MSc Thesis

Nautical safety on Dutch inland waterways, a study to detect anomalous vessel behaviour based on AIS data

B.G.H. van Engelen

Delft University of Technology

TU Delft

Witteveen + Bos

MSc Thesis

Nautical safety on Dutch inland waterways, a study to detect anomalous vessel behaviour based on AIS data

by

B.G.H. van Engelen

to obtain the degree of Master of Science
at the Delft University of Technology,

Student number:	4399811	
Project duration:	December, 2022 – December, 2023	
Thesis committee:	Prof. dr. ir. M. van Koningsveld	TU Delft
	Ir. L. de Boom	Witteveen+Bos / TU Delft
	Dr. F. Baart	TU Delft
	Ir. S.E. van der Werff	TU Delft
	Dr. ir. P. Mares Nasarre	TU Delft

Cover: Tim Altmann Unsplash (Modified)
https://unsplash.com/@timaltmann

TU Delft Witteveen+Bos

Preface

This thesis marks the final stage of my study at the Delft University of Technology, as the last part of my master's to obtain the Master of Science degree in Civil Engineering within the master track Hydraulic Engineering. This research was carried out in collaboration with Witteveen+Bos by means of a graduate internship.

Selecting a thesis topic was a challenging journey, with numerous options to choose from. I doubted the amount of Python programming with my initially limited experience. However, I am grateful that I chose this topic, and I actually like the unlimited possibilities of coding when it works. As with many life choices, do not hesitate to long to take the step. In the end, you will get there.

Reaching the end of this process was only possible with the guidance and support of my committee. Firstly, I thank Lex de Boom as the daily supervisor and for introducing me to Witteveen+Bos. I gained valuable insights into the topic during the weekly meetings and entertaining moments with the haveninrichting en scheepvaartwegen group.

I also want to thank Fedor and Solange for their effort and weekly meetings to push me in the right direction. Their assistance with the code is very much appreciated; I could never have reached this end product. Fedor's introduction to DigiShape and the opportunity to present on the DigiShape day, hosted by Mark at van Oord, was a great experience. Mark offered valuable feedback and insights during the process, and his enthusiasm for the subject gave me more energy to continue. At last, I want to thank Patrica for joining the committee and providing valuable feedback at the end stage of the report.

Finally, I would like to thank the colleagues at the Rotterdam office for the joyful coffee moments and lunch walks. A special mention goes to my close circle: family, roommates, and friends, who provided constant support and fresh perspectives and, most importantly, helped me not overthink my thesis from time to time.

Bart van Engelen
Rotterdam, December 2023

Abstract

Inland shipping is an essential mode of transport in the Netherlands, representing 19% of the total volume transported. Substantial growth is anticipated in the sector, with estimated growth rates of 34% relative to 2014. Up to 2023, this growth is not necessarily observed in the number of vessels navigating the inland waters, but a trend towards larger vessels has been visible over the past years. With the amount of cargo transported over water and vessel size increasing, the pressure on nautical safety in Dutch inland waters is expected to grow over the coming years.

The current safety assessment method for inland waters relies on accidents reported in the Scheepsongevallendatabase (SOS-database) and is supplemented with expert opinions. From 2009 to 2022, an annual average of 1110 accidents was registered, with 14% categorised as significant accidents, indicating severe consequences. The current method is based on lagging indicators, relying on the examination of historical accident data. Alongside the retrospective approach, in the accident data, under-registration is evident, and expert views may be influenced by biases or personal interests. Consequently, essential qualitative and quantitative information, necessary for an objective and comprehensive assessment of nautical safety, is missing.

Conversely, much data on shipping activities exists in the form of Automatic Identification System (AIS) data, providing detailed trip information for nearly all vessels navigating Dutch inland waters, given the system its mandatory usage for vessels over 20 meters since 2016. This data source holds the potential to complement the information on nautical accidents. The principal aim of this research is to assess whether historical AIS data can fill in the gap in accident reporting on inland waterways, identifying unusual shipping patterns, and ultimately uncovering incidents or near misses to enhance insight into nautical safety.

In the proposed method, definitions of normal behaviour are established, and patterns deviating from this are identified as anomalous. Vessel behaviour is characterised by extracting features from vessel trips based on the AIS data logs. This characterisation include speed, acceleration, direction, manoeuvrability, and position-related features. The vessel behaviour per trip, defined by all individual features is subsequently reduced into two dimensions by the Uniform Manifold Approximation and Projection (UMAP) algorithm. This approach retains the underlying behaviour, preserving similarities in the original data within a 2-dimensional embedding. Following this, K-means clustering is used to generate clusters of vessels exhibiting similar behaviour patterns. The hyperparameters are chosen based on the elbow method assessed using multiple scoring metrics.

The methodology is applied in two cases of ship-infrastructure interaction, covering 2322 vessel trips at the Moerdijkbrug on the Hollands Diep and 2248 trips near the Schellingwouderbrug at the IJ. Across the clusters, similarities in main navigation direction and vessel paths are clearly observable in most instances. Furthermore, the same vessel types are predominantly clustered together. Examination of the individual underlying features reveals distinctions among the clusters, such as higher velocities observed in one cluster compared to others.

This approach helps in identifying vessel trajectories that deviate from expected norms, classifying them as potentially dangerous. Within this category, distinctions can be drawn among false alarms, actual accidents, and near misses, which are defined as just not accidents. For instance, in the context of bridge interactions, near misses might occur when the minimal distance with respect to the bridge pillar is only slightly larger than the distance between the AIS transmitter and the vessel's outermost boundary.

In the case of the Schellingwouderbrug, out of the 2248 trips, 20 were flagged as suspicious, leading to the identification of two potential accidents. This highlights the method's capacity to pinpoint accidents

by delving deeper into just 1% of the data. This efficiency underscores its capability to effectively target anomalous vessel behaviour and incidents, thereby enhancing insights into nautical safety. Such findings can complement the SOS-database, currently serving as the primary data source for nautical safety assessment in the Netherlands.

A notable observation is the influence of additional infrastructure elements on the outcomes of the clustering. The inclusion of a lock complex in the investigation of bridge interaction at the IJ region resulted in a different outcomes compared to the scenario without the lock. Hence, careful consideration should be given to defining the domain used for trajectory generation. Additionally, the length of trajectories impacts the evaluation method employed in this study.

Moreover, the quality of input data holds importance, particularly in preprocessing steps involving outlier removal. Eliminating outliers is crucial for generating credible vessel trajectories. Excessive smoothing might eliminate minor deviations in trajectories that could serve as indicators of potentially dangerous situations within the proposed methodology.

This study, however, excludes external hydraulic factors such as currents, waves and weather conditions, despite their known influence on shipping behaviour. Their incorporation is essential for a comprehensive analysis. Furthermore, while some steps have been made to scale the data in terms of time and location, future research should delve deeper into this aspect, given the scarcity of data on accidents and near misses. Additionally, this study underscores the necessity of transitioning toward a more proactive safety assessment approach, as the current approach remains retrospective.

The method's application to the Moerdijkbrug at the Hollands Diep and Schellingwouderbrug at the IJ revealed clusters displaying similar vessel behaviour, particularly in direction and paths. Several atypical patterns were observed and further investigated, analysing less than 1% of the data set in both cases. Within this limited portion of input trajectories, two distinct patterns were identified and classified as probable accidents in the IJ case.

In conclusion, the study suggests that the implementation of the proposed method to detect anomalous vessel behaviour based on AIS data holds the potential to enhance insight into nautical safety.

Contents

1 Introduction

”On November 2nd at 11:56, a potential disastrous incident at the Spijkenisserbrug in Rotterdam was observed. As a cargo laden sea vessel approached, the Spijkenisserbrug began to open but abruptly halted, leaving the movable bridge deck at the side of Spijkenisse opened for only one meter before it stopped. By a quick response from the control room the other part at the Hoogvliet side opened just in time. The vessel changed course and passed the bridge on the alternate side. Due to the alert response a collision is avoided!” (De Havenloods, 2023).

An article like the one above, describing an almost accident or near miss, is not easily found, as most articles focus on actual incidents. Yet it is reasonable to assume that numerous similar cases of near accidents occur on inland waters.

In this chapter, general background information on inland shipping in the Netherlands and nautical safety is presented. The problem is described together with the objective and the scope. Subsequently, the research question is introduced, and to conclude the chapter, an outline of the report is presented.

1.1. Nautical safety in the context of Dutch inland shipping

Inland waterway transport plays a vital role in the Netherlands transportation network. With over 6,000 kilometres of interconnected waterways linking cities and industrial regions across the country, transport over water is essential for both domestic transportation and international export.

In 2021, inland waterway transport accounted for almost 19% of the total transported volume. For outgoing cross-border transport, an even greater share, reaching 28%, is transported by inland vessels, particularly those travelling from the Port of Rotterdam to the hinterland (CBS, 2023c). To put these percentages into perspective in terms of vessel movements, approximately 7,500 weekly trips were conducted. When extrapolated to annual figures, this translates to roughly 400,000 movements for transportation via inland waterways alone (CBS, 2023b).

A study conducted for Rijkswaterstaat predicts a growth range of 18 to 34% in inland shipping, relative to the volume of goods transported in 2014. Furthermore, the same report expects an increase in recreational and passenger shipping (Rijkswaterstaat WVL, 2021). Until 2023, this growth does not necessarily affect the total number of vessels yet, but a trend towards larger ships has been visible over the past years (CBS, 2023a).

This growth in inland waterway transport can be attributed, in part, to initiatives of the European Commission, which have been designed to promote and provide incentives for this mode of transportation. These measures are intentionally formulated to make the transition from conventional modes of transport, such as road and rail, to the more environmentally friendly, per unit transported cargo, option of inland waterway transport easier. This shift significantly contributes to a reduction in energy consumption and emissions, aligning with environmental sustainability goals (European Commission, n.d.).

Besides environmental benefits, inland waterway transport is a relatively safe method of transport compared to other modalities (Rijkswaterstaat WVL, 2021). However, ensuring safety remains a critical

concern. Many measures are taken to improve or guarantee nautical safety. Regulations in the Netherlands are defined in the Scheepvaartverkeerswet (2021), which has to be followed by all waterway users. Safety, traffic flow and reduction of pollution are important aspects of the regulations. The traffic rules are stated in six different shipping regulations for the inland waterways and one for the coastal waters, containing rules on priority and the meaning of traffic signs (Rijkswaterstaat, 2023c).

Despite the rules and regulations, accidents on the waterways are unavoidable and have a broad variety. Two vessels colliding, a collision between a vessel and waterway infrastructure, motor failure, fires, and minor accidents aboard are all examples of accidents occurring on or around waterways.

Additionally, frequent dangerous situations arise on waterways, such as vessels crossing paths or overtaking, leading to unsafe conditions. Such situations can be categorised as a 'near miss' if the hazardous condition nearly escalates into a full accident. Various definitions on near misses are found in the literature, according to Al Shaaili et al. (2023) a near-miss can be defined as "an incident that is close to happening and has the potential to cause harm or damage but lacks a few of the ingredients or generating mechanisms that may eventually lead to an accident". For example, when two vessels cross on a canal and come dangerously close to each other, this can be considered as a near miss, in most cases, the near miss is not reported, both vessels typically continue their journeys.

However, near misses can be of great importance in nautical safety, assuming the common cause hypothesis holds, which suggests that causal pathways leading to near misses are similar to those of actual accidents (Wright & Van Der Schaaf, 2004).

With the amount of cargo transported over water, vessel size increasing, and climate change resulting in more draughts and less navigational space, the pressure on nautical safety in Dutch inland waters is expected to grow over the coming years.

1.2. Problem description

To gain insight in the state of safety of a process, transportation mode or business sector safety assessments are conducted. Numerous standard models exist to assess safety, the most common methods used in maritime context are discussed in the work by Huang et al. (2023). This work provides a comprehensive review of maritime risk literature from 2002 to 2021. The study analysed risk assessment methods, giving insight into frequently used methods.

A classic model in risk assessment, as mentioned in this work, is the bow-tie model by Nielsen (1971). In this model, hazardous events are positioned at the centre, their causes on the left side, and consequences on the right side. This arrangement forms a graphic resembling a bow-tie, as illustrated in Figure 1.1.

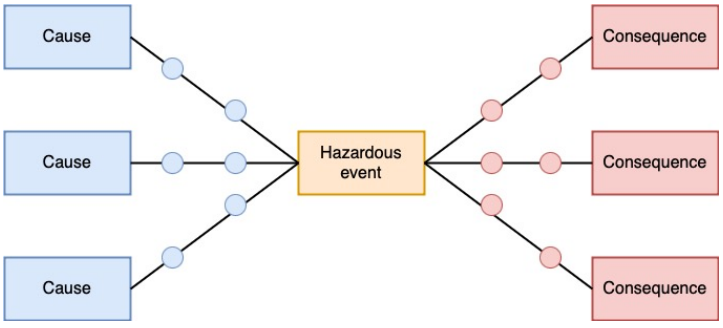


Figure 1.1: Schematic representation of bow-tie diagram used in safety assessment (Nielsen, 1971)

In a bow-tie model, the events leading to a hazardous event are referred to as leading indicators, incorporating safety checks and inspections. Those following the event are known as lagging indicators, involving the accident or incident that occurred. Proactive safety measures are positioned on the left side of the diagram, while reactive measures are on the right side, indicated by the circles in the diagram (Ehlers et al., 2017). The concept of leading and lagging indicators is widely used in safety assessments

as indicated by Dilkhaz (2019). Within this context, the method used to determine nautical safety on Dutch inland waters will be evaluated.

1.2.1. Method to determine nautical safety on Dutch inland waters

The current method to determine nautical safety on inland waters does not consistently give a complete or accurate perspective. The status of nautical safety is described in the Monitor Nautische Veiligheid, which is based on reported accidents in the Scheepsongevallendatabase (SOS-database) and expert opinions. In the context of the bow-tie diagram this relies on lagging indicators, the accidents already occurred with their associated consequences. These are valuable for assessing past safety performance but may not provide a comprehensive view of safety on their own, since these data sources suffer from incompleteness or inconsistencies (Hofmeijer, 2020; Rijkswaterstaat, 2023b).

The SOS-database, a national shipping accident database managed by Rijkswaterstaat. This contains information on shipping accidents and other incidents on water that have taken place within the Netherlands. Input for the database comes from reports filed by the involved skipper or the responsible waterway manager. However, this manual input method is prone to errors and may lack completeness. An example is the exact location of occurrence, nearby a bridge the Global Positioning System (GPS) coordinates of an accident could be registered, but the precise orientation of the passage is complex to determine afterwards.

On the other hand, there is significant under-registration concerning shipping accidents, in most cases less severe accidents. During bridge inspections, damages are occasionally discovered that do not correspond to known accidents, indicating potential under-registration (Beenhakker & Schelling, 2020). Additionally, near misses, as mentioned earlier often go unreported. This could be of great importance in the overall insight into nautical safety. Besides the incomplete or missing input files, managing and combining the reported accidents in the database is a human job, which can introduce faulty documentation.

Furthermore, the Monitor Nautische Veiligheid also relies on expert opinions, a common practise in safety assessments (Sklet, 2004). Information is collected during expert meetings, involving participants from Rijkswaterstaat, the police, port authorities and relevant professional associations who give their views based on field experiences. However, it should be noted that these opinions can be biased. In the context of the bow-tie diagram, the expert opinions can be seen as leading indicators when insight and predictions about potential safety issues are provided. On the other hand, can be viewed as lagging indicators when consulted after with the incident being analysed, and opinions on what went wrong being provided for future improvement.

In summary, reliable information as input for the safety assessment, both qualitative and quantitative, is lacking, which is essential for an objective and reliable assessment of nautical safety. Given that subsequent actions and safety measures are based on the Monitor Nautische Veiligheid, it is essential that this assessment provides a comprehensive view of the current safety status.

1.2.2. AIS data to fill information/ AIS data to complement missing information

With the increasing availability of data, accident data is used in many safety assessments across various mobility domains, such as the railway sector. However, not all assessments have to start with accident data, where in the the work by Di Ciccio et al. (2016) flight data is used to detect anomalous trajectories of air planes, which is used to predict diversions in freight transportation, these anomalous patterns can provide insight in safety related issues as well.

Whereas information on shipping accidents is missing in this case, significant shipping data can be accessed through the Automatic Identification System (AIS). AIS is a system designed to enhance the safety of navigation; AIS facilitates communication between skippers and traffic stations. The device automatically transmits radio waves at regular intervals based on the vessel's speed. The radio waves carry location, speed and ship data related to the voyage. AIS devices automatically receive all information broadcast by other AIS devices on other ships and ashore within transmission range. Since

2016, the system has been required on all commercial vessels and recreational craft with a length of over 20 meters navigating Dutch inland waters (Rijkswaterstaat, 2022).

Since the introduction of AIS, initially for seagoing vessels and later for inland vessels, numerous studies have been conducted utilising this data across various domains such as marine environments, security, safety and other areas. Fournier et al. (2018) summarised research carried out from 2004 to 2016, indicating that in coastal waters, probabilistic safety assessments and statistical analysis based on AIS data are commonly employed to estimate the frequency and risk of collisions and groundings. These studies hold promise for application in inland waters.

1.3. Research objective and scope

The primary objective of this project is to determine if the use of historical AIS data can fill in the gap in accident reporting on inland waterways, accurately determine the accident occurrences, search for anomalous shipping patters, and ultimately identify incidents or near misses to enhance insight into nautical safety.

To achieve this, the AIS data should be transformed into patterns of shipping behaviour. In these patterns, anomalous activities are searched for, which could indicate potential accidents or near misses. With this additional information on accidents, the SOS-database can be supplemented, and a more complete view of nautical safety can be generated. It becomes possible to conduct in-depth investigations into dangerous situations or areas not currently classified by the existing method.

The research primarily focuses on three main accident types, as defined by Rijkswaterstaat. These include single-vessel accidents, specifically, ship-object and ship-infrastructure incidents, which will be evaluated in depth. On the other hand, there are accidents where two vessels are involved. While the typical behaviour descriptions of vessels involved in these accidents will be discussed, the ship-ship accidents will not be incorporated in the modelling steps. Detailed definitions of the accident types, as introduced by Rijkswaterstaat, can be found in Section 2.2.

This study excludes certain accident types, such as fires, explosions on board, puncturing and sinking. While these categories are part of the current safety determination method, they are excluded from this study due to the absence of a clear correlation between these accidents and specific location. Rijkswaterstaat categorises accidents like fires as non-shipping accidents, including near misses in this category. However, near misses are addressed and considered in this research.

Unlike the non-shipping accidents, incidents involving ship-ship, ship-infrastructure and ship-object collisions often exhibit correlations with specific locations. These type of accidents can potentially be mitigated through navigation rules or modifications in the waterway layout. However, it is important to note that that this is not always the case, as occurrences within these categories might result from issues like loss of steering due to poor maintenance or a fire for example.

Moreover, external hydraulic influences, such as currents or waves, impact shipping behaviour. Additionally, weather conditions like wind, rain, and fog should also be considered, as they have implications for shipping behaviour according to Shu et al. (2017). However, these factors are outside the scope of the current project.

1.4. Research question

In order to achieve the stated research objective, the following research question has been formulated:

"How can insight into **nautical safety** on **Dutch inland waterways** be improved by the detection of **anomalous vessel behaviour** based on **AIS data**?"

To come to an answer to the main research question, four sub-questions are defined:

1. "What defines the current state of nautical safety on Dutch inland waterways, how is this measured, and what aspects could be enhanced for improved safety assessment?"

2. "What is the definition of "normal" and "unusual" vessel behaviour, and how can near misses be defined in this context?"

3. "How can the transition from AIS data to vessel behaviour be made to identify anomalous behaviour, including near misses?"

4. "What is the added value of the proposed method for detecting anomalous vessel behaviour compared to the existing method, and how does it enhance insight into nautical safety?"

1.5. Report outline

The report is subdivided into three parts with nine chapters, as presented in the Table below. This table offers an overview of the report's structure and the individual sub-questions addressed in each chapter.

	Chapter 1: Introduction	
Part I	**Materials and Methods**	
	Chapter 2: Nautical safety on Dutch inland waters	sub-question 1
	Chapter 3: Conceptual framework and background on methods	sub-question 2 & 3
	Chapter 4: Available AIS data and its characteristics	
	Chapter 5: Implementation of method to detect anomalous vessel behaviour	
Part II	**Results**	
	Chapter 6: Results	sub-question 4
Part III	**Discussion, Conclusion and Recommendations**	
	Chapter 7: Discussion	
	Chapter 8: Conclusions	
	Chapter 9: Recommendations	

Part I

Materials and Method

2

Nautical safety on Dutch inland waters

Chapter two provides background information on the Dutch inland waterway system and its users. Followed by an in-depth description of the current method used to determine and quantify safety, including definitions and types of accidents.

The chapter proceeds with an evaluation of the SOS-database, covering the method of data collection and assessing data quality, supplemented with illustrative examples. Historical accident data from inland waters, as well as near misses, are presented, based on information from the SOS-database. To conclude the chapter, the limitations of the current safety assessment method are considered.

2.1. Shipping on Dutch inland waters

As a basis for a safety study of the Dutch inland waters, understanding the system and the users is needed. The Netherlands has an extensive waterway system of 6.3 thousand kilometres, primarily concentrated in the west and north of the country. These waterways were historically created to serve the need for agricultural land or the drainage of low and high-moorland areas for peat production. Connecting these canals and waterways to existing rivers to drain excess water resulted in an extensive network of channels and rivers (CBS, 2022).

The central transport axis consist of; the major rivers, the Amsterdam-Rhine Canal and the Rhine-Scheldt connection are mainly used by inland vessels for the transit of goods from the ports of Amsterdam and Rotterdam to the hinterland, Germany and Belgium. This main transport axis significantly contributes to the overall accessibility and efficiency of waterborne transport throughout the country (CBS, 2022).

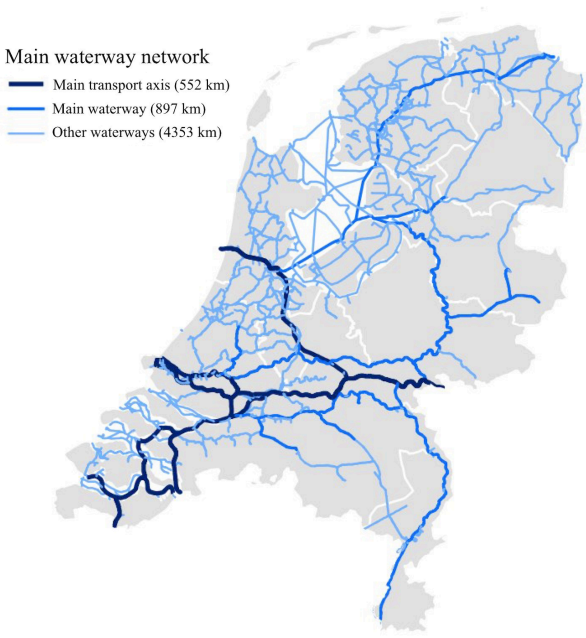


Figure 2.1: Main waterway network (CBS, 2022)

2.1.1. Waterway users

In addition to the vessels for inland waterway transport, various other types of vessels navigate Dutch rivers and canals. The main categories are seagoing, cruise, ferries, and recreational vessels. Marine vessels are found near major sea harbours like Rotterdam and Amsterdam. Cruise vessels, often navigating the larger rivers, Rhine or Meuse. Ferries are used where bridges or other connections between two land parts are not possible or economical; this can be the case for rivers, channels or lakes. Varying in size from small vessels for pedestrians and bikes towards larger ships carrying cars or trucks.

The last category, the recreational vessels, differ significantly from the others in size and the skippers' experience. The combination of relatively small recreational boats and other waterway users can lead to dangerous situations; small ships are not well visible from the perspective of a larger vessel. Besides, the skippers on recreational vessels are, in most cases, less experienced and can make unexpected choices and manoeuvres. With about 400,000 recreational boats, this is a serious group of waterway users, and extra attention is needed, especially when commercial and recreational vessels share a waterway (Waterrecreatie Nederland, 2016).

To increase the safety of recreational boating, specific parts of the waterways are allocated for pleasure craft. This segregation is aimed at minimising the potential for accidents between these two categories of vessels. In the 'Varen doe je samen!' project, Rijkswaterstaat collaborates with various partners to improve safety in interactions between pleasure and commercial shipping. The primary focus lies in creating awareness among professional skippers and recreational skippers, about each other's behaviour and limitations on the water. Additionally, the campaign identifies potential bottlenecks on waterways, and safe routes are suggested (Rijkswaterstaat, 2023a).

Since the vessels navigating the inland waters have their typical characteristics and specific sailing areas in the system, a distinction in vessel type should be made when usual or unusual patterns are defined. Furthermore, potential interesting locations to set up the model can be determined based on the knowledge of the main waterways.

2.2. Current method of nautical safety determination

To enhance insight into nautical safety, it is essential to understand the current approach to its determination. First, definitions used in the present studies are addressed, followed by an exploration of the risk determination process.

Periodically, on behalf of Rijkswaterstaat studies are conducted to gain insight into nautical safety, with a specific focus on different accident types. The Monitor Nautische Veiligheid study covers accidents related to ship-ship and ship-object collisions. Ship-object accidents involve collisions with jetties, buoys, and quay walls. The study is divided into two parts, one concentrating on the North Sea the other addressing inland waterways. Additionally, regional studies are also undertaken.

Accidents involving collisions with ship-infrastructure, such as bridges, locks, weirs, and barriers, are addressed in a separate study (Beenhakker & Schelling, 2020). An overview of the distinctions and description of these accident types as classified by Rijkswaterstaat is provided in Table 2.1.

Table 2.1: Accident types as defined by Rijkswaterstaat (Hofmeijer, 2020)

Type	Example
Ship-ship	Collision between two or more vessels.
Ship-object	Collision with a buoy, mooring dolphin, ice floe or other objects in the water
Ship-infrastructure	Collision with sluices, bridges, weirs, embankments or groundings
Non-shipping	Fire or explosions on board, sinking, puncturing or near miss

Engineering and consultancy firm Witteveen+Bos conducted the latest edition on inland waterways, covering 2009 to 2018. In this report, nautical safety is defined as "the extent to which the risks of maritime accidents are controlled to an acceptable and preferably negligible level". The nautical safety risk is described as the probability of the occurrence of a shipping accident times the effect of the accident (Hofmeijer, 2020).

The report defines a shipping accident as an accident on the water in which unintentional damage occurs and at least one sailing or stationary vessel is involved. Damage in this context is described as casualties, damage to one or more ships involved, damage to infrastructure or objects, environmental damage, blockage of a waterway, and residual damage such as loss of time and other damage. Accidents are defined in two different degrees, shipping accidents and significant shipping accidents, where the

impact of a significant accident is more severe. When one or more consequences in Table 2.2 are met, the label significant is applied.

Table 2.2: Criteria of significant shipping accident (Hofmeijer, 2020)

Consequence	Description
Casualty	Dead, missing or severely injured.
Waterway damage	Immediately (within 7 days) after the accident action needed to repair infrastructure or object.
Vessel damage	A vessel involved in the accident can no longer sail or is not allowed to sail.
Cargo damage	In the event of 10 tonnes of cargo or more or the loss of one container.
Environmental damage	As a result of an accident, chemicals (packed or unpacked) or oil (fuel or cargo) spilled into the water. Visible consequences such as fish mortality.
Blockage	Complete blockage of the waterway of 1 hour or more.

2.2.1. Method of risk determination ship-ship and ship-object

The latest Monitor Nautische Veiligheid by Hofmeijer (2020) states a table of the ten most important risks on inland waters as outcome of the report. Table 2.3 includes the risks, classification, and typical situations or locations of occurrence. This list is made up based on accidents registered over the years 2009 to 2018 in the SOS-database. Additionally to the data, input from experts is used to complete the list and define the top 10 risks. Classification of the risk is from very high, high to medium.

Table 2.3: Top 10 risks outcome classified in latest Monitor Nautische Veiligheid (Hofmeijer, 2020)

#	Risk	Classification	Situation
1	Collision recreational and commercial	Very high	Bifurcations and intersections with high shipping intensity, high-speed shipping areas and waterways shared by commercial and recreational shipping
2	Single-vessel accident recreational	High	Open waters
3	Accident service vessel (sinking/capsizing)	High	Harbour areas
4	Accident ferry	High	Places where ferries sail, cable ferries pose an additional risk
5	Accident passenger vessel	High	Open waters and rivers
6	Collision with swimmer	High	No specific situation
7	Collision recreational-recreational	Medium / High	Bifurcations and intersections with high shipping intensity
8	Collision commercial-commercial	Medium	Junctions and intersections with high shipping intensity
9	Grounding commercial vessel	Medium	National waters
10	Accident commercial vessel (sinking/-capsizing)	Medium	Open waters

The following method is used to identify and rank the risks; all accidents in the database are scored in three categories, namely:

- Safety, health and society;
- Environmental damage;
- Economic loss: to Rijkswaterstaat, vessel owner and/or due to waterway blockage.

In the three categories, an accident is ranked in an effect class, where the effect class represents the impact severity. The effect classes correspond with an effect score, the small impact gets zero points, and the highest impact results in a effect score of 100,000.

Within the safety, health and society category, the lowest effect score corresponds to no casualties and the highest to multiple deaths or missing. If applicable, one accident can score in all categories, and the total score will be the summation of the three scores. Scores for financial and environmental damage are derived from available descriptions, which may not always be complete or accurate. The complete table used to score the accidents can be found in Appendix B.

After scoring each case, the accidents are grouped into risk categories; for example, all collisions between two commercial ships are combined in one group. The type with the highest score can be determined by summing all individual accidents in a group.

Applying this method results in the highest risk score for a collision between recreational and commercial vessels, indicated in Table 2.3. In this overview, recreational vessels are involved in four of the ten risks, a significant contribution. Moreover, the classification of recreational vessel accidents is high to very high. An explanation can be the relatively high probability of severe casualties on the smaller boats. The latter is also why a collision with a swimmer is in the top 10, it does not occur often, but when it does, the chances of serious injury are significant.

2.2.2. Method of risk determination ship-infrastructure

To completely analyse the safety of waterways, it is essential to consider shipping accidents that are not directly related to other traffic. Infrastructure collisions can have a severe impact on inland shipping. For instance, the accident with the weir in Grave in 2019, where a motor tanker rammed the weir with a lowering of the water level at a large part of the river Meuse as a result. This incident resulted in navigation problems for vessels on the river (De Gelderlander, 2016).

From one perspective, infrastructure is essential for maintaining navigability and providing land connectivity through bridges. However, structures on and around the waterway can influence safety. Visibility along a waterway can be limited near bridges or special manoeuvres should be carried out when passing through locks, potentially leading to dangerous situations. Therefore, it is crucial to account for the interaction with infrastructure in safety studies.

Arcadis studied this part of nautical safety. In this study, an analysis is made of collisions between ships and bridges, locks, weirs and barriers. Table 2.4 gives an overview of the ship-infrastructure accident types and their corresponding classification (Beenhakker & Schelling, 2020).

Table 2.4: Risks ship-infrastructure by Beenhakker and Schelling (2020)

#	Risk	Classification
1	Bