location.

This calculation is performed by rasterising the polygons with a 1x1 m resolution. 360 finite rays are then cast with a certain maximum distance of 1 km. The 360 finite rays will have an angular resolution of 1 degree, meaning 1 ray will be cast each of the 360 degrees. These rays will be intersected with the raster so the lines will contain the information of the raster. A cumulative distribution computation follows which can then describe the space around the ship. If 10% of the total records are used to draw the ship domain, then a point is drawn on each ray if it reaches 10% of the total raster records counted on the line from the centre. The 360 drawn '10% points' can then be drawn into a polygon using ArcGIS resulting in the 10% ship domain.

3.3 Risk detection model

Once the ship domain is generated, the risk detection model generation process can be started. As this research aims to categorise and analyse the identified risks with a geographical aspect, the model should output a risk factor with a geographical aspect per risk encounter. The International Maritime Organisation (IMO) have previously performed a formal safety assessment (IMO, 2018) where risk has been calculated and categorised in a matrix-based way. Each of the risk factors were categorised from minor to catastrophic. Low frequency and low consequences will lead to high risk and high frequency and high consequences lead to high risk. These matrix values can be calculated by dividing all the parameters' (whichever fits best) into 6 categories. The matrix will then have a 6x6 size with each of the input values having a place on the matrix and thus having an individual risk factor. This factor can be calculated by multiplying the two risk identification factors. This will give a 'risk score' between 1 and 36. If these values are divided again between the categories 1 through 6, the risk factor is divided into the risk categories (IMO, 2018). This method of categorising and dividing the factors into matrixes will be applied in this risk detection model. This method will output risk scores for each of the encounters on which an analysis can be performed looking for certain connections between ship parameters and the risk factor.

3.3.1 Potential parameters

The literature review has given an overview of potential risk identification parameters (table 1) which could be used in the risk detection model. Because this research can only be of a certain capacity, not all the parameters can be used in the risk modelling of this research. Not all parameters are feasible to analyse in this short term and the available data would also not allow for all of the parameters to be analysed. A selection will have to be made create the best fit model for the research. Therefore, some values will be added or removed based on model results. This will be a small iterative process where the models will be compared to other versions of the model and findings of other risk related research in the Strait of Istanbul (Altan, 2017). Table 2 presents the previously gathered potential parameters together with a degree of useability for this research. The most suitable parameters will be included in

the research and the less useable will be used to further modify the model if needed. Choices made for these degrees of useability will be discussed below.

Table 2: Useability of potential risk identifiers

Risk parameter	Useability for risk identification
Ship domain violations	Very useful
Speed	Potentially useful
Ship dimensions	Not very useful
Weather conditions	Requires a lot of data compared to effect
Time of day	Not very useful
Distance	Very useful
Time of violation	Very useful
Angle of approach	Difficult to implement

The parameters ship domain violations, distance between ships and the time of the violation are three parameters that are very important for risk identification (Qu et al., 2011, Altan & Meijers, 2019, Szlapczynski & Szlapczynska (2016). These parameters have been mentioned in most of the ship collision-based literature in some way. Therefore, will these be used in the determination of the risk factor in the initial model. The speed, ship dimensions and time of day are three factors which could be added in the iterative process depending on the findings of the initial model. Speed as a parameter is also mentioned in different ways and can therefore be of importance, yet the time of violation and the distance to the ship can also indicate some form of speed. This could cause these parameters to strengthen each other in the analysis resulting in skewed results. The time of day and ship dimension parameters are not very useful because these are very different parameters compared to the earlier mentioned parameters. These could, however, be used later as a potential relation to ship collision risk in the analysis. The angle of approach would require the data to be manipulated in a way in which trajectories would have to be made for each of the encounters. The choice was made to not perform this type of analysis as this would result in too much work which can decrease the quality of the overall results. Finally, the weather conditions could possibly be used in the later analysis stages of the research, but do not fit as well for the risk identification model. Together with that, it would require a new set of weather data that fits the times of the data of this research.

3.3.2 Ship domain violations

Now that the parameters that will be used in the model have been determined, it is important to define how these parameters will be measured for each of the encounters and what is needed to achieve these values.

The matrix method of standardising and combining the risk factor values requires single values for each individual ship encounter. Ship encounters that are considered are encounters where a target ship enters the ship domain of the own ship. The recorded AIS messages within the ship domain will be used for determining the values of the parameters. If an encounter does not contain a domain violation, the encounter will be identified as ‘safe’ and will not have to be considered a potential risk.

Distance and time of violation

As the IMO (2018) have stated in their formal safety assessment, frequency and consequences are important factors in risk identification. In a ship domain encounter perspective, consequences can be higher when ships are closer to each other and the amount of time the ship is within the dangerous area could indicate the frequency of the risk. Therefore, the distance and time of violation parameters are defined as the minimum distance between the two ships during the domain violation and the duration of the domain violation. This is a more simplified approach of Szlapczynski & Szlapczynska (2016) as the complete DDV and TDV approach would be too complicated for this research. Instead, for the time component, the time to closest point of approach (TCPA) is used.

The distance parameter will be determined by using the previously made space-time segments and calculating the minimum distance between the own ship and the target ships' space-time segment. This will result in a distance value which can be used in the risk matrix calculation where a lower minimum distance indicates more risk.

The time component will be determined using the TCPA. Calculating this requires the time of the target ship entering the own ship's domain, and the time when the ships reach their minimum distance. These timestamps will be determined using the same space-time segments, if the targets' space-time segments are intersected with the ship domains, the interpolated times can be used to calculate the time of violation. Time of entering the domain T_e and the time minimum distance T_x will be subtracted leaving entry time to minimum distance T_v as a useable value. The time in seconds will be used in the matrix calculation where shorter entry time to minimum distance value will score higher on risk as this indicates a faster approach towards the minimum distance. The choice of the TCPA value over the total domain violation time was made as the total violation time cannot accurately implicate risky ship behaviour. A longer domain violation time (domain time) could be riskier but does not necessarily indicate this as it could be possible that a longer violation time is caused by ships travelling slowly and safely past each other. Paired with this is the thought that whenever ships have reached their minimum distance in the encounter, the time they travel away from each other will not impact the risk of collision between the vessels, thus rendering this time insignificant. Applying the time parameter together with the approach speed using TCPA is deemed as a better solution.

Speed

The speed of a travelling ship can be interpreted in different ways. In an encounter scenario, there are two ships which have their own velocity. Therefore, there is an interplay between these two velocities and would it not be wise to choose one of the velocities. An accurate way of determining this speed is by using the relative velocity between the ships. Silveira et al. (2014) defined the relative velocity between two ships as:

$$\sqrt{V_{own}{}^2 + V_{target}{}^2 - 2 \times V_{own} + V_{target} \times \cos\left(cog_{own} - cog_{target}\right)}$$

Here V is the average velocity of the own ship and target ship, and the COG is the average course over ground of the own and target ships. This will lead to relative velocity between the ships during the domain violation which can then be used in the matrix calculation where a higher relative velocity indicates more risk, and a lower velocity indicates a lower risk.

Probability

The matrix-based risk analysis, which is performed in this research consists of two parts which are weighed against each other to assess the risk of a situation. It weighs the consequences of the risk situation, against the probability of the situation (IMO, 2018). The distance, time of violation and the speed variables represent the consequences part of the analysis. These variables assess how risky the situation which has occurred is and if the chance of consequences is high. So, a high relative speed, a low minimum distance, and a low time between entering the ship domain and reaching the minimum distance indicate that the there is a high chance of risk consequences in the particular situation. This, however, does not always represent the actual risk of a location or situation. A risky situation can occur often in a certain location for instance, but this can be caused by a high frequency of vessels traversing this location. Therefore, there is a need for the including probability of the risky situation happening, which will give information on the chance of risk.

The basic thought is that if two ships meeting each other in the waterway, at a certain location in a certain timeframe always results in a risky situation, then the probability of risk is higher than if this only occurs once in 100 ship meetings at that location and time. The probability of risk is derived from this thought and uses the total amount of ships found in the initial search of target ships around the own ship, before filtering the ship domain violations. The probability is calculated by dividing 1 by the

total amount of ships around the own ship, in and around the time of the domain violation described above. Ships that the own ship encounters around the same time and location which do not cause a risky situation, can indicate that these ships do not cause a risky situation in that scenario. Meaning that the more ships found in that scenario, reduces the probability of this risky situation.

The way in which this is measured is done by taking the timeframe of the own ship and adding two minutes of extra time to both sides of the timeframe for extra search distance and measuring all possible target ships around the own ship within 2500m. The pairs of own ships and target ships will be counted per ship domain violation encounter, counting all unique target ship MMSI numbers. If we then divide 1 by the count of unique target ship MMSI numbers, the result will be the probability of this a risky situation happening as this calculation is only performed on the ship domain violation encounters so this 1 value represents the potential risky situation. This probability is then categorised to fit in the risk calculation matrix.

3.3.3 Division into ship encounter types

Before calculating the required variables for the risk score calculation, the data needs to be split into three different parts. Maritime vessel encounters are not always similar and therefore require different calculation parameters to correctly calculated the risk score in an empirical manner. This is due to the fact that ship encounters differ in the way in which they approach each other. Vessels who encounter each other head on, will behave differently than vessels overtaking each other (Zhang et al., 2016). Therefore, the encounters are divided into three encounter types: crossing, head on, and overtaking encounters. These are based difference in course over ground between the two ships (Chang, Hisao & Wang, 2014)

- Head on: $170° < \Delta COG < 190°$
- Overtaking: $0° < \Delta COG < 67.5°$ or $292.5° < \Delta COG < 360°$
- Crossing: ΔCOG outside of the above-mentioned ranges

The reason this is important is that ships in different types of encounters behave differently than in another encounter, thus requiring different thresholds to determine whether a situation is risky or not. Ships in an overtaking encounter for instance, keep less distance than ships in a crossing encounter as overtaking ships travel in the same direction thus not being likely to cross paths. Crossing encounters, however, do potentially cross each other's trajectory and therefore keep more distance to be considered 'safe' than ships in an overtaking encounter. For this reason, the ship pairings will be given an encounter type value, separating the three types throughout the analysis. The ship domain violations will be determined by three different fitting ship domain contours which have been created by Altan and Meijers (2019) (figure 4). The encounters that are determined through these ship domain violations, are used to calculate the risk score variables. In this way, the fitting risk thresholds are determined per encounter type, resulting in more accurate results.

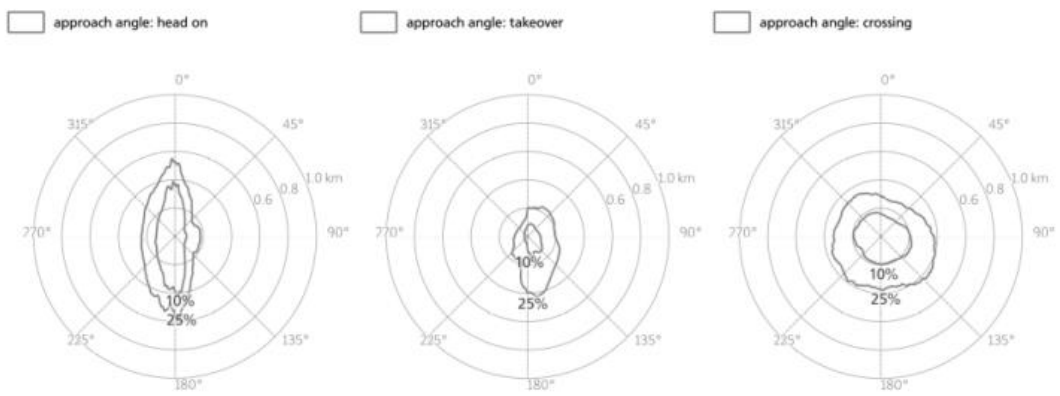

Figure 4**:** Ship domain contours per encounter type (Altan & Meijers, 2019)

3.3.4 Risk indexing

For the combination of the variables to be converted into a risk score, the risk will have to be indexed to combine the variables into a final score. The risk indexing will be based on risk evaluations of the International Maritime Organization (IMO) (2018). This method of risk indexing uses a matrix-based approach where the risk is categorised with a score of 1 to 6 (very low risk – very high risk) (IMO, 2018). The main matrix consists of a 'consequences' score between 1 and 6 and a 'probability' score between 1 and 6. These two are multiplied to a score between 1 and 36 and then divided again to standardise back to a 1-6 score. This will be the final risk score for the ship domain violation encounters in this research

The probability score is directly derived from the probability values, the consequences score is formed based on the three consequences variables mentioned in the previous paragraphs. The entry time to minimum distance variable, the minimum distance variable and the relative velocity variable are all categorised into the 1 to 6 categories before being combined with the product of the three to give each of the ship encounters a consequence score of 1 to 216. Before this consequence score can be combined with the probability, the score will have to be standardised back to the 1-6 categories. So, the consequence score is divided by 36 which leaves a 1-6 score for both the consequences and the probability scores to be put into the final risk matrix calculation.

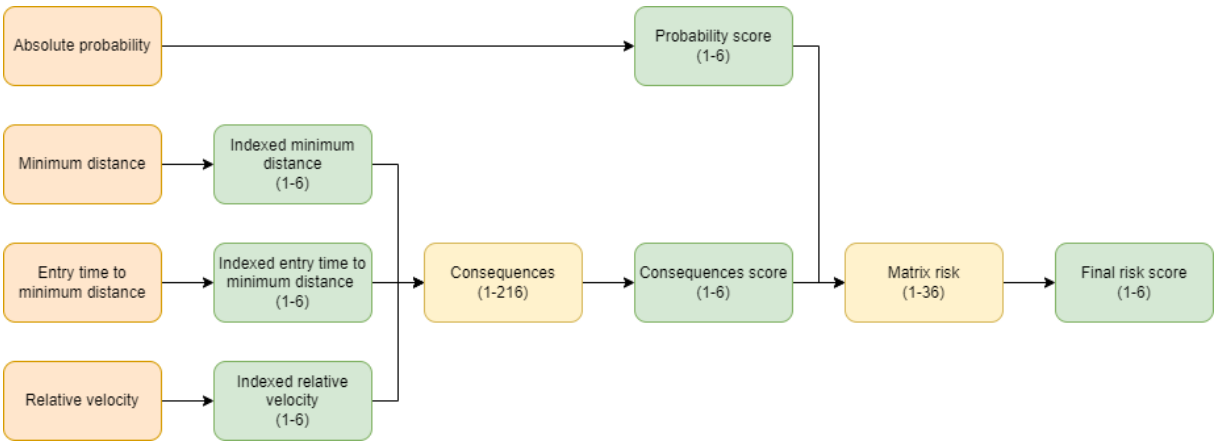

Figure 5: Schematic overview of risk indexing

3.3.5 Categorisation

The previous paragraph mentioned the indexing of the risk scores and the categorisation of variables into risk categories. The values of 4 variables: the entry time to minimum distance, minimum distance, relative velocity, and probability will have to be categorised into 6 categories ranging from very low risk to very high risk (IMO, 2018). As the spread of the values differ along the 4 different variables and the 3 different encounter types, there is a need for a categorisation method which accounts for this differing spread in values. Normally, a quantile or equal interval categorisation can be used to categorise values. However, these methods always cut of categories based on certain thresholds and this risk analysis requires cut-off points which are more based around the spread of the values. Only the high risk set of values should be categorised as category 6 (high-risk) and outliers which are not as high risk, should not be included in this high-risk category. Instead of a quantile or an equal interval categorisation, the categorisation is performed using a K-means clustering approach. K-means clustering (Lloyd, 1982) (Macqueen, 1967) aims to divide *m* observations into *n* clusters which are based on clusters and the surrounding data with the nearest cluster centre (or mean value). K means clustering is useful as it allows for a custom number of clusters to be found (in this case 6) and it can find accurate clusters is a large dataset. K-means clustering is not a perfect method however, the custom range of *n* clusters could yield poor results and the nearest distance methodology could also cause some clusters to not be completely accurate. Nevertheless, K-means clustering is a better alternative than using simpler categorisation methods, so it is taken as a good way to categorise the initial variables into the 1-6 risk indexes.

3.4 Analysis

The previous paragraphs have gone in depth into the steps taken up to creating the risk identification model. The following step in this research is to analyse the produced results and see as to what extent these results can be mapped, analysed, and validated.

Calculating the risk scores

Once the consequences variables and the probability values have been calculated and categorised into the six categories, the total risk score can be calculated. The way this is calculated follows the path of the risk matrix methodology used by the IMO (2018). A score of 1 to 6 is required for both the consequences and the probability, these two values can then be used to derive the actual risk score. As mentioned above, the consequences variables each have a score of 1 to 6 and will therefore have to be combined beforehand. This is done by taking the product of the scores for each encounter. This results in number of each ship violation encounter between 1 and 216 (1x1x1 up to 6x6x6). Dividing this total number by 36 results in an average risk consequence score of each of the ship domain violation encounters between 1 and 6. These scores can then be multiplied by the representative probability of the ship domain violation encounter, resulting in a number between 1 to 36. Dividing this number by 6 again will result in the final risk score of the encounter. These scores can then be analysed for potential patterns. Due to the nature of the matrix calculation, a score of 6 in the final risk score is low and many of the ship domain violation encounters will not score high in the risk scoring. Though this makes sense, as not all of the domain violation encounters have a high risk of collision, it is important to determine what is deemed as high-risk. Encounters scoring a risk score of more than 3, are seen as high-risk, this is a relatively low chance and indicates that the combination of the consequences score, and the probability score is high enough to have a risk of collision worth mentioning.

Hot spot analysis

With the risk scores established, these scores can be analysed to find certain patterns in the results. An important factor that can be analysed is location of the encounters. Patterns in these locations could indicate places and sections in the Strait of Istanbul which are risky. A way in which these patterns can be found can help to interpret the risk scores.

For finding these patterns, a hot spot analysis is used. Finding significant locational clusters of high-risk score values can show where these risk values are high and whether it is possible to find a certain risky area in the Strait of Istanbul. The cluster analysis that is used in this research is an Optimised hot spot analysis. An optimised hot-spot analysis is used to find the most optimal parameters for the dataset. The hot spot analysis uses the Getis-Ord GI* statistic to find significant hot and cold spots in the data. This statistic will show where the high-risk scores are clustered in a hot spot.

High risk analysis

The hot spot analysis which will be performed will give clear insight in the locational characteristics of the high-risk ship encounters. Additionally, a qualitative analysis will be performed on the high-risk records to analyse other characteristics of these ship encounters. Firstly, the high-risk encounters will be further analysed based on their movement in the Strait of Istanbul. Movement lines of each of the ship pairings have been made in paragraph 3.2.1. Selecting the sections of the high-risk encounters is possible using a query. Visually analysing these lines will give more insight in the exact movements of the ship encounters and give a clearer view of the risk dispersion in the Strait of Istanbul. Secondly, as mentioned in the theoretical framework, additional characteristics are worth analysing after the risk scores have been established. The ship size, speed and rate of turn are three parameters which differ per ship and could show possible connections with the ship score. The high-risk analysis will qualitatively compare the above-mentioned parameters between the high-risk selection of data and other selections of the AIS dataset. This could give insight in a possible correlation between the risk and the parameters which could later be analysed statistically.

3.5 Conceptual model

The findings in the theoretical framework and the research methods presented in this chapter, have been presented a conceptual model which is shown in figure 5. This will give a schematic overview of the steps taken to answer the research question.

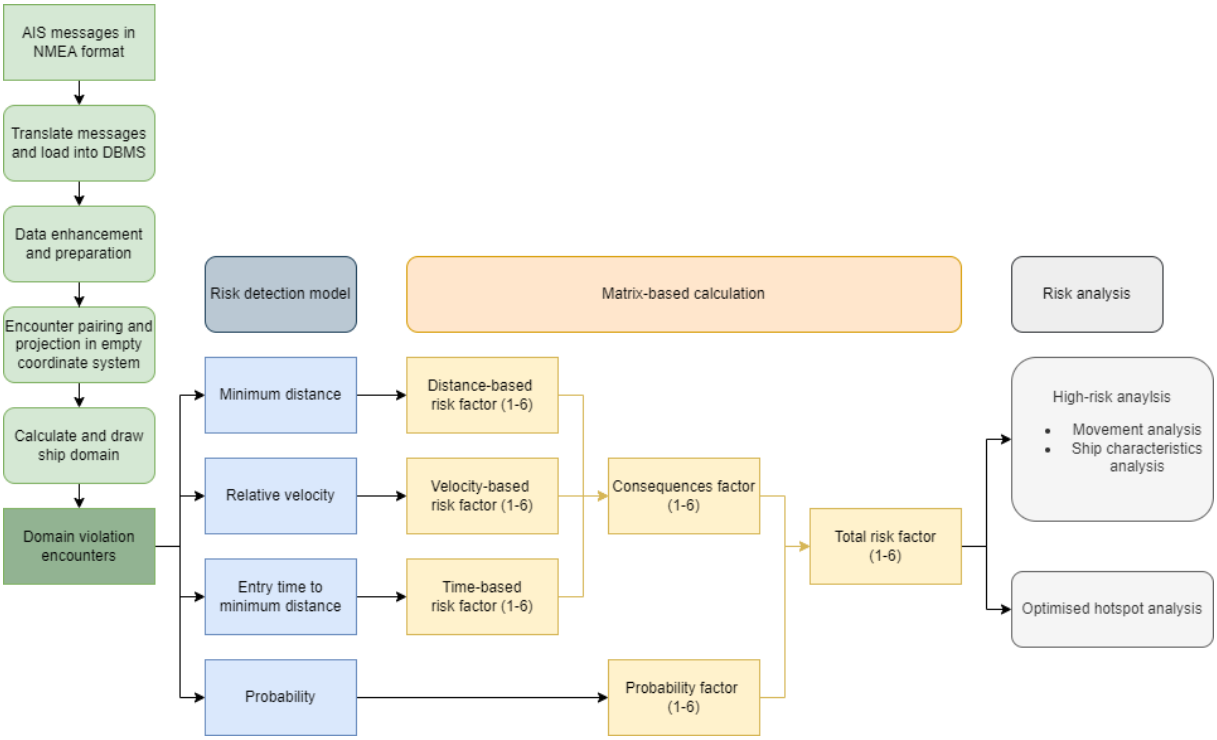

Figure 5: Conceptual model

4. Results

4.1 Ship domain

As explained in the methodology chapter, the ais data will be divided into three parts based on the different approach types that apply to the ship encounters. As the method of drawing the ship domain is an empirical one, the ship domain contours can also be drawn accordingly to the encounter type. Figure 6 shows the target ship's dimensions of the ship pairings that violate the ship domain contours.

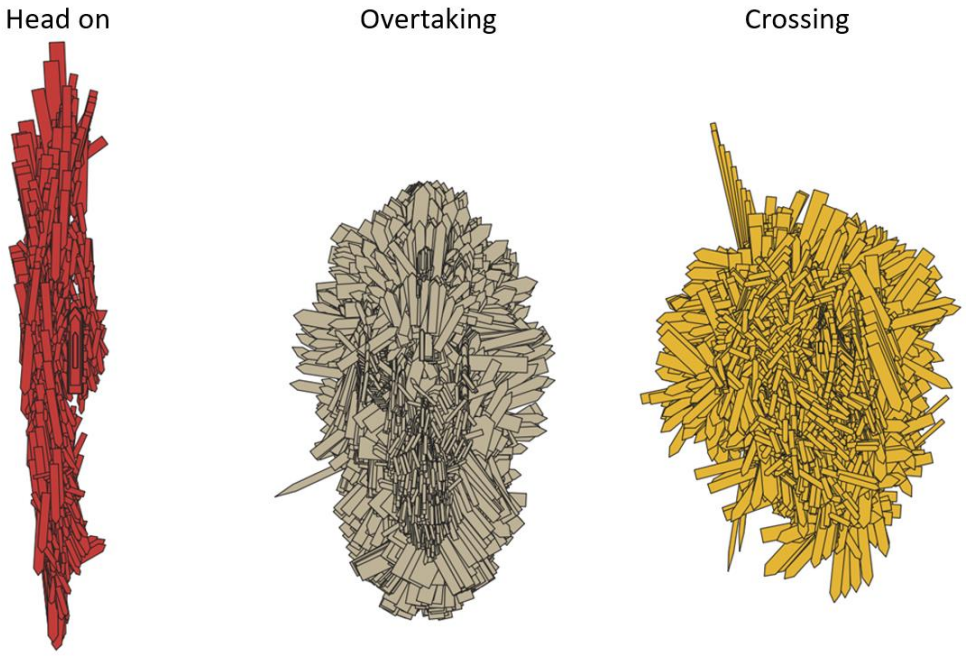

Figure 6: Target ships within ship domain contour per encounter type

The records shown in figure 6 are projected into a coordinate system where the own ship's centre is the 0.0 point of the system, resulting in the target ships surrounding the own ship values. All the records are overlayed to show the way in which the three encounter types differ. The head on contour records are much thinner than the overtaking or crossing encounters. The head on encounter contour also clearly shows that the target ships are all travelling in the opposite direction of the own ship. The overtaking and crossing encounters look more similar, the crossing records are more spread out than the overtaking records. This is due to the fact that the 10% ship domain contour contains the 10% closest records within the encounters. And as the overtaking encounters can keep a smaller distance to remain 'safe', the domain contour is also smaller. The difference in travel direction between the overtaking and crossing encounters is clearly seen as the target ships of the crossing encounters travel in very differing directions whereas the target ships of the overtaking encounters tend to travel in a similar direction as the own ship. It has to be noted that some target ships have a different angle than expected in the encounter type it is categorised in. This is because an encounter between two ships can have a differing relative approach angle. An entire encounter is categorised based on the majority of encounter angles, so, if the majority of own ship and target ship pairs are overtaking pairs, the entire encounter is classified as an overtaking encounter.

4.2 Risk identification model

The previous section shows the ship pairings being filtered to find the domain violation contours, which will be used to analyse the potential risk of these encounters. In this paragraph, some numbers on the results will be shown. These include the total number of datapoints from one month of AIS ship

data which are shown in table X. This data shows that there is a large number of total ship points which are filtered down along the steps of the model. First being reduced into the ship pairs set which are pairs of ships that are within 2500m of each other at a certain timeframe of the own ship. In the next step, these 10 million pairs are reduced to 148,059 ship domain violation pairs. The empirical ship domain contours derived from Altan and Meijers (2019) are based on a percentage of nearest ships. This means that the, in this case 5% nearest ships of the total ship pairs, should be inside the ship domain contour. In the case of this research, the total percentage of domain violations derived from the pairs is only around 1%. This is possibly caused by the categorisation into encounter types and calculating based on the different domain contours for each of the encounter types (see figure 6) Ship pairs that would possibly be included in the overtaking or crossing contour could not be included if they are classified as a head on encounter. Therefore, reducing the total percentage of ship domain violation pairs. The ship domain violation pairs are finally grouped to a total of 5903 ship domain violation encounters. There are relatively little head on encounters in the strait compared to the number of overtaking and crossing encounters. This can be explained by the case that there is a rule in the Strait of Istanbul where large ships can only travel in one way in the strait, and that large ships travelling in the other direction, will have to wait for the other ships to exit the strait. Because of this rule, there are very little head on encounters to be found. It can also be seen that the crossing encounters have the highest percentage of encounters relative to the number of domain violation pairs. This indicates that per encounter, crossing encounters have the least amount of ais records per encounter. This could be due to a higher approach speed or the nature of the approach angle but the exact cause of this cannot be confirmed.

Table 3: Total number of datapoints per encounter type

Type	Total AIS points	Total ship pairs	Total ship domain violation pairs	%dom	Total encounters	%enc
Head on	N/A	244235	1654	0.68%	61	3.69%
Overtaking	N/A	7966231	115887	1.45%	3930	3.39%
Crossing	N/A	2503095	30518	1.22%	1912	6.27%
Total	14496428	10731561	148059	1.38%	5903	3.99%

The final encounters have had the main risk calculation variables calculated which are shown in t**able** X. The domain time: the time between the target ship entering and exiting the own ship's domain is also included in this table. This is however to give further insight in the data, this variable is not used in the risk scoring process.

This table shows the relative differences between the mean values of the main variables. There are a few interesting things to be seen in this table. One thing that can be noticed is that the head on encounter type has the lowest average minimum distance. This is likely because of the narrow shape of the head on domain contour (figure 6). The overtaking encounter type has the highest mean minimum distance which could be because ships overtaking travel alongside each other for a certain time and therefore keep a higher average distance to avoid collision. The way in which overtaking ships travel also explains the low average relative velocity and higher domain time of this encounter type. As the ships travel alongside each other, it will take longer for the target ship to exit the ship domain on average. The domain time and the entry to minimum distance time is on average lower in crossing and head on encounters. The difference between these is relatively small but the crossing encounter type has a higher domain time, but a lower entry to minimum distance time compared to the head on encounters. This could indicate that on average, crossing encounters reach the minimum

distance quicker within the domain time frame than head on encounters, indicating a potentially riskier approach.

Table 4: Mean values of relevant variables per encounter type

Variable	Value	Head on	Overtake	Cross
Minimum distance	Mean	107.1	189.6	123.1
Relative velocity	Mean	16.8	2.2	15.4
Entry to min distance time	Mean	32.9	58.0	26.8
Domain time	Mean	50.4	99.4	55.9

The result output from the risk identification model is further shown in the section below. Here, the individual encounter types will be analysed mainly on their risk score and location in the Strait of Istanbul. Later in the research, these results will be statistically tested for potential correlation.

4.2.1 Head on encounters

The first encounter type that will be focussed on, is the head on type encounter. After the data enhancement phase, one week of data resulted in 1654 head on encounter ship pairings of which the target ship violated the own ship's domain. After grouping these records based on their encounter group (will further on be known as an encounter), a set of 61 encounters was left to analyse. All 61 encounters have been visualised in figure 7 based on the minimum distance location of the ship domain violation encounters.

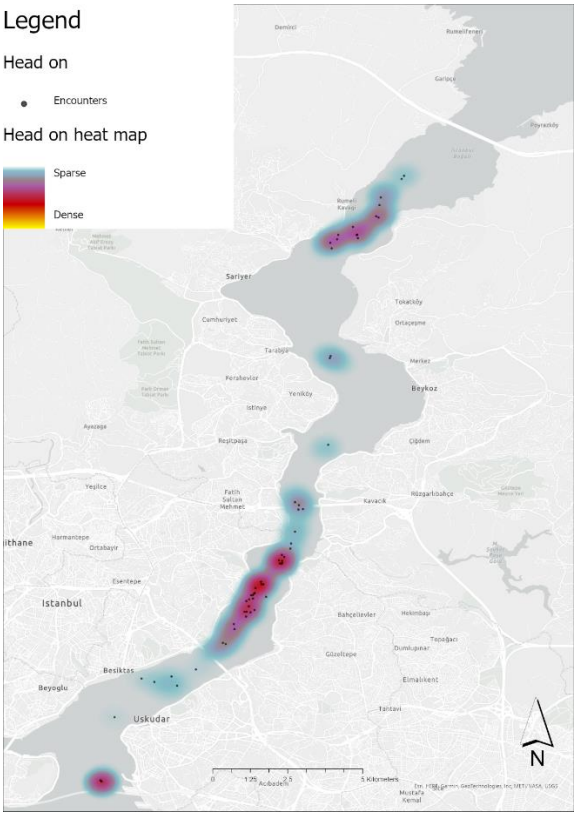

Figure 7: Heat map of head on encounters

The spread of the head on encounters is not very concentrated around one particular section of the strait. Most of the points are concentrated in the more straight and narrow sections which make sense as narrow sections have a higher chance of close encounters and straight sections will likely make the

approach either head on or overtaking. It has to be noted however, that the number of head on values is relatively small, so it is difficult to make hard assumptions out of these results.

Minimum distance

To be able to make assumptions out of this data, the risk scores will first have to be calculated, starting with the consequence variables minimum distance, relative approach time, and relative velocity. Starting with the minimum distance values, these values are derived from the grouped encounter pairings. Each of the encounter pairings have been grouped into one record in the dataset, containing the location of the minimum distance record. Figure 8 shows the minimum distance records for the head on encounters spread throughout the strait. As these minimum distance records are often within the middle of the ship encounter, the patterns of these records are similar to the not grouped dataset of 1654 records.

After the calculation of the minimum distance, the values have been categorised using a K-means clustering method, the results of this, are shown in figure 8 values have been categorised in values from 1 to 6, 1 being not risky, and 6 being risky. From this categorisation is assumed that records in category 6, can be seen as risky, and the lower values cannot immediately be seen as risky.

As the number of head on ship domain violation encounters is relatively low, it is more difficult to analyse than the other two encounter types. The minimum distance scores show high (cat 6) risk points in different sections of the strait and compared to the total dispersion of the head on encounters, there is little to be concluded from these.

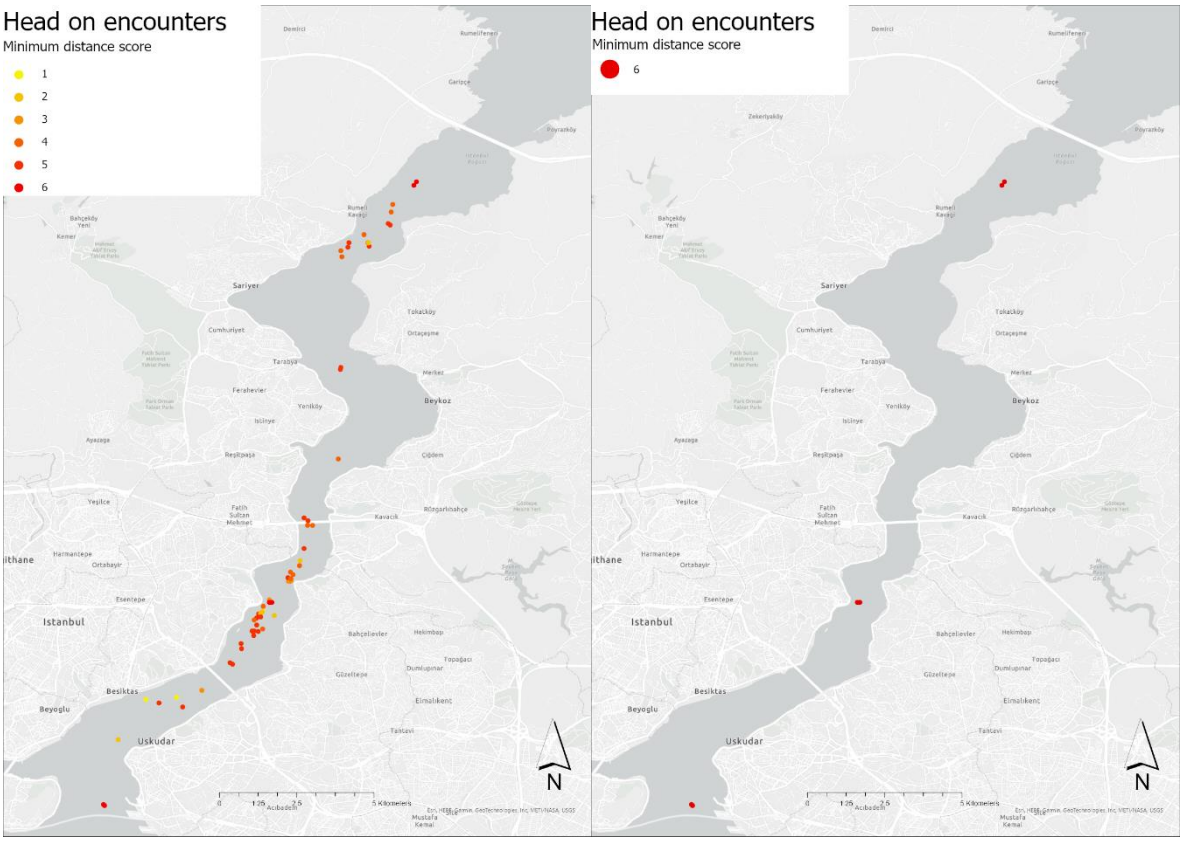

Figure 8: Minimum distance maps of head on encounters (left: all categories, right: category 6 only)

Relative velocity

Another to be calculated variable is the relative velocity of the two ships. In absolute values this is expected to be relatively high as head on approaches will not always have to match speeds, which is

more common in overtaking scenarios. Whether the risk of head on encounters increases in a certain location, we have to look at the findings from the K-means clustering of the values (Figure 9)**.** The location of the encounters is also based on the point of minimum distance of the encounters.

Figure 9: Relative velocity maps of head on encounters (left: all categories, right: category 6 only)

The categorised records in figure 9 show different results from the minimum distance points. The high-risk (cat 6) points are however, mostly centred in the higher point density areas when compared to the total head on encounter set. Therefore, there is no remarkable pattern in these scores aside from an even dispersion along the density.

Entry to minimum distance time

The third parameter focusses on the approach and time factor of the ship violation encounter. The time between the target ship entering the own ship domain and reaching the point of minimum distance is calculated. The assumption of the literature (Szlapczynski & Szlapczynska, 2016) states that a lower time of entry to minimum distance indicates a riskier situation. The results of this analysis for head on encounters is categorised using the K-means and is show in figure 10.

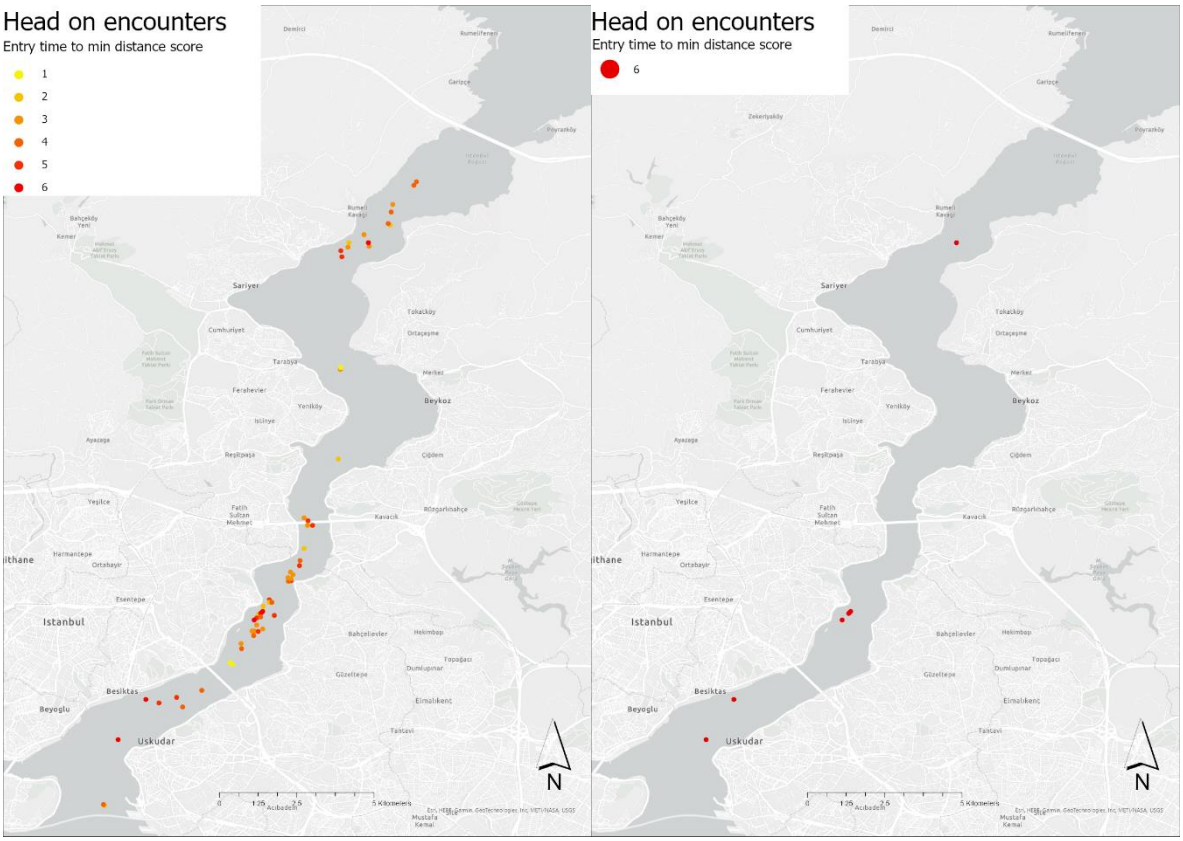

Figure 10: Entry time to minimum distance maps of head on encounters (left: all categories, right: category 6 only)

The results in figure 10 show that the highest risk category is highly represented. Compared to the minimum distance and relative velocity variables, this variable contains more high risk (cat 6) values. A small cluster can be seen in the middle of the strait which could indicate that ships have a short entry time to minimum distance here. Whether there is a significant relation to this location, will have to be determined further in the research.

Probability

The final calculated value that has to be included for the head on analysis, is the probability factor. This probability factor determines the probability of a ship pairing being a ship domain violation pairing. The ships around the own ship during the domain violation of a risky encounter are used to determine the probability of a domain violation occurring. If multiple ships do not violate that own ship's domain, it can be said that the probability of a domain violation happening in that situation is lower than if there are no other ships around that do not violate the domain of the own ship. The probabilities, categorised 1 to 6 using the K-means, are mapped in figure 11 below.

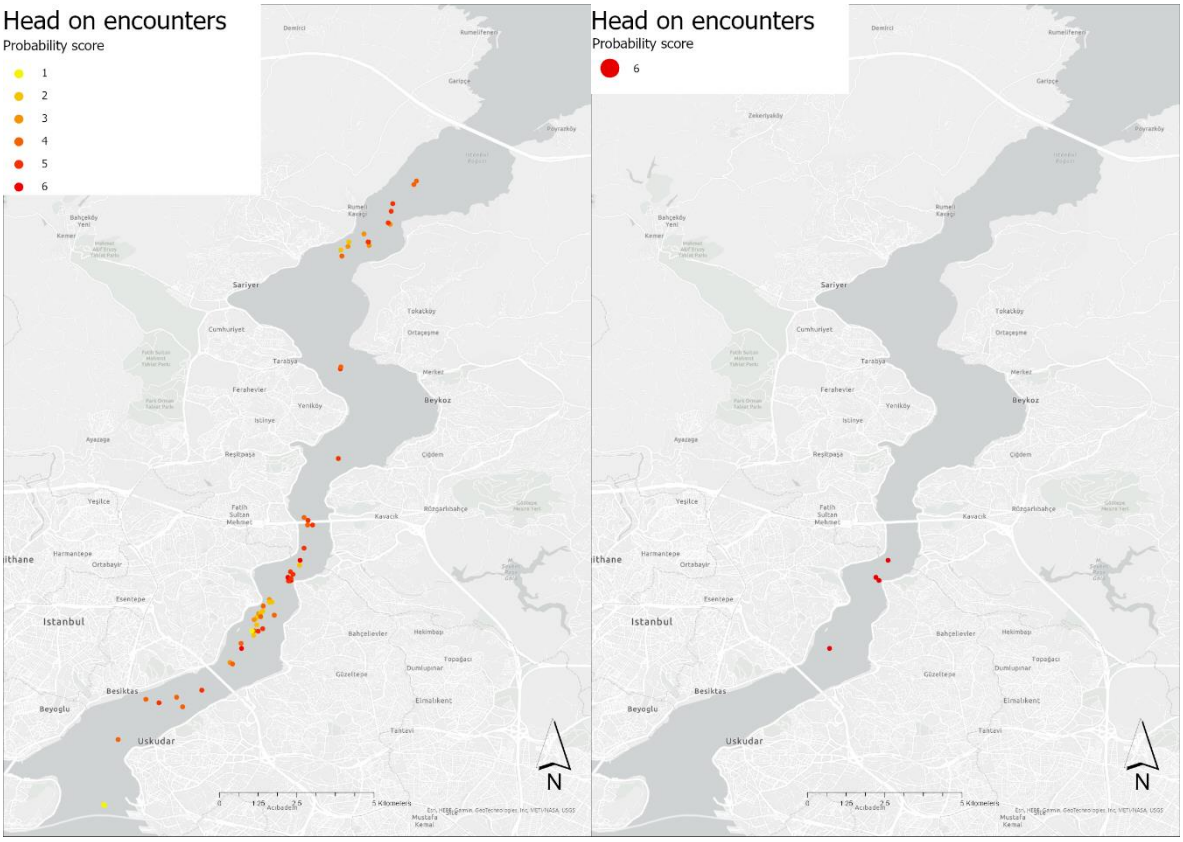

Figure 11: Probability maps of head on encounters (left: all categories, right: category 6 only)

Figure 11 shows that the high probability (cat 6) points are not highly represented in the head on encounters. Though lowly represented, the higher probability points are fairly clustered in the middle of the strait. This indicates that when ships meet here, there is higher chance of a head on ship domain violation occurring.

Risk score

The 3 variables mentioned above can be combined with the probability factor of each of the encounters to calculate the risk score (1-6).

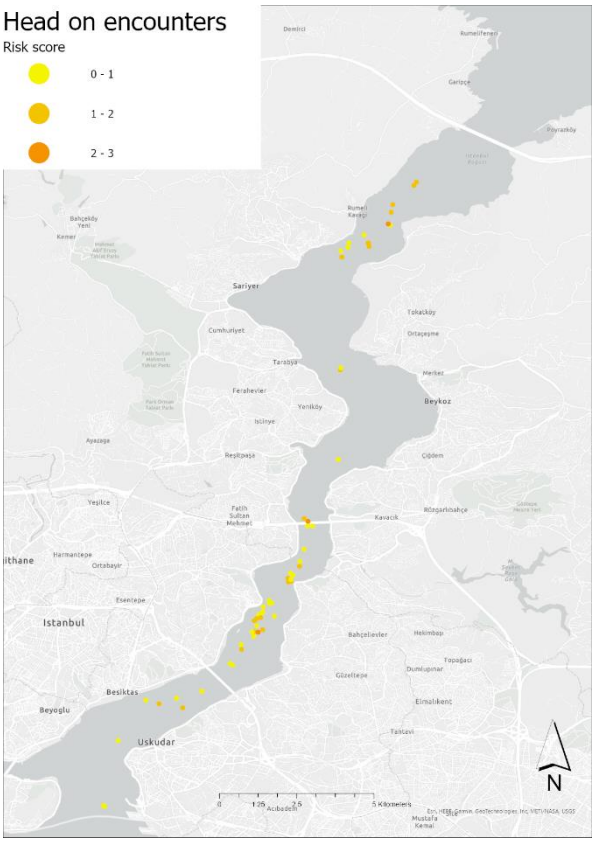

Figure 12: Risk score map of head on encounters

Figure 12 shows the risk score of the head on encounters. The first thing that can be seen is that the calculated risk score is relatively low for head on encounters. There are no risk scores higher than 3 which is understandable as there are only 61 head on ship domain violations in total. The probability that one of these encounters is very risky, is low. As there are encounters with a risk score of over 3. The head on encounters do not have any high-risk results and, as only the high-risk points are of importance, it is not needed to look at the results on this map.

4.2.2 Overtaking encounters

The second encounter type that will be covered is the overtaking encounter type. The selection of values resulted 115,887 usable records where the target ship violated the domain of the own ship. After grouping these values by encounters, 3930 encounters remained.

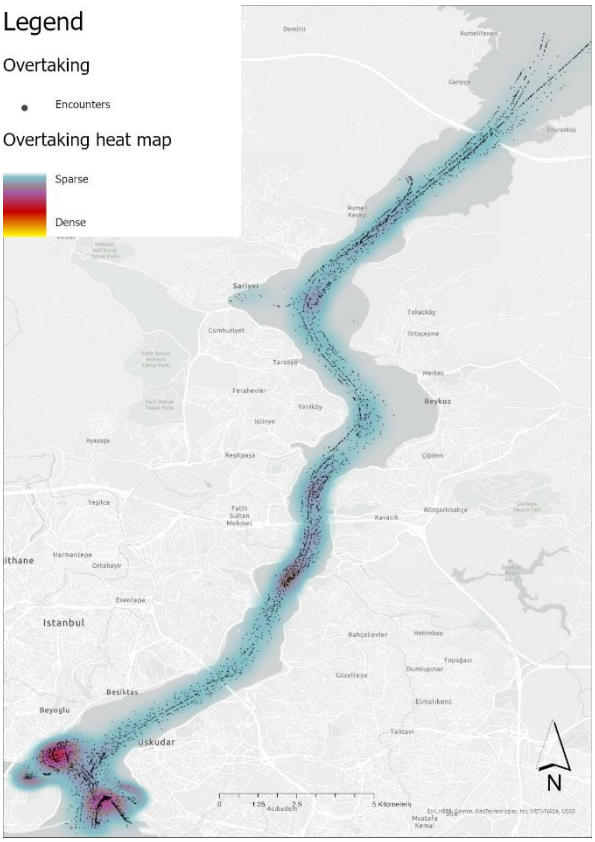

Figure 13: Heat map of overtaking encounters

Figure 13 shows the dispersion of the overtaking encounters in the Strait of Istanbul and a heatmap of the points in this map. This map shows that the clustering of ship domain violation encounters is present in different parts of the strait with the densest clusters in the bottom of the strait. Although, it stands out that the main corridor flow of the bottom of the strait is less dense in ship domain violation encounters. Most of the encounters occur near the docking areas in the bottom of the strait. Aside from the clusters in the bottom, the middle bridge section, a straight and narrow section of the Strait of Istanbul, also contains a cluster of overtaking ship domain violation encounters. These violation encounters potentially be risky encounters, whether these encounters are risky and where potentially risky encounters occur, will be discussed in the sections below.

Minimum distance

The minimum distance scores for the overtaking encounters are shown in figure 14. The mean minimum distance of all encounters is 189.6, which is the highest of the three encounter types. This means that the ships in a domain violation, keep a relatively larger distance from each other. This could be due to the shape of the ship domain, presented in figure 6. The domain is relatively large, and the ships pass each other parallelly, meaning the shorter sides of the ship will determine the minimum distance. The points with a very low minimum distance are concentrated around the bottom of the strait and near the middle of the strait. Comparing the minimum distance points of the overtaking data to the minimum distance points in the head on data, the main findings are that the number of points in the overtaking data is much larger. This makes the overtaking data easier and more reliable to analyse.

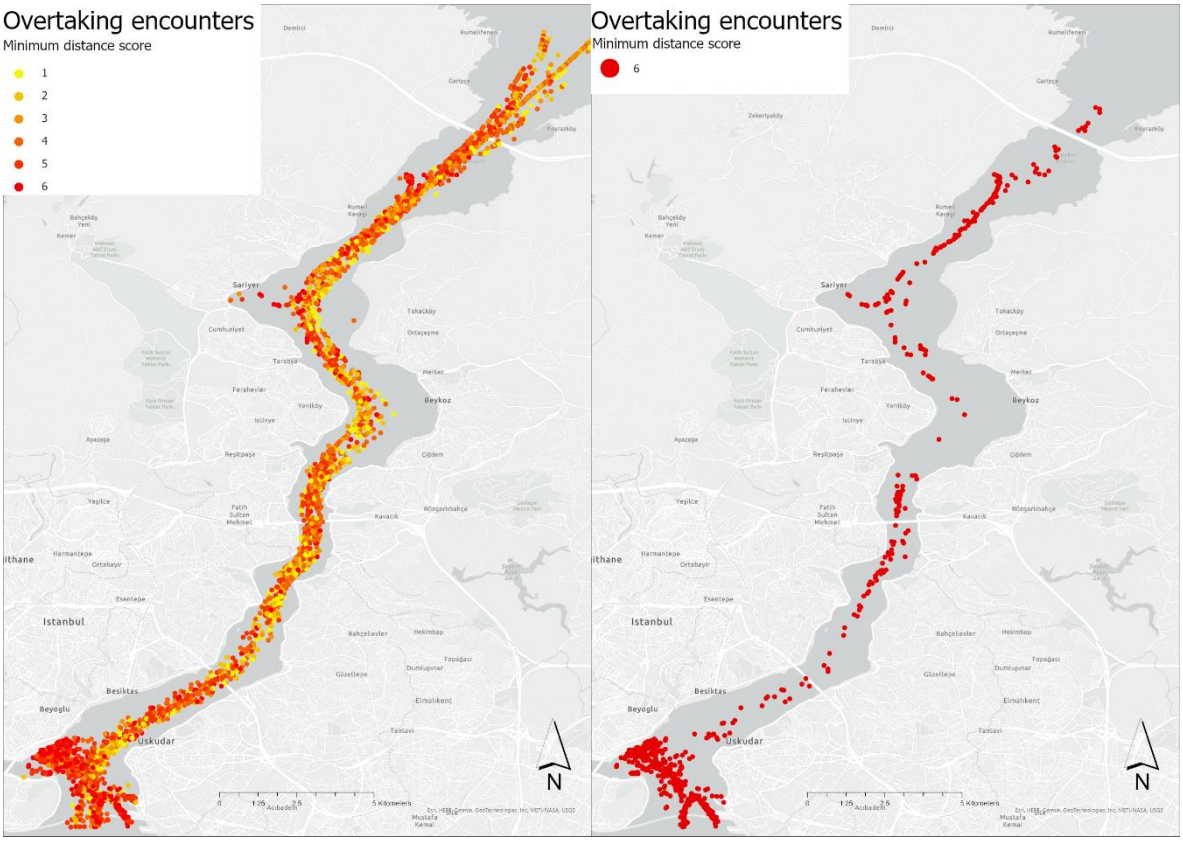

Figure 14: Minimum distance maps of overtaking encounters (left: all categories, right: category 6 only)

The points in figure 14 show that encounters with low minimum distance occur throughout the entire strait but that a large portion of these occur in the bottom of the strait. In fact, most of the points in the bottom of the strait have a minimum distance in the high-risk category (cat 6). This indicates that the bottom section of the strait results in more risky minimum distance scenarios than other sections of the strait. Despite this finding, it is not yet clear whether this is significantly riskier for ships.

Relative velocity

The relative velocity values of the overtaking encounters are shown in figure 15. The relative velocity values between overtaking ships are relatively low, which is understandable as ships overtake each other and tend not to have very differing speeds. Encounters with category 6 relative velocities are found throughout the strait and again mainly in the bottom though not as much as the minimum distance data (figure 14). Different sections in the middle of the strait also contain hotspots of higher relative velocity values. These could be explained by ships slowing down due to bends in the waterway. This can however not be confirmed.

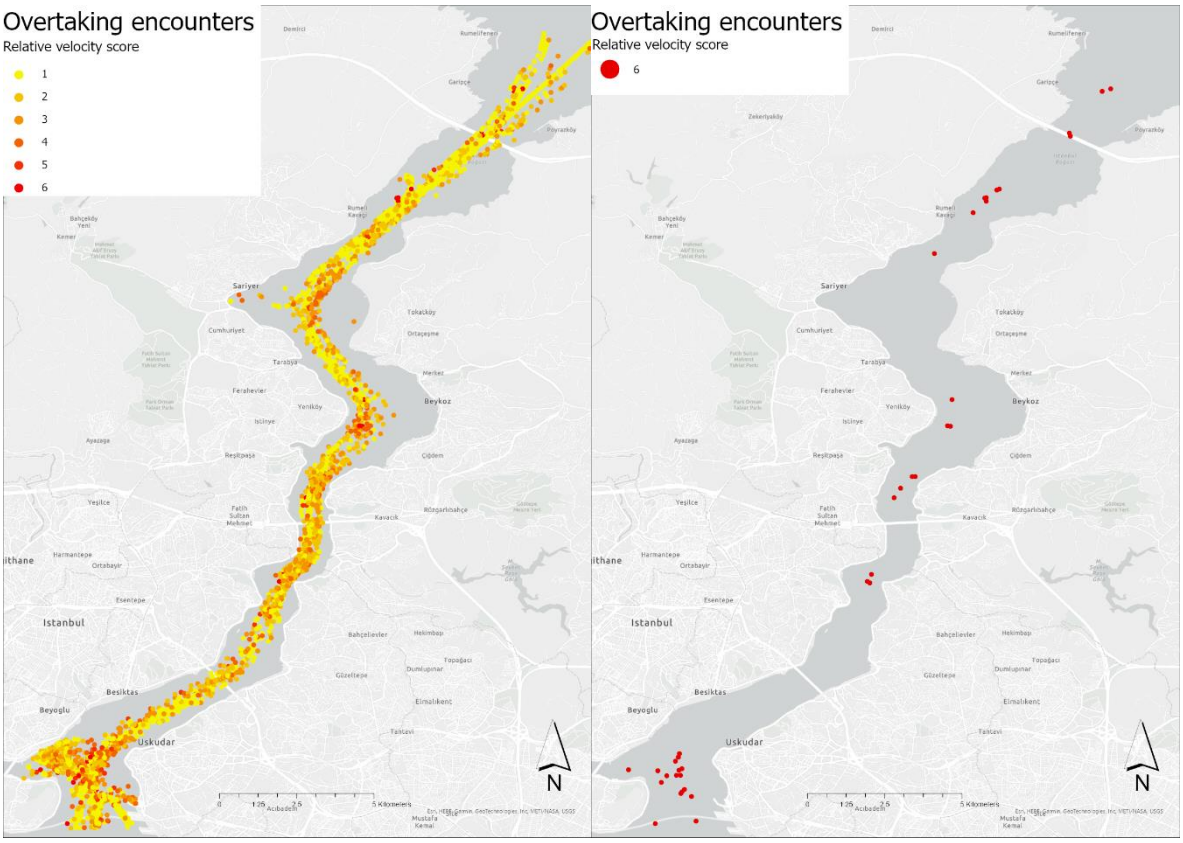

Figure 15: Relative velocity maps of overtaking encounters (left: all categories, right: category 6 only)

Entry to minimum distance

Looking at the entry time to minimum distance for overtaking encounters, the overtaking encounters have the highest mean time value of the three encounter types. This means that, on average, target ships in this encounter type category reach the minimum distance to the own ship slower upon entering the own ship's domain compared to the head on and crossing encounter types. This goes paired with the finding that the overtaking encounters also have the highest mean domain time, meaning that the time the target ships within the encounters are inside the own ship's domain, is higher compared to the other encounter types. These findings can also be explained by the nature and shape of the domain and the encounter type. An overtaking encounter with lower relative velocity is likely to have a longer duration than a head on or crossing encounter.

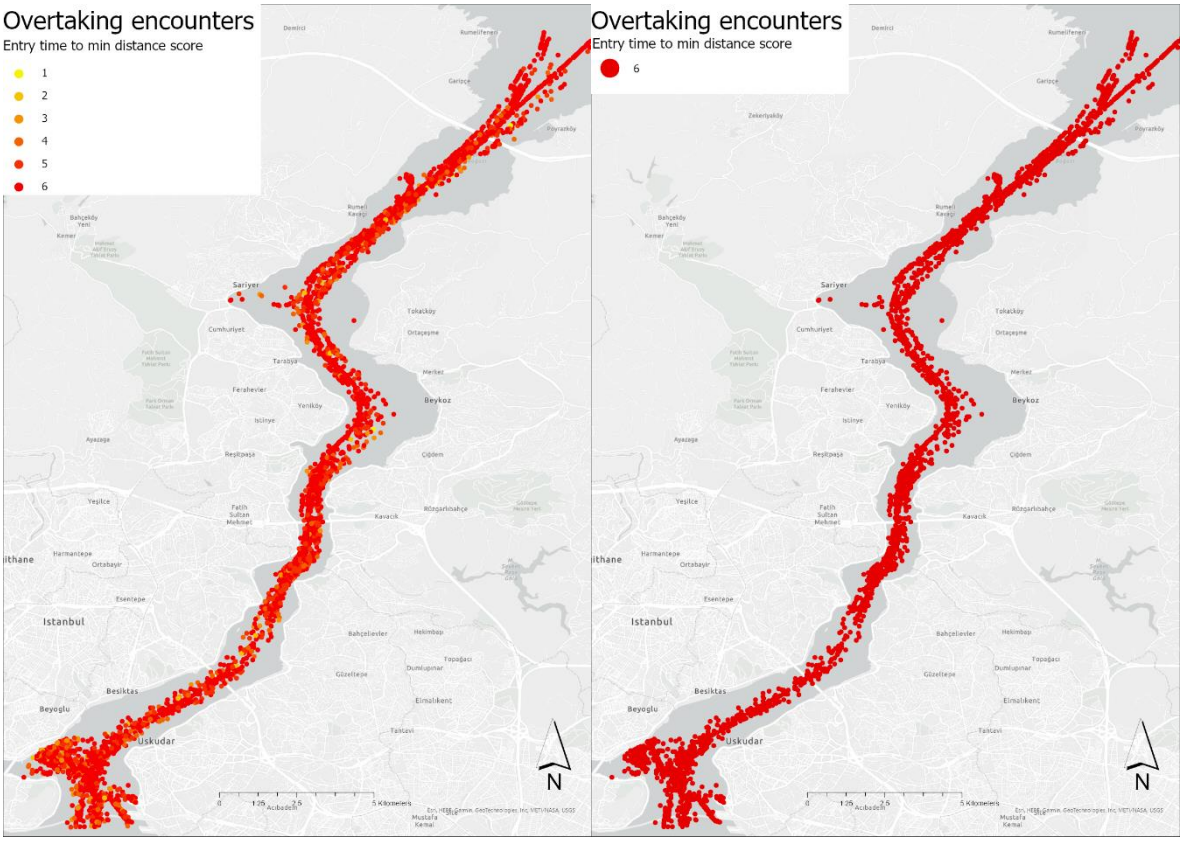

Figure 16: Entry time to minimum distance maps of overtaking encounters (left: all categories, right: category 6 only)

The entry to minimum distance score values (1-6) and the 6 score values are shown in figure 16. The entry to minimum distance variable has a high K-means clustering in at the 6 score value, which results in the following result that high percentage of the total domain violation encounters, get a high-risk score for this variable. This could indicate that most of the overtaking counters have a fast entry time to minimum distance and that values that do not have this, are particularly not risky.

Probability

The probability values of the overtaking encounters are shown in figure 17. The main finding of the dispersion of high probability values compared to the total set of overtaking encounters is that though there are many total points in the bottom of the strait, there are little high probability points in the bottom of the strait. High probability hotspots are found in the middle to top section of the strait. This means that, many of the high-risk values found in the consequences variables in the bottom of the strait, do not always occur and have a lower probability of occurring than the potentially risky sections in other parts of the strait. This probability will influence the total risk score of the overtaking encounters.

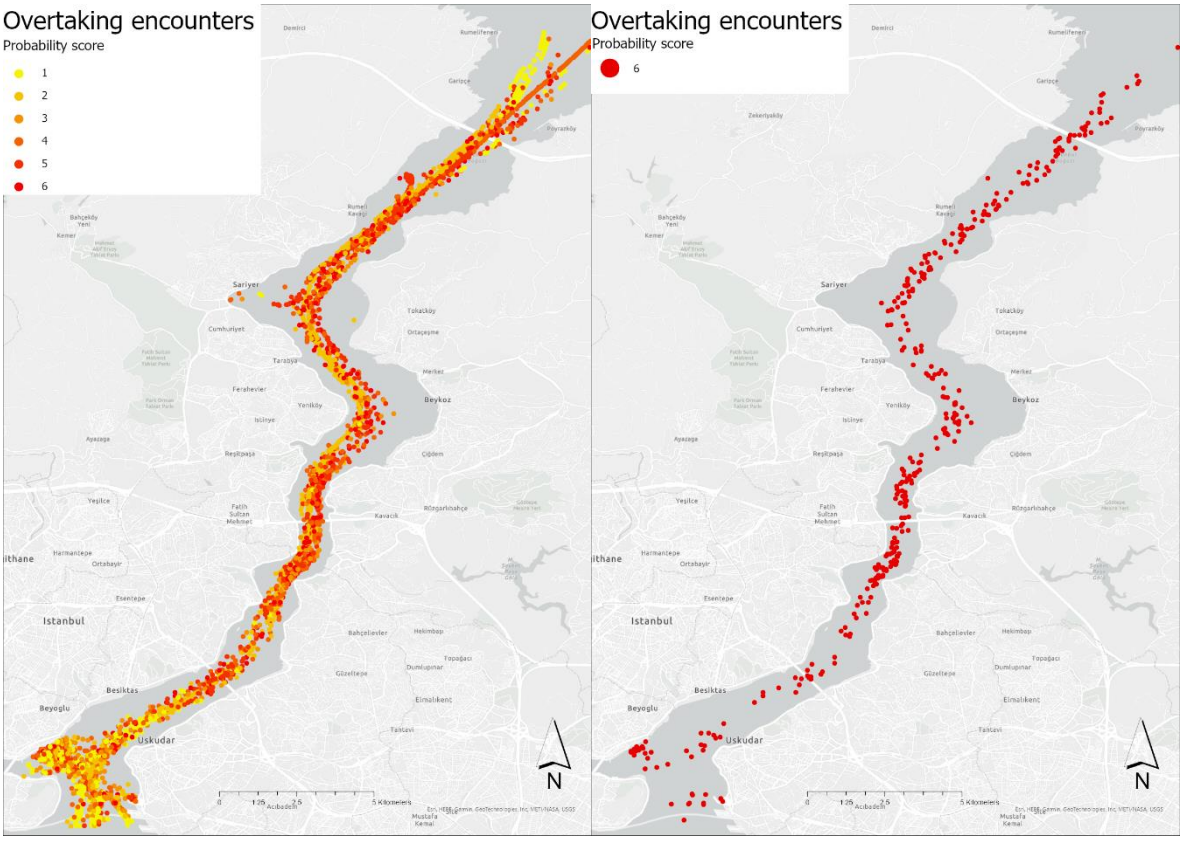

Figure 17: Probability maps of overtaking encounters (left: all categories, right: category 6 only)

Risk score

With all the required variables calculated and analysed, the risk scores of the overtaking ship encounters are calculated and shown in figure 18. Just like the risk scores of the head on encounters, no encounters result in a risk score between 5 and 6. The overtaking encounters do contain risk scores between 3 and 5. The high-risk scoring points are in the middle and bottom of the strait. As the probability scores were relatively evenly distributed, it can be said that the risk consequences values are higher in these areas where the risk score is highest. Especially compared to the other high risk probability locations. These scores show that the bottom part of the strait and the middle bridge section are possibly the high-risk areas for overtaking encounters. But whether this is also statistically the case, will be determined later in this research.

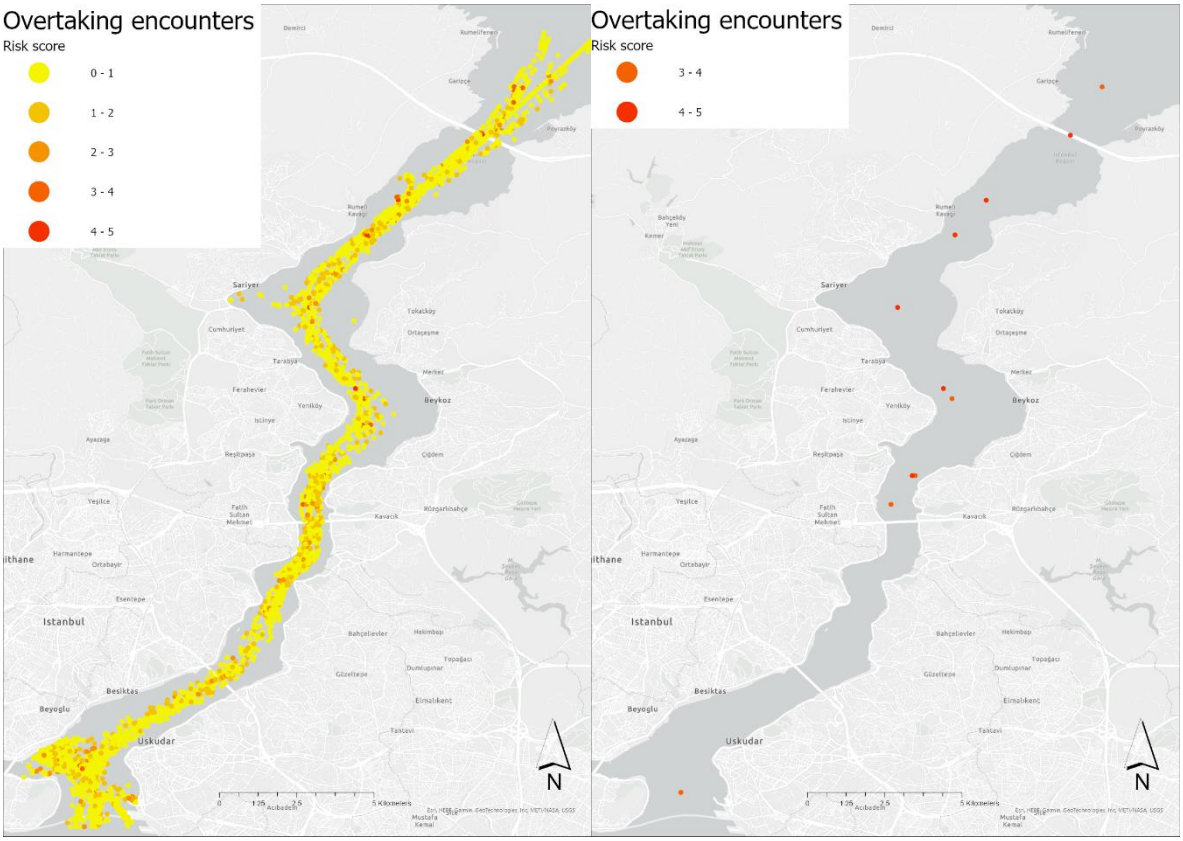

Figure 18: Risk score maps of overtaking encounters (left: all categories, right: high scores only > 3)

4.2.3 Crossing encounters

The final encounter type to be analysed, is the crossing encounter type. Records with this encounter type that fit the model and where a ship domain violation was measured, are a total of 30518. The number of encounters derived from this, are 1912.

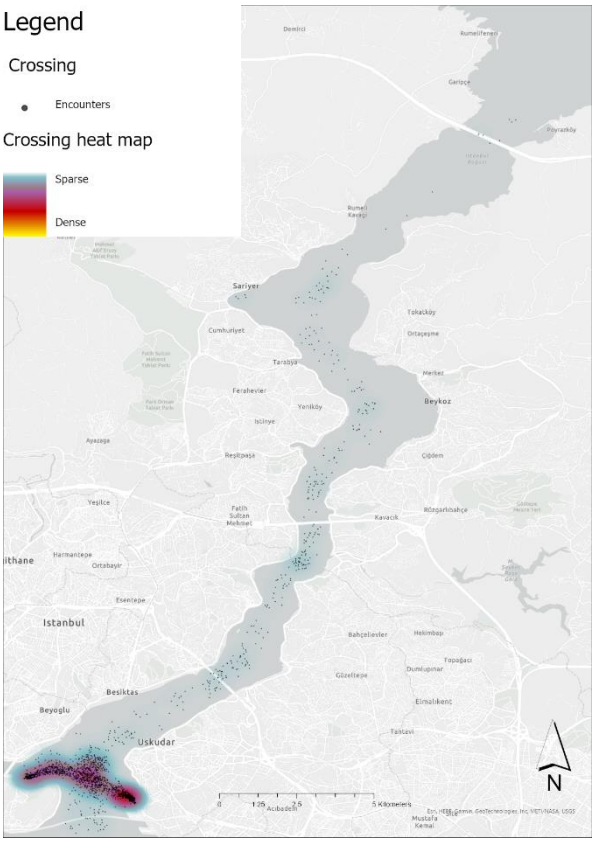

Figure 19: Heat map of crossing encounters

The crossing encounters in figure 19 show that most of the crossing ship domain violation encounters occur in the bottom of the Strait of Istanbul. In the case of the crossing encounters, this is an expected outcome as the bottom of the strait holds an intersection of docking areas and the continuous passage of the strait. Hence why ships cross each other more often and have more ship domain violations there. This bottom area is the main location of clustering of crossing encounters. Further up the strait, there are fewer crossing encounters, the top section of the strait is very sparce in crossing domain violation encounters. This is likely because this area is less urban and thus has less local traffic, local traffic crossing large passing vessels is more likely to cause a crossing encounter. This combined with this section being only straight, results in little to crossing encounters. The following section will analyse the consequences variables and the probability variable to calculate the risk score or the crossing encounters.

Minimum distance

The mean minimum distance of the crossing encounters is the 123.1 metres. Compared to the two other encounter types, this is an average distance. Higher than head on encounters and lower than overtaking encounters. The crossing encounters ship domain shape is similar to the overtaking shape (figure 6**)** but as ships approach each other more perpendicular, the centres of the ships can be on a relatively similar distance, but the bow of the target ship can be closer to the own ship in a crossing encounter.

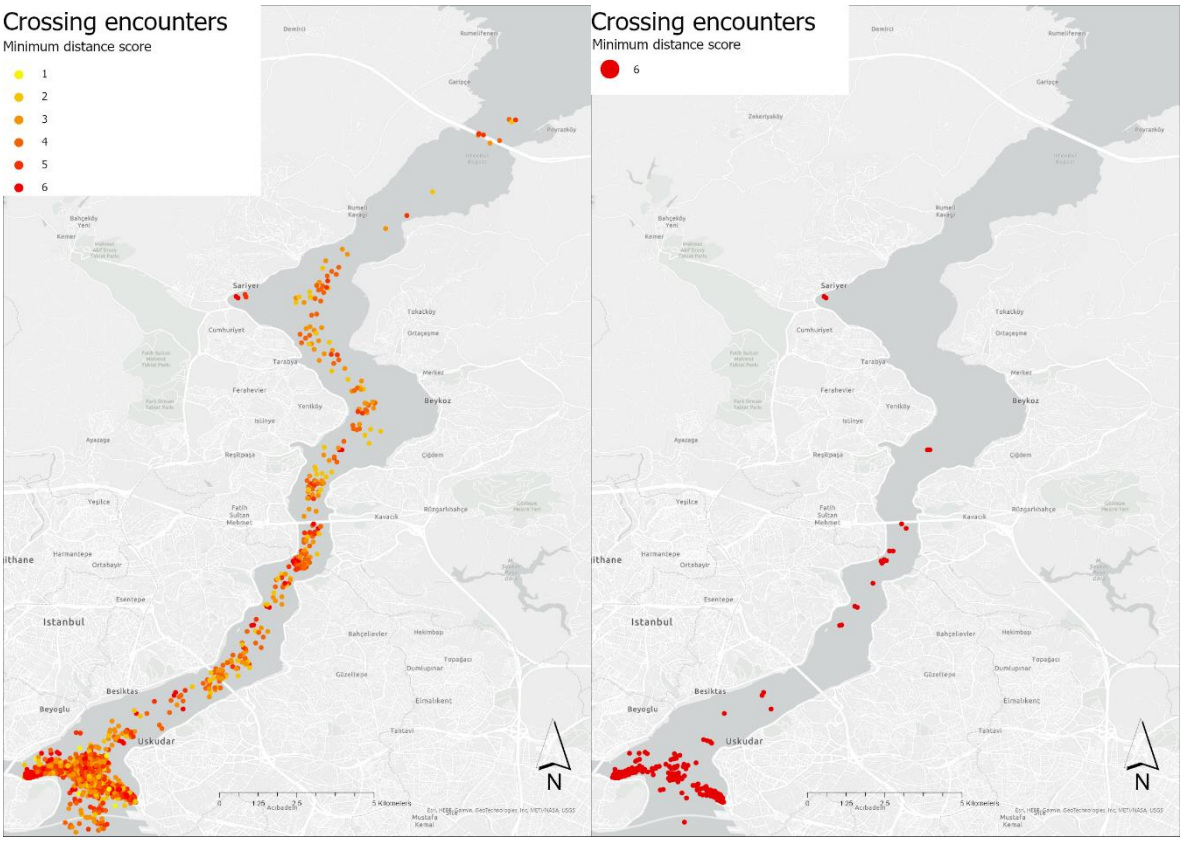

Figure 20: Minimum distance maps of crossing encounters (left: all categories, right: category 6 only)

The recorded minimum distance scores of the crossing encounters are shown in figure 20. The records show a clustering of records in the bottom of the strait and especially closer to the quay. This indicates that ships leaving these docking areas tend come within a minimum distance of each other that can be deemed as risky.

Relative velocity

The relative velocity of crossing encounters is relatively average compared to the head on and overtaking encounters, with the mean relative velocity lying between the other encounters. The mean relative velocity of crossing encounters is higher than the mean of the overtaking encounters, but still much lower than the mean of the head on encounters. Crossing ships could likely keep more speed if they can cross each other without having to interfere but the speed will have to be managed in a way in which the ships do not risk colliding, thus keeping a safe relative speed.

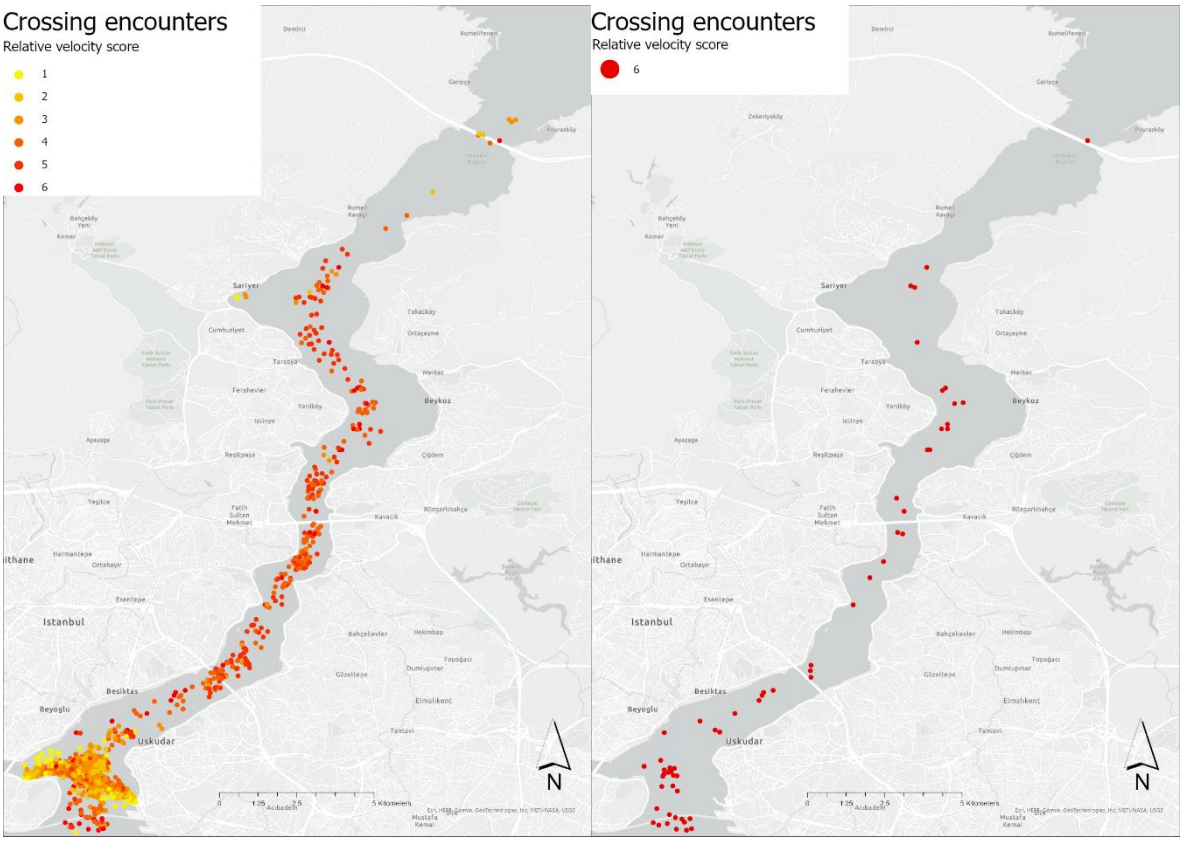

Figure 21: Relative velocity maps of crossing encounters (left: all categories, right: category 6 only)

Figure X shows the relative velocity risk scores of the crossing encounters. The amount of category 6 relative velocity points is relatively low and do not follow an exact locational pattern in the strait. The pattern in figure X is very different from the minimum distance map. Where the minimum distance map shows many risky distances at the bottom of the strait, the relative velocity map shows very little risk in this section of the Strait of Istanbul.

Entry to minimum distance

The entry time to minimum distance has a mean which is the lowest for all the three encounter types, the mean domain time however, is lower than the overtaking encounters. This means that the ships reach the minimum distance relatively quicker in crossing encounters than in other encounters if you approach it by time relative to the domain time of the encounter. The scores for entry to minimum distance time are shown in figure 22.

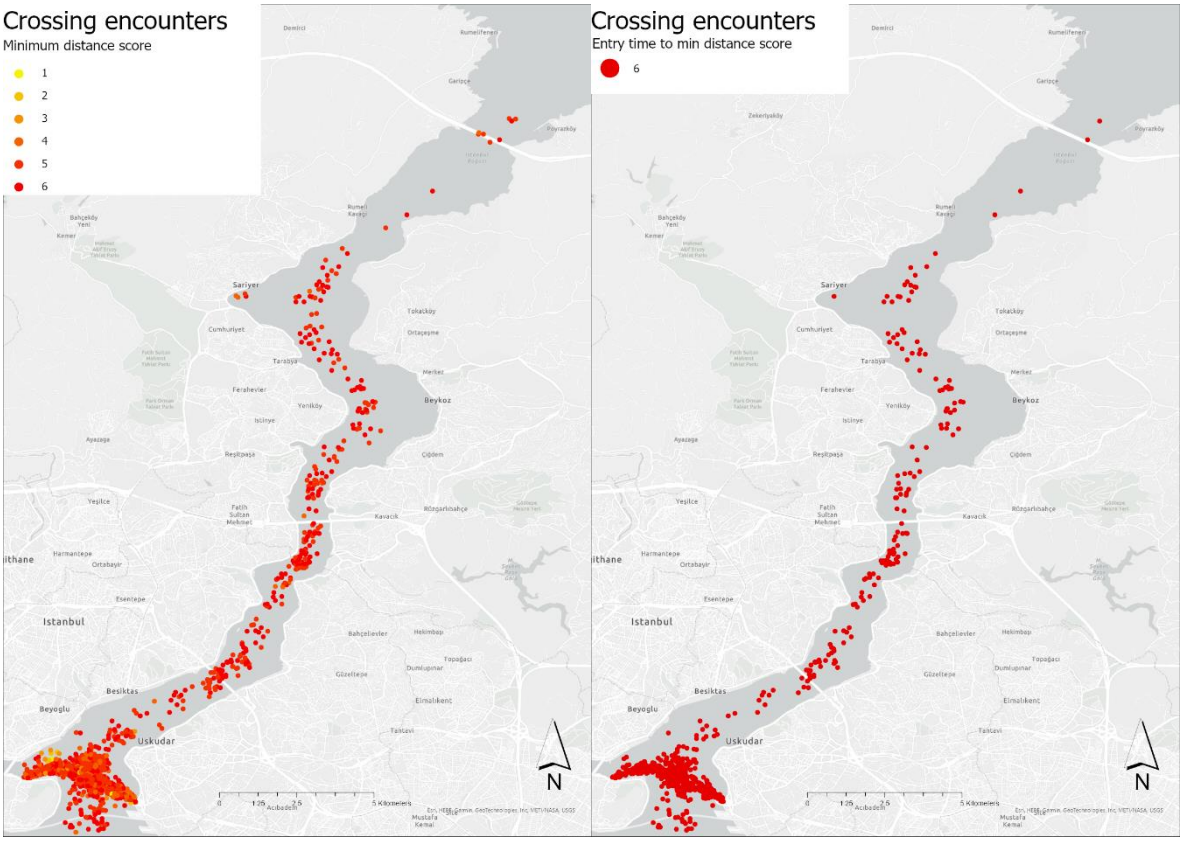

Figure 22: Entry time to minimum distance maps of crossing encounters (left: all categories, right: category 6 only)

The scores for this variable are, similarly to the overtaking scores of this variable, high in category 6 clustering throughout the entire strait. This indicates a cluster of low entry to minimum distance time values. This could be explained by the finding in the data that many recorded encounters are relatively short. So, these short domain violations, also have a short time to minimum distance value.

Probability

The probability of a risky ship encounter in crossing encounters is shown in figure 23. Many of the high probability encounters occur in the bottom of the strait. This is still also due to the high volume of ship domain violation encounters taking place in the bottom of the strait. Relative to the density of total ship domain violation encounters along the strait, the probability of a ship domain violation is higher in other sections than the bottom section of the Strait of Istanbul.

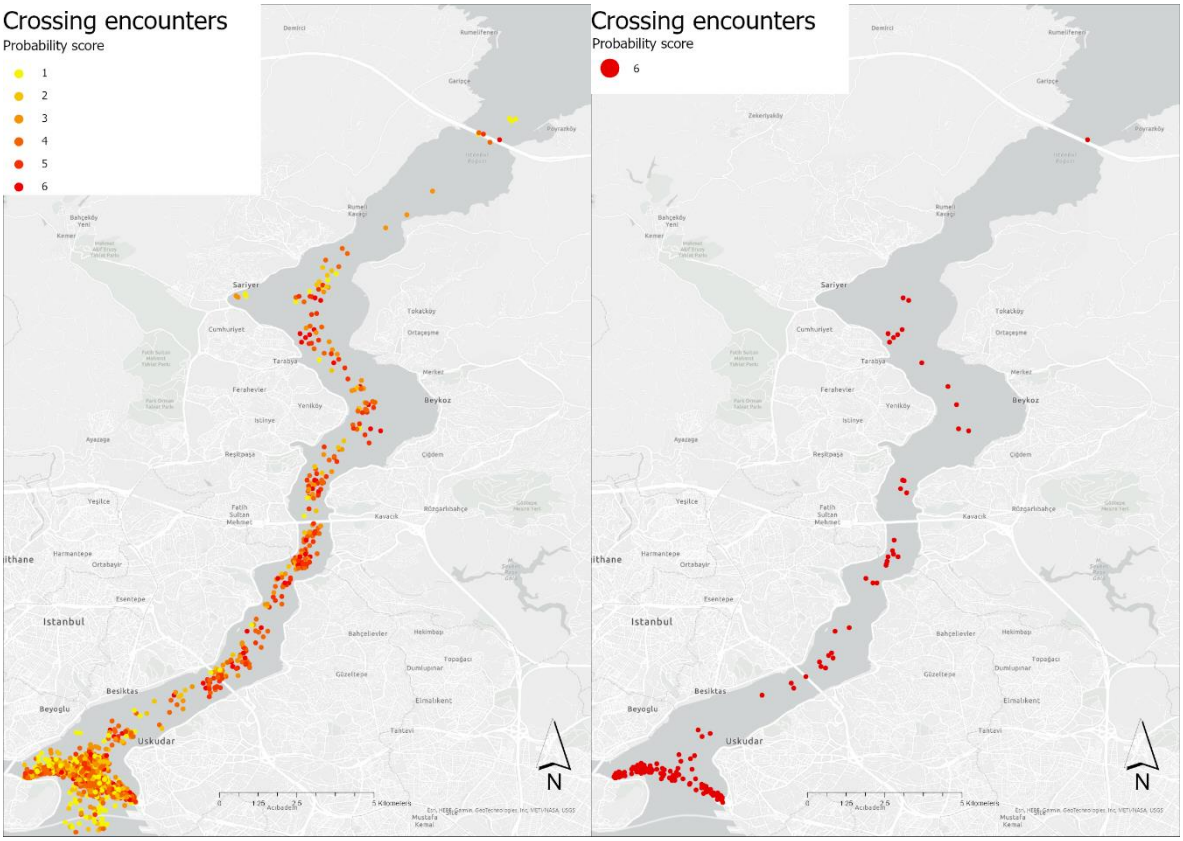

Figure 23: Probability maps of crossing encounters (left: all categories, right: category 6 only)

Risk scores

The risk score maps in figure 24 show that there is an even dispersion of high scoring encounters through the Strait of Istanbul. In contrary to the balance in total datapoints of the crossing domain violation encounters (figure 19), the bottom of the does not contain many high-risk encounters. This can be explained by the relatively low probability values of the crossing encounters in the bottom of the strait. Crossing encounters occur often in that area of the strait but very often, they are not risky or not even a domain violation encounter. This results in a relatively low number of high probability encounters in this area which then results in a relatively low number of high-risk encounters.

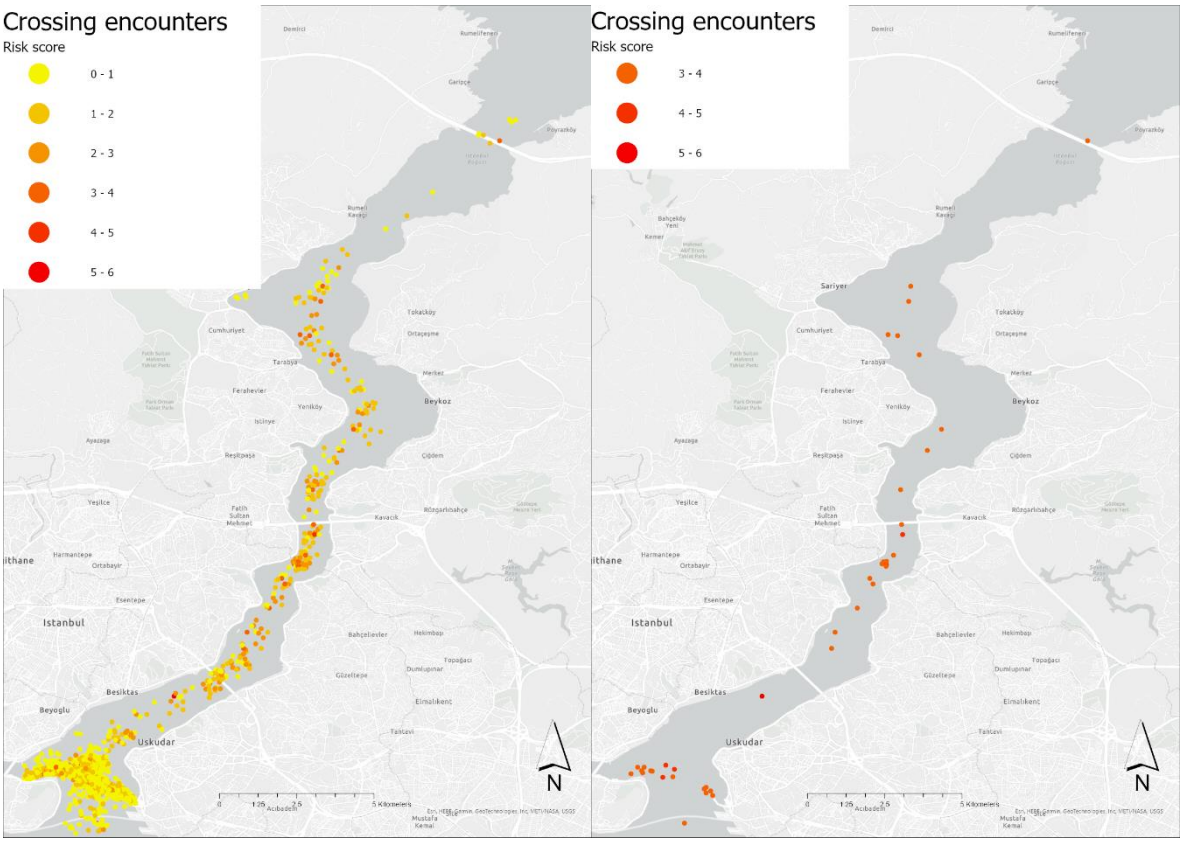

Figure 24: Risk score maps of crossing encounters (left: all categories, right: high scores only > 3)

4.2.4 Total risk scores

The previous paragraphs have given insight in the absolute values and findings of the encounter type calculations. Looking at the individual encounters types gives information on the way ships act in different parts of the strait in different encounter type scenarios. To be able to answer the research question however, all the values will have to be analysed and interpreted. This paragraph will investigate the combined values of the risk scores. Figure 25 shows that the combined spread of potentially risky ship encounters is centred in the bottom of the strait. The middle also represents a fair part in the spread of the encounters, but the majority of encounters occur in the bottom section of the strait.

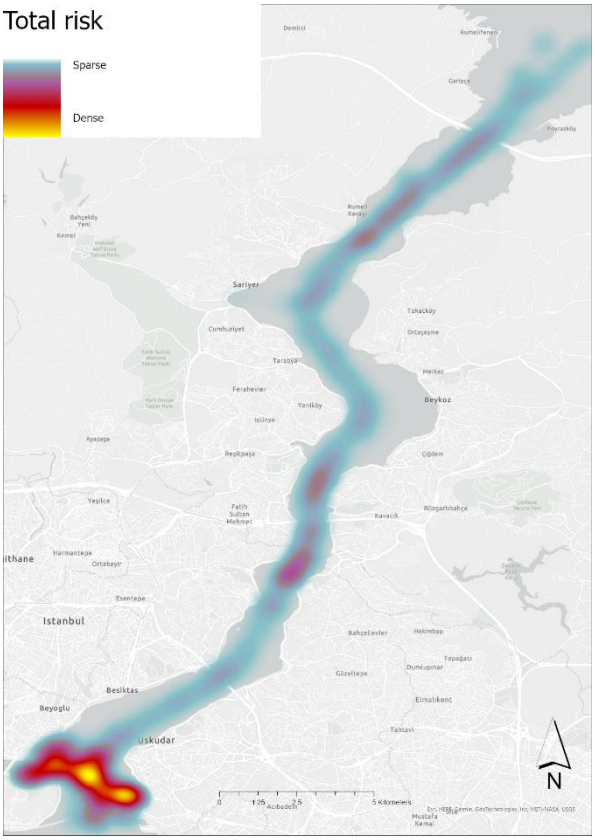

Figure 25: Heat map of total risk

Consequences

Combining the three variables with the indexation method stated paragraph 3.3.4 will result in the consequences score of the head on violation encounters. This score is one half of the final matrix indexing and shows the severity of the potential risk. A higher score means that, in that case, a higher risk situation has occurred. This score does not yet consider the likelihood of this situation occurring; therefore, the probability variable will be the other half of the matrix calculation to result in the final risk scores for each of the ship domain violation encounters.

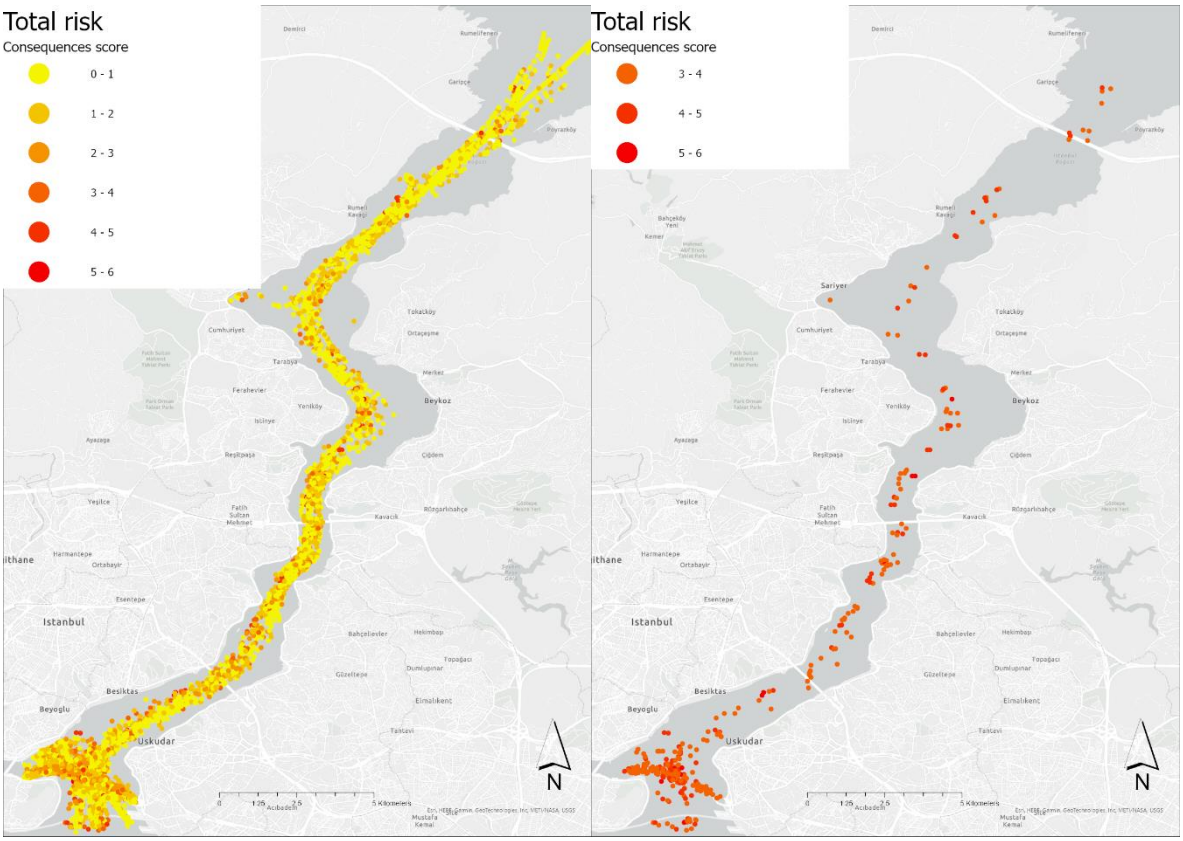

Figure 26: consequences maps of total risk (left: all categories, right: high scores only > 3)

The maps above show that the consequences scores are relatively high in the bottom area of the strait. Moving further north, there are less high scoring encounters which indicates a few interesting findings. Many of the high scoring points (consequences score > 3) are concentrated in the bottom, so in that area, the largest number of risky situations occur. The northern section of the strait on the other hand, has very little high risk consequence situations occurring. This finding can be explained by multiple things. Firstly, the number of high scoring encounters is high as there is an intersection with docking areas and a ferry which increase traffic and could cause ships to have to traverse in a riskier manner. Secondly however, the total number of ship encounters is much higher in this area. As there is a higher volume of ships meeting and thus more ship domain violation encounters, the chance of a high-risk situation happening is higher as well. The high traffic volume in the bottom of the strait results in the fact that it cannot be said that this area is immediately risky as it is not clear what the chance is of an individual risky situation occurring. To be certain of the chance of the risk occurring, the probability will also have to be considered.

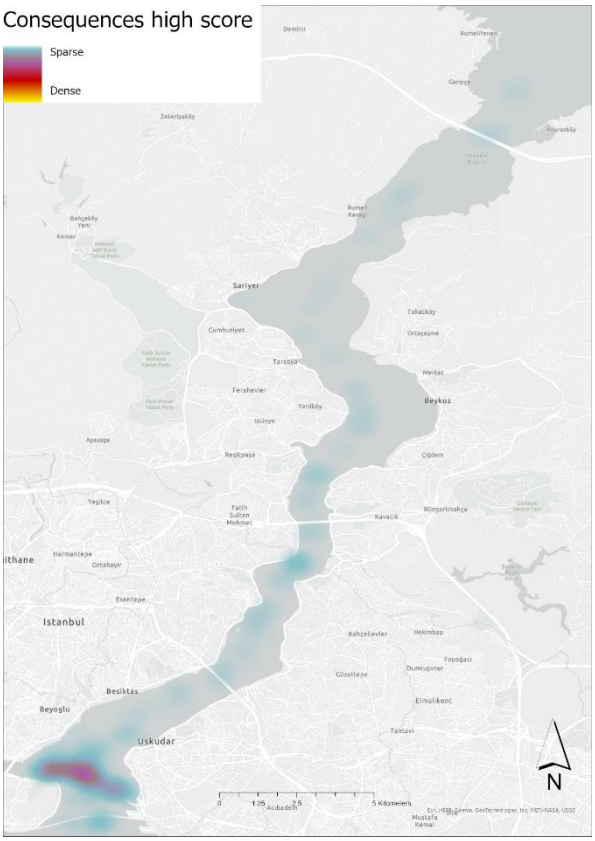

Figure 27: Heat map of high consequences scores (score > 3)

Probability

The second variable in the risk matrix is the probability of a risky situation occurring. As explained in the previous paragraph, this is an important step to find whether a situation is actually risky or whether it is just a rare risky occurrence that happens over time. This paragraph will go into the final findings of the probability scores for the risk score calculation. Comparing the spread of probability values with the spread of the encounter records gives insight in the balance between volume of records and actual high probability risk encounters. As seen in figure 27 above, the ship domain violation records are centred in the bottom of the strait. An area with many records has a higher chance to have more high-risk records than areas with a lower volume of total ship domain violation records. Therefore, it is of interest to look at the spread of the high probability values of the ship domain violation encounters. These are shown in figure 28. Here you can see that the probability of a ship domain violation is more evenly spread along the strait than the spread of the records. High risk probability records still occur a lot in the bottom of the strait, but also more in the middle of the strait and the top part of the strait is also more represented. This shows that the probability of ship domain violation situations occurring is overall more evenly divided than the total set of records. So, though the bottom section of the strait has the most ship domain violations recorded, not all the records have a high probability of being a ship domain violation, due to the many total AIS records in that area, the amount of ship domain violations is higher. Whether the spread of these records is also a significant factor in the actual location of risky encounters, cannot yet be said based on these findings.

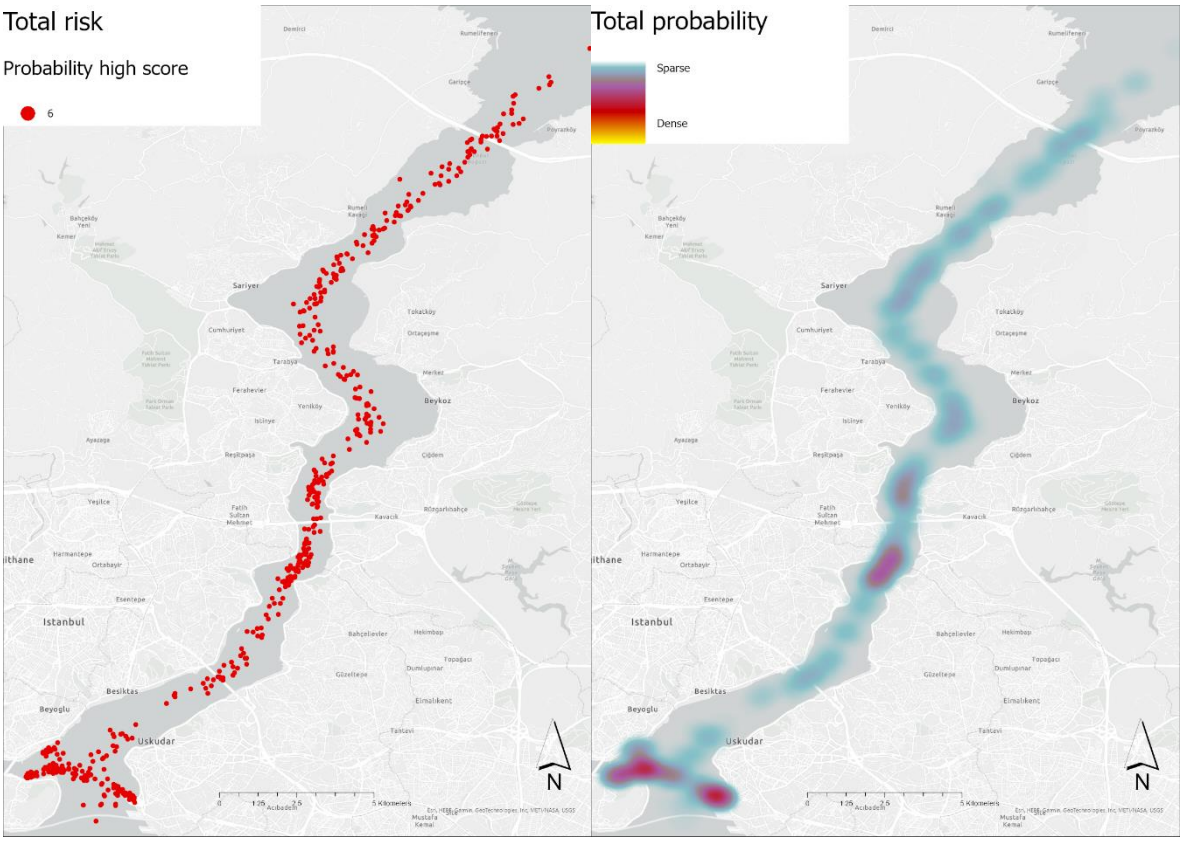

Figure 28: Probability maps of total risk (left: category 6 only, right heat map of category 6)

Risk score

A following step towards mapping the risk of ship collision in the Strait of Istanbul is looking at the total risk scores and the spread of higher risk values in the strait.

Figure 29 shows the final risk scores in the strait and the risk scores from category 4 and above. From these results can be found that there is only one encounter record that scores within category 6 of the scoring matrix. This is understandable as this requires a category 6 score in almost all the required variables. Figure 30 is a heatmap of the high risk (cat. 4,5 and 6) scores. This shows that, in contrast to the probability scores, the higher risk scores are centred in the bottom of the strait again. So, both keeping into account the probability and the risk consequences variables, the high volume of ships in the bottom of the strait seem to score highest in the risk calculation. To map the risk in the strait, these findings will have to be analysed statistically before being able to draw a conclusion on the locational spread of risk in the strait.

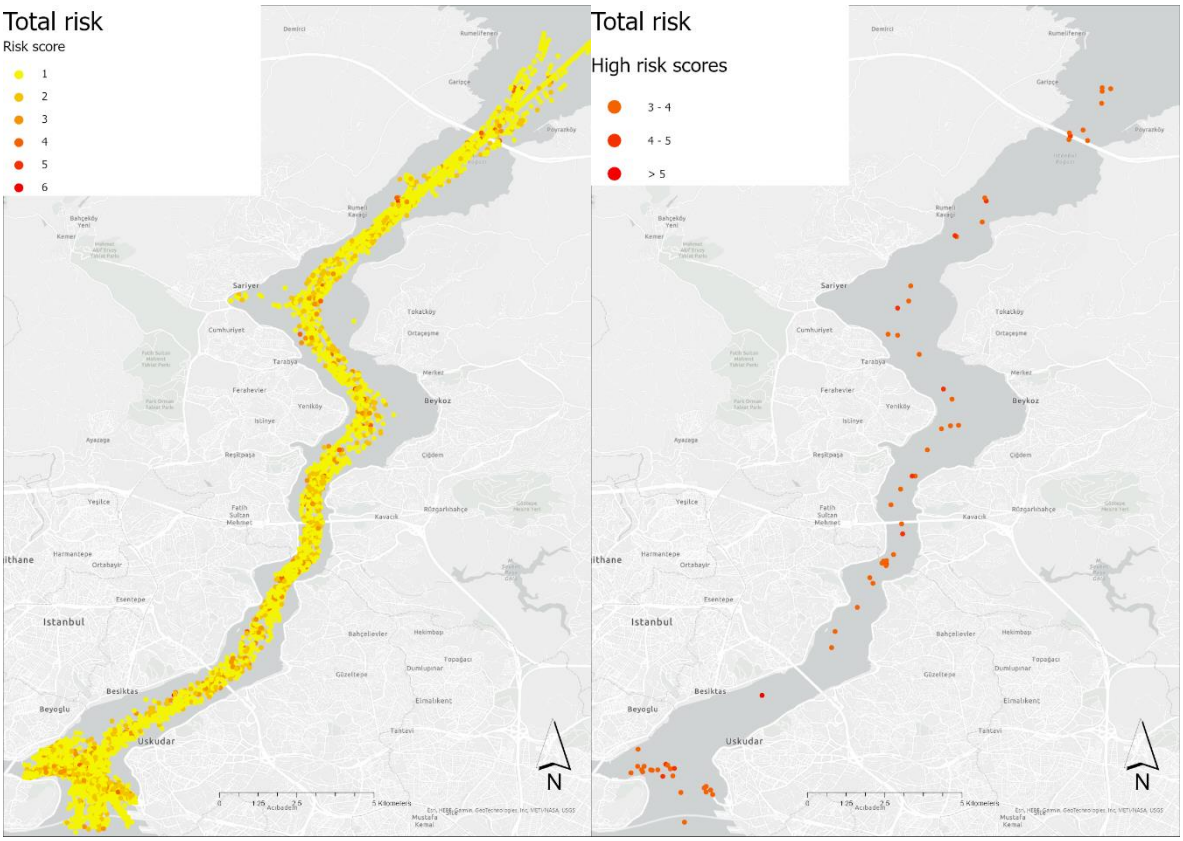

Figure 29: Risk score maps (left: all categories, right: high scores only > 3)

Figure 30: Heat map of high-risk scores (score > 3)

4.3 Hotspot analysis

The previous paragraph shows the spread of risk scores in the Strait of Istanbul. However, as these results are all based on a set of ship domain violation encounters, a statistical analysis will have to be done to be able to conclude statements of the spread of risk in the strait. This will also help to visualise the occurrences of risk in the strait better. For statistically analysing the spread of high-risk scoring ship encounters, a hotspot analysis is performed. This Getis-Ord Gi* analysis will be used to find hotspots where the risk score is statistically high compared to the other encounter records. The findings are shown in figure 31. The results of the hotspot analysis show

The results from the hotspot analysis show that the bottom of the strait contains many statistically high hotspot records, but that there are high risk hotspots found in all areas of the strait. The bending section in the middle of the strait seems to be an area which contains several higher risk records which are statistically significant to the rest of the data. This shows that this area has a significantly higher chance of collision risk than some of the other sections of the strait. Still, the highest hotspot clustering is in the bottom part of the strait. Interestingly, the centre of this bottom area is not a significant hotspot, this could indicate that the ships travelling through this middle section have less risk of collision compared to the other areas. The higher risk is likely also caused by the ships entering and exiting the docking areas of this part of the strait. There is a clear flow towards these two areas on the bottom east and bottom west of the strait. Though there is a statistically significant hot spot in this bottom section of the strait, a sidenote has to be made: the data is designed to leave out stationary ships to prevent a lot of records being significant at port locations, but it is difficult to filter out slow moving ships entering and leaving a docking area when they do move inside the strait. This is a point of attention which will be considered in concluding remarks in this research.

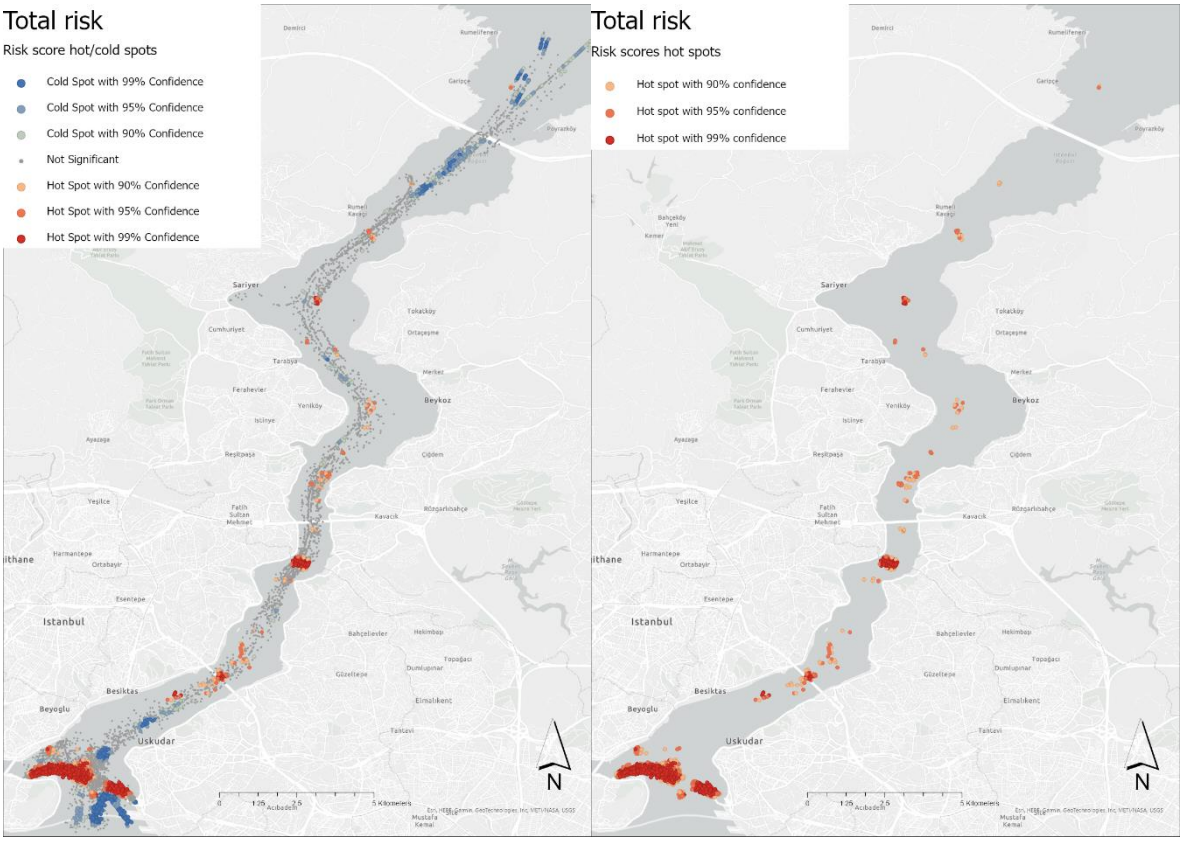

Figure 31: Hotspot analysis of total risk (left: hot and cold spots, right: hotspots only

4.4 High risk analysis

So far, the collision risk analysis has focussed on locational characteristics of the potentially risky ship domain violation encounters. Using point data, the locations of risky encounters has been found and analysed. This paragraph focusses on the high-risk data points and aims to further analyse these results outside of just the point location of the encounters. This analysis will look at the movement of the ships in the high-risk encounters and will look at the other data that is present in the AIS dataset attempting to qualitatively interpret possible connections between the ships' characteristics and the high-risk score given by the risk identification model.

4.4.1 Movement analysis

Earlier in this research, the ship domain violation encounters used the point of minimum distance between the own ship and the target ship to indicate the location of the ship encounter. In reality, the ship domain violation encounter does not consist of a single point, instead it consists of two ships moving alongside each other in the Strait of Istanbul. It could therefore be of interest to look into the movement of these ships to give further insight into why a certain point in the Strait of Istanbul could cause a higher collision risk.

Figure 32 shows that there is a clear difference in length of ship domain violations throughout the entire strait. Many of the high-risk encounters have shorter movement lines but there are several encounters present where the encounter is much longer in length. The right side of figure 32 and figure 33 zoom in further on the case to get a clearer view on the movements of the individual encounters. Several hotspots that are derived from the hotspot analysis show some interesting movements of ships. The northern section (figure 32) shows that the bend of the strait in the bottom of this figure is a wide bend where ships still move very close to each other. This is likely due to the depth of the strait in this section, most of this section is likely too shallow to traverse with a large vessel and thus, despite of the wide waterway, ships have to travel in close proximity of each other, causing a higher risk of collision.

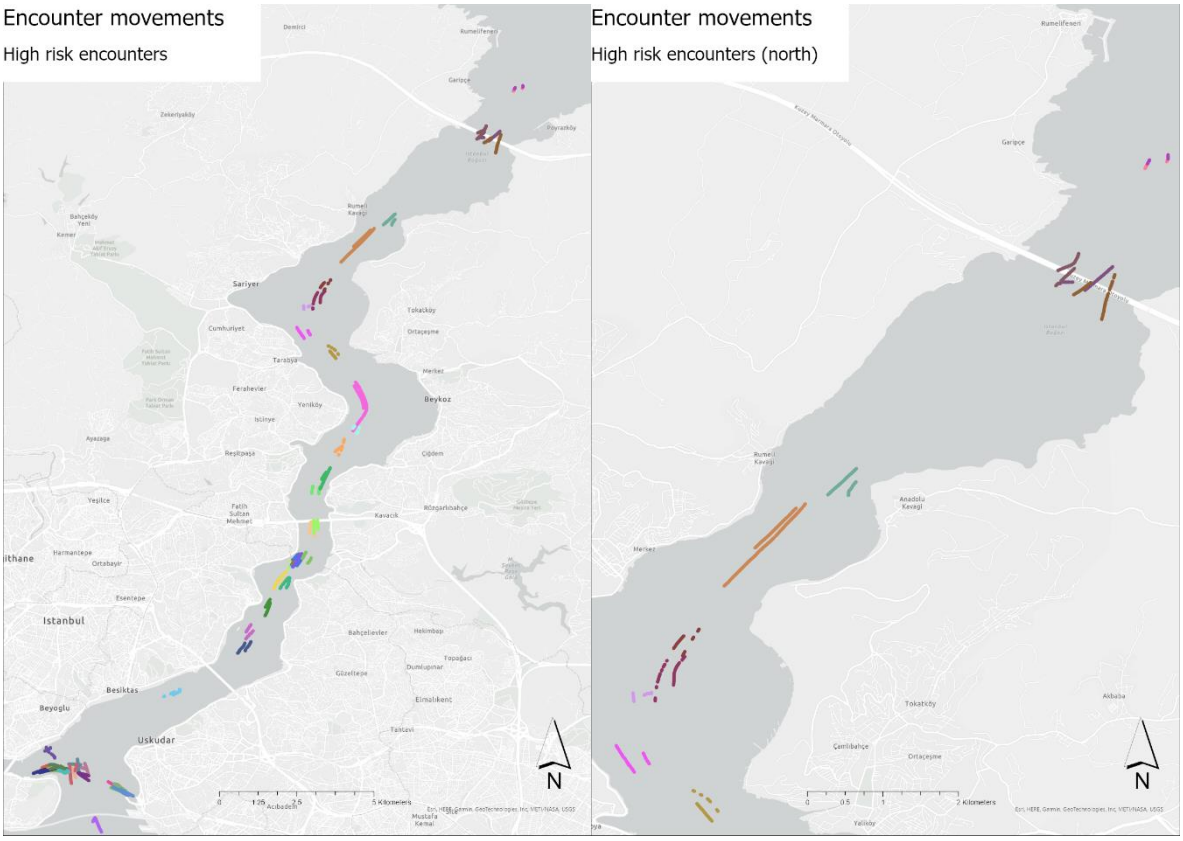

Figure 32: Ship pairs movement analysis (left: full view of Strait of Istanbul, right: north section of Strait of Istanbul)

The middle section shown in figure 33 is one of the narrowest and curvy areas of the Strait of Istanbul. It is therefore also a section where a high collision risk cluster is present. This figure shows a large cluster of high-risk movement south of the bridge. The nature of the curves in the strait causes the ships to have to move close to the east of the shore which is likely the reason for the high number of high-risk encounters at this location. The movement lines of these encounters are also short which indicates that the ships move close to each other and when they have room, they will move away from one another again. These high-risk movement lines show characteristics of movement in and around a chokepoint in the strait.

Figure 33 also shows the bottom section of the strait. As stated earlier, this is the busiest area in the strait containing an intersection between docks (east-west) and the cargo vessel travel route (north-south). Interestingly, few of the high-risk encounters are caused by the crossing of east-west and north-south traffic. Most of the high-risk situations occur in near the docks which could be caused by the fact that ships are very close to one another when entering or leaving a dock. The encounters found here are mostly crossing encounters which cross paths slightly near a dock.

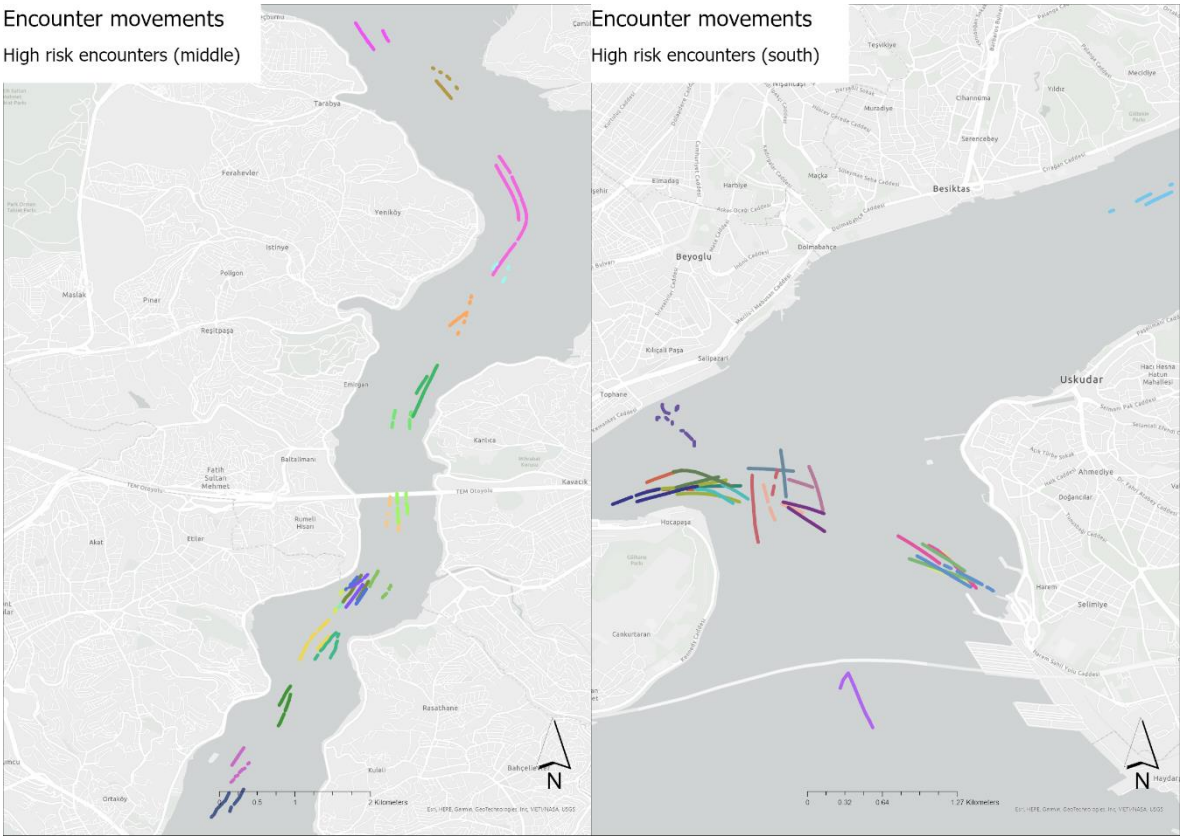

Figure 33: Ship pairs movement analysis (left: middle section Strait of Istanbul, right: south section of Strait of Istanbul)

Qualitatively analysing the movement lines of the high-risk encounters has shown that the curves in the strait have a high impact in the movement patterns of the ships and that these curves can be the cause of a higher risk of collision. Additionally, the cluster of high collision risk encounters in the bottom of the strait is mostly caused by the movement in and out of the docks at the eastern and western shores of the strait. There are relatively little high-risk encounters recorded which are caused by the crossing of dock traffic and cargo traffic moving along the strait.

4.4.2 High-risk ship characteristics

The previous section discussed movement patterns of the high-risk ship domain violation encounters. It is now necessary to go into the ship characteristics of the encounters to see whether there are

patterns to be seen which can explain the higher risk of the ship encounter. Previously, most of the analyses have been performed on an encounter level, looking at characteristics of the encounter such as the relative velocity and minimum distance between the ships. Besides this, the ships also have certain individual characteristics which could differ vastly between different encounters. This section aims to analyse these characteristics and see what the different ship-related aspects are of the high collision risk ship encounters.

Table 5: Averages of ship characteristics per risk type selection

Risk type	Rate of turn	Speed (knots)	Ship size (m^2)
Total ships	61.4	9.3	2873
Domain violation encounters	74.1	7.8	2449
High risk encounters	109.8	18.7	1159

Table 5 shows the difference averages of several ship characteristics between the total set of recorded ais information, the total ship of the ship domain violation encounters and the high-risk encounters. These values show that there are noticeable differences between these averages. Firstly, the average rate of turn is the lowest for the total dataset. The selection of ship domain violation encounters shows an increase in average rate of turn and the selection of high collision risk encounters results in the highest average rate of turn. This means that, in the case of this research, when ships get closer and up to a risky situation, the average rate of turn increases. This can be explained by the need for ships to manoeuvre away from one another when they get into a situation with a higher risk of collision thus having an increased rate of turn. The average speed of the total set of ships is higher than the selection of ship domain violation encounters. However, the average speed of high-risk encounters is substantially higher than the other two averages. Ship encounters that score high in the risk identification model, have on average a much higher speed than ships in a less risky scenario. This could be due to if a ship has a high speed, it is more likely to have a higher relative velocity compared to the other ship in the encounter and as the relative velocity is used in the risk calculation, a higher relative velocity is likely to contribute to a higher risk score. On the other hand, it can also be said that ships that travel at a higher speed take more risk and thus have a higher risk score. The final average is the ship size. Measured in square metres, the ship size is large on average for the initial dataset. The ship domain violation dataset contains a set of ships which are smaller on average. The high-risk ship selection is even smaller with an average of 1159m^2. This would indicate that smaller ships have a higher chance of getting a high risk of collision score. Whether that is truly the case, cannot be determined from these findings. What can be said is that, in the case of this research, the average ship size is lower in high-risk cases than in the total ship dataset. It is important to state that the findings of this analysis serve the use of exploring potential results and do not have any statistical evidence. Hence, why the possible relations between high risk and the above-mentioned ship characteristics cannot be verified. To be able to draw concrete relations between the ship's characteristics and the risk of the encounter, a statistical analysis will have to be performed which would be suggested for further research.

4.5 Validation: comparison with previous research

The model created in this research is derived from variables that have seen significance for risk identification in earlier researches. This method however, being a new interpretation of identifying risk, cannot be immediately defined as a valid model. Therefore, the model will be validated with collision risk findings of earlier research in the Strait of Istanbul. The research of Altan (2019) has used the collision diameter method to analyse collision probability in the Strait of Istanbul.

The model of Altan (2019) is represented in a spread of collision probability. This is visualised with points for different sectors of the strait which leads to a grid-like view of the strait with each point having a collision probability value assigned. The characteristics of these results are different from the ship domain-based results presented in this research. Therefore, it is not possible to directly calculate the similarities and differences between the models for validation. The cluster results from the domain-based analysis performed in this research does not output evenly spread results along the strait and is based on the actual points of encounters occurring. The clusters do however provide an aggregated way of showing where the risk is highest according to the model. This can be manipulated to be useful for comparison of the two models

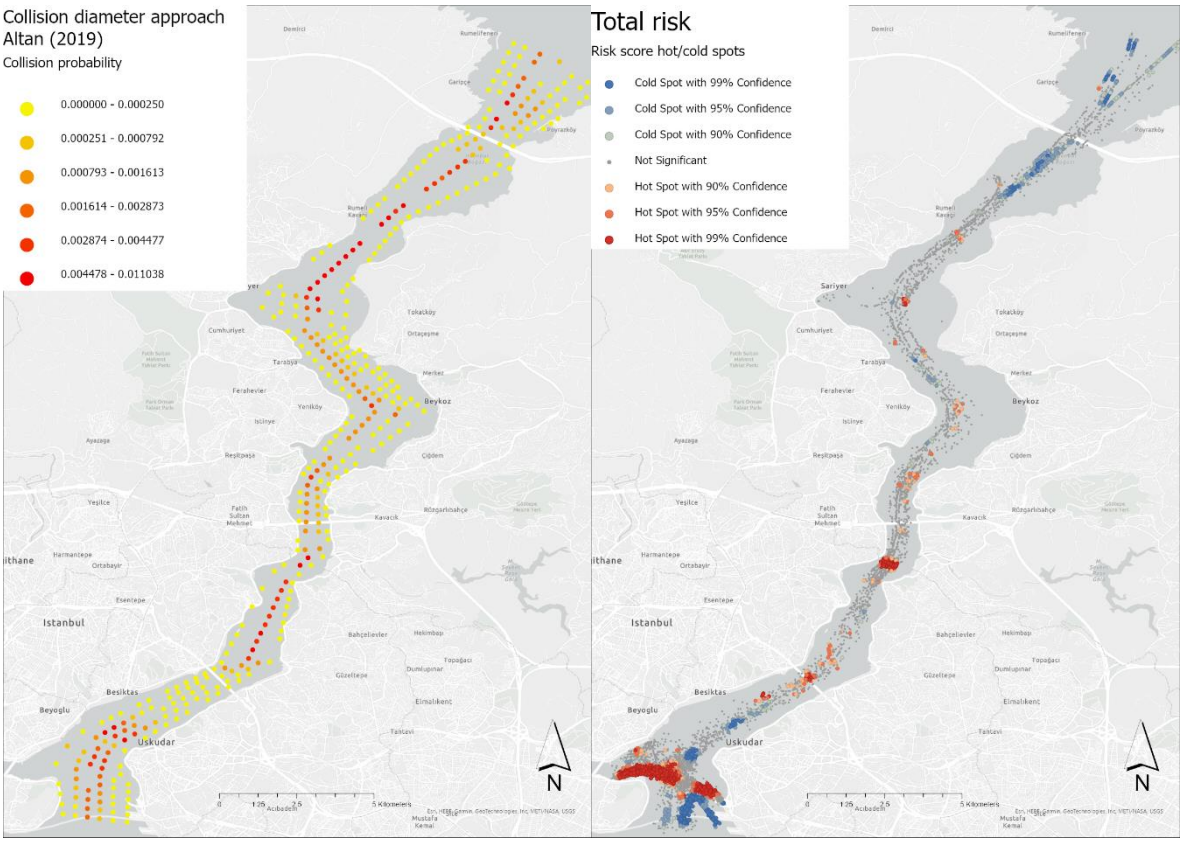

Figure 34: Collision diameter approach (Altan, 2019) and hotspot analysis

To be able to compare the two models, the results are overlayed. As both results sets are initially point data, one of the results will have to be converted to a different datatype to be able to visually compare the two. This is performed in figure 35. Here, the results of Altan (2019) have been interpolated using inverse distance weighting (IDW). The reason this method is applied to the results of Altan (2019) is that the data is evenly spread which allows for a safer interpolation between the points. In this way, a rasterised contour of the entire Strait of Istanbul can be made for the purpose of overlaying the domain-based results on to this.

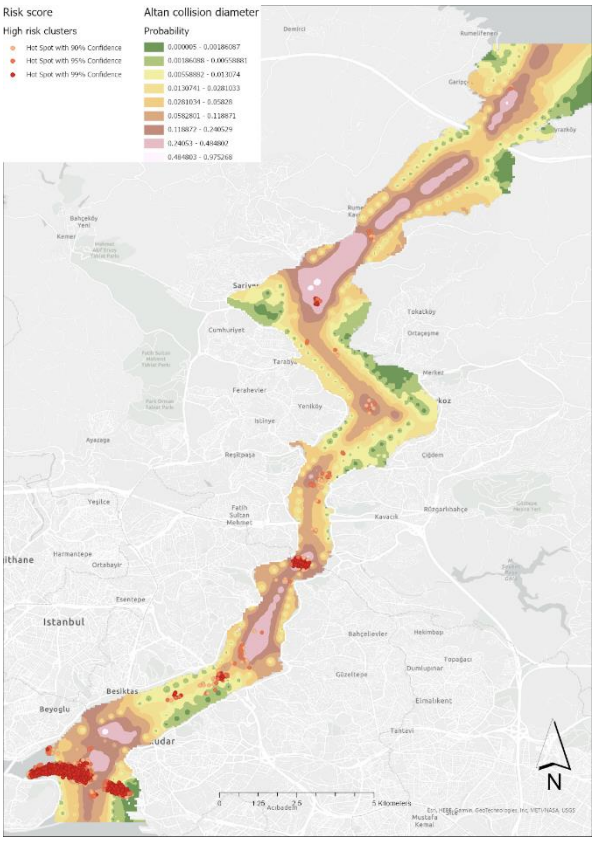

Figure 35: Comparison collision diameter and risk identification model

The comparison between the two results shows that there are many similarities in the findings of both researches. The risk clusters appear in the areas where Altans (2019) risk probability is highest which indicates that the results from the domain-based approach performed in this research are reliable. Despite the similarities, there are also some differences to be found in the comparison. One of the differences is that the cluster movement from east-west is not present in Altans (2019) results. The nature of this research applies distance and relative velocity as important indicators for risk whereas the collision diameter approach does not weigh these as highly. This can explain the cluster of east-west traffic. Whether this traffic is actually risky is difficult to say, but the bottom of the strait is still a high collision risk cluster in both researches. Another finding is that there is a cluster of risk in just north of the bottom section of the strait where there is little to no risk in Altans (2019) research. But given these points and given that the goal of the validation is to see whether the results in the model are reliable, it can be said that the model is a valid representation of reality. There are enough similarities in the output of both models to say that the model does make sense to interpret. The differences between the two outputs can also be seen as interesting points to see whether these are also risky and not included in Altans (2019) model.

4.6 Conclusion of results

In this chapter, all the relevant results have been analysed and interpreted. It has been found that the ship domain violating encounter records have a differing pattern of risky variables per encounter type. This shows that ship domain violations with different encounter types, tend to occur at differing locations within the Strait of Istanbul. The bottom section of the strait has the most occurring ship domain violation encounters in all three of the encounter type analyses. However, within this section, the location of the ship encounters also seems to differ per encounter type. These differences become more prevalent whilst looking at the category 6 values for each of the variables. The results for the different encounter types can however not be compared very easily as the difference in number of records differs greatly between the three encounter types.

Combining the three encounter types has given more insight in the total spread of risk in the Strait of Istanbul. The combination of the three encounter types saw a concentration in the bottom of the strait which can also be seen when only the higher risk score points are mapped. Further assumptions can be made when the total score data is analysed with the cluster analysis, which shows that clustering of high risk is present in different parts of the strait, but it mostly appears in the bottom of the strait.

The high-risk score analysis has shown that the high-risk encounters occur most often in two scenarios, moving in an overtaking or crossing encounter in a bend in the strait, or moving from east to west (or west to east) in the bottom of the strait. The movement lines show that most of the encounters occur in this pattern. Furthermore, the high-risk score analysis found patterns in averages of ship characteristics. High risk scoring encounters have on average a higher speed, higher rate of turn and a lower ship size if compared to the averages of the total data and the dataset of the ship domain violation pairs.

Finally, the validation of the results with Altans results from the collision diameter approach research show that results occur in similar areas which indicates that the model creates results that are likely to be a correct representation of the situation.

5. Discussion

This research has been an explorative process of looking for the potential of mapping maritime collision risk using AIS. The following section will reflect on the process of creating a model and analysing the results with the goal of reflecting on this process and finding further improvements and suggestions for further research. As this is explorative research where the research question revolves around the potential of a new risk identification method, the discussion chapter will help to answer the research question in the concluding chapter.

The focus of this research was the creation and analysis of a risk identification model for collision risk in the Strait of Istanbul. An iterative process was ran using an empirical ship domain contour. The results have shown insight in potential collision risk for the analysed records in the Strait of Istanbul. The validation process has shown that the created model is able to present an accurate representation of reality and that there is potential in further analysing collision risk using this approach.

The iterative method of creating a model which analyses the ship domain violation encounters does come with some shortcomings. Looking at these shortcomings can help to answer the research questions and improve future research and is an important step to take in reflecting on the exploration of this risk identification approach. One of the first discussion points is the translation of AIS messages, which have many points per ship voyage, to a single point encounter location on a map of the Strait of Istanbul. Two ships meeting each other results in the encounter of an own ship and a target ship. These both move in the waterway over time making it difficult to define the exact location of a risky ship encounter. Analysing the risky ship encounters required the encounters to be point data. Other data types such as line fragments along long parts of the strait are hard to perform analyses on like the hotspot analysis. Thus, the encounters required a transformation from a multipoint set to a single point on the map. The choice was made to use the point where the minimum distance between the own ship and target ship is reached as this often indicates a middle point for the ship domain violation, is an important factor for calculating the risk score and is the point where the potential collision is closest. As many of the ship domain violation encounters are relatively short and do not cover a large section of the waterway, this transformation is interpreted as a good representation of the encounters.

The used variables for creating the risk consequences scores have been derived from literature and the availability in the AIS dataset. These variables, though of importance do have some important sidenotes which have to be addressed after the creation of the risk model. This is the case with the ‘entry time to minimum distance time’ variable. This variable calculates the absolute time between the target ship entering the own ship ‘s domain and reaching the point of minimum distance. As this is the absolute time, ship encounters with a short total duration, have an overall lower entry to minimum distance time compared to an encounter with a longer total duration. The model states that a short time for this variable results in higher risk, meaning that encounters with a shorter total duration have a higher chance of being seen as risky and as most of the encounters have a relatively short duration, many of the encounters score the maximum risk score for the entry time to minimum distance time variable. To make this variable more accurate, the total encounter duration could be taken into consideration as the variable focusses on the relative approach speed of the encounter, which could result in a higher risk. However, dividing the entry time by the total duration of the ship domain violation does give insight in the relative approach time, but also comes with another shortcoming. Namely that a relative approach value like this does not give any information on the actual speed of approach of the encounter. Also, if the domain violation is very short, the results can be skewed as a little change of removing one point could totally flip the result of the variable.

The results above also conflict slightly with the method of categorisation of the risk score categories. The K-means clustering performed well in selecting clusters of results causing a natural categorisation into the 6 risk score categories. However, it is expected for 10% of the total records to be categorised

as risky (cat 6 for a variable). The values of the entry time to minimum distance variables, however, result in a lot of values in the high-risk category, which is a less accurate representation of collision risk.

The scoring calculation is also found to be a point of attention as the total risk score is a product of 4 separate variables scoring 1 to 6. As little records are expected to score high on just one of the variables, very few of the encounters analysed in this research has scored a total score of 6. This does not have to mean that there are not enough very high risks encounters to represent reality as it could be possible that in one month of data used in this research, little to no very high risky situations have occurred in the Strait of Istanbul.

As this research is explorative, it is difficult to compare this research to other related researches. The validation process shows that the model is valid, but the ship domain-based model and the collision diameter model differ in such a way that it is difficult to fully compare the two and as this research is explorative, there were few expected results set in this research. Expectations were that some areas would be riskier than others, which has been confirmed in the validation process, but there were little expectations on the extent in which risk could be identified and mapped. This will be further discussed in the next chapter.

6. Conclusion

This research focussed on finding and mapping risk of collision between ships in the Strait of Istanbul. An iterative approach is used to create a model which uses from AIS data deducted variables to give ship domain violation encounters an individual risk score. This approach has been performed to ultimately answer the research question of this research:

To what extent can situations of maritime risk of collision be mapped in the Strait of Istanbul using AIS data?

The research has used an empirically based ship domain contour for each of the three different encounter types to find potentially risky ship encounters which had their risk score calculated. The risk score approach used three risk variables which had been found in the literature study and which have proven to be able to help define risk of collision. A matrix-based calculation has been used to calculate the risk scores by also applying a probability value to the model which was needed to determine the chance of the potentially risky situation happening. The results of this score calculation have been analysed to give to model statistically proven results.

The model has been created with AIS data as a basis. AIS data contains a broad range of ship and voyage related information of each of the ships in the Strait of Istanbul. This data can be used to create the model for defining risk in the strait. The findings in the research however show that there are potentially more variables that influence the risk of collision between ships in the strait and that the variables found in the AIS dataset do indeed find patterns and clusters of potential risk but are not strong enough to fully simulate potential risk of collision.

The application of the matrix-based risk calculation gives the model an opportunity to categorise risk into risk scores from 1 to 6. This scoring based on an aggregation between a probability score and a consequences score is a fair approach to scoring risk in this situation. However, because the consequences score is already an aggregation of the three AIS based risk variables, a high-risk final score is unlikely as it would require an encounter to have a very high-risk score on all variables.

The mapping of maritime collision risk is done with the usage of the final risk scores per ship encounter. The mapping process, however, does find some complications regarding different factors. A ship domain violation encounter contains the dynamic data of two ships travelling individually in the strait. Making mapping of the location of the encounter occurring complicated. The best approach to be found has been to map the encounter with a point of the location where the minimum distance between the two ships has been recorded. This location is most likely the highest risk area of the ship encounter. These points have been analysed using a hotspot analysis which shows that the clusters of risk in the strait can be statistically significant. These hotspots show different areas of the Strait having a higher chance of high collision risk than others.

The findings from the hotspot analysis and high-risk analysis have shown promising findings indicating that tighter areas in the strait and intersections of waterways result in a higher clustering of risky situations compared to other sections in the waterway and that a high average speed, high average rate of turn and a low average ship size are present in the selection of the high-risk results. These findings indicate that there is potential to be found in the AIS based collision risk mapping approach and that there is room to further analyse the risk scores on statistically significant findings.

So, the extent in which maritime collision risk can be mapped in the Strait of Istanbul using AIS is that it is a complex process which will require more iterative steps to be able to accurately appoint risk score definitions to the AIS data encounters. This research does however show enough potential in this methodology and provides a new step towards fulfilling the potential of risk definition and mapping.

6.1 Further research

The explorative research performed here has the goal to set a basis for an improved version of this research approach. The subjects addressed in the previous section gives a selection of ideas to consider in future research. This section will go into the subjects that can be advise for future iterations of this research methodology.

Firstly, the data size of the research could be increased from to a larger dataset with multiple months of data for instance. This would cause the results to be more accurate as it is possible that other months of data contain more risky ship encounters and that these could show new findings.

Secondly, to further improve the risk identification model, the methods of Szlapczynski and Szlapczynska (2016) could be used to replace the distance and time values with their time to domain violation (TDV) and degree of domain violation (DDV) parameters. The application of this method could reduce errors due to discrepancies in the data and improve the overall accuracy of the model.

Future research is also advised to look for fitting data beyond the AIS dataset. AIS contains several useful attributes to use as variables for defining risk, but the collision risk of a ship cannot only be entirely derived from these variables. Therefore, it is advised to experiment with external variables that could indicate risk of collision. This could include other factors mentioned in table 2 like weather and time of day.

This research can also form a basis for other types of ship risk analyses. This research has only focussed on ship collision risk and not on other types of risk. Ships colliding with other ships is just one section of different ship voyage accidents that can occur. Grounding and stranding for instance, is another risk that is present for ships travelling through the Strait of Istanbul. By including data on water depth of the Strait of Istanbul and changing certain risk scoring parameters, this matrix method can be used as a basis for ship grounding and stranding risk identification.

Finally, further research could expand the model by including different research areas. Applying this method of ship collision risk identification in other busy waterways could improve the overall research field in two ways: firstly, the method used in this research is widely applicable as the empirical domain and the K-means based scores can be derived from all types of AIS data. So, if the AIS data is available, the method is applicable. This gives the opportunity to apply this method to other areas, giving insight in the collision risk of other areas. Secondly, if this research is applied to different areas, risk related research in that area can be used to compare and validate the model further. More iterations of validation can help to improve the model on a qualitative basis and makes the reliability of the model stronger.

7. Literature

Altan, Y. C., & Meijers, M. Ship Domain Variations in the Strait of Istanbul.

Altan, Y. C., & Otay, E. N. (2017). Maritime Traffic Analysis of the Strait of Istanbul based on AIS data. The Journal of Navigation, 70(6), 1367-1382.

Altan, Y. C. (2019). Collision diameter for maritime accidents considering the drifting of vessels. Ocean Engineering, 187, 106158.

AP news, (2021). Turkey: 1 dead in Bosporus Strait container ship collision. Retrieved from: https://apnews.com/article/bosporus-middle-east-turkey-europe-business-34a2b236edba24eff6f51968bb2ac898

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.

Chang, S. J., Hsiao, D. T., & Wang, W. C. (2014, April). AIS-based delineation and interpretation of ship domain models. In *OCEANS 2014-TAIPEI* (pp. 1-6). IEEE.

Eliopoulou, E., Papanikolaou, A., & Voulgarellis, M. (2016). Statistical analysis of ship accidents and review of safety level. Safety science, 85, 282-292.

Fournier, M., Hilliard, R. C., Rezaee, S., & Pelot, R. (2018). Past, present, and future of the satellite-based automatic identification system: Areas of applications (2004–2016). WMU journal of maritime affairs, 17(3), 311-345.

Friis-Hansen, P. (2007). IWRAP MK II: Basic Modelling Principles for Prediction of Collision and Grounding Frequencies.

Fujii, Y., & Shiobara, R. (1971). The analysis of traffic accidents. The Journal of Navigation, 24(4), 534-543.

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.

Goodwin, E. M. (1975). A statistical study of ship domains. The Journal of navigation, 28(3), 328-344.

Harati-Mokhtari, A., Wall, A., Brooks, P., & Wang, J. (2007). Automatic Identification System (AIS): data reliability and human error implications. The Journal of Navigation, 60(3), 373-389.

Im, N., & Luong, T. N. (2019). Potential risk ship domain as a danger criterion for real-time ship collision risk evaluation. Ocean Engineering, 194, 106610.

International Maritime Organisation, (2018). Formal Safety Assessment. Retrieved from: https://wwwcdn.imo.org/localresources/en/OurWork/Safety/Documents/MSC-MEPC%202-Circ%2012-Rev%202.pdf

Kaluza, P., Kölzsch, A., Gastner, M. T., & Blasius, B. (2010). The complex network of global cargo ship movements. Journal of the Royal Society Interface, 7(48), 1093-1103.

Kim, D. H. (2020). Identification of collision risk factors perceived by ship operators in a vessel encounter situation. Ocean Engineering, 200, 107060.

MacDuff, T. (1974). The probability of vessel collisions. Ocean industry, 9(9).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).National Marine Electronics Association, (n.d.)

Pedersen, P. T. (1995). Collision and grounding mechanics. Proceedings of WEMT, 95(1995), 125-157.

Qu, X., Meng, Q., & Suyi, L. (2011). Ship collision risk assessment for the Singapore Strait. Accident Analysis & Prevention, 43(6), 2030-2036.

Silveira, P., Teixeira, A. P., Guedes Soares, C., & Santos, T. A. (2015). Assessment of ship collision estimation methods using AIS data. Maritime Technology and Engineering, 195-204.

Szlapczynski, R. (2006). A unified measure of collision risk derived from the concept of a ship domain. The Journal of navigation, 59(3), 477-490.

Szlapczynski, R., & Szlapczynska, J. (2016). An analysis of domain-based ship collision risk parameters. Ocean Engineering, 126, 47-56.

Tetreault, B. J. (2005, September). Use of the Automatic Identification System (AIS) for maritime domain awareness (MDA). In Proceedings of Oceans 2005 Mts/Ieee (pp. 1590-1594). IEEE.

United States Coast Guard (n.d.) Maritime Mobile Service Identity. Retrieved from: https://www.navcen.uscg.gov/?pageName=mtMmsi#format

Zhang, W., Goerlandt, F., Kujala, P., & Wang, Y. (2016). An advanced method for detecting possible near miss ship collisions from AIS data. *Ocean Engineering*, *124*, 141-156.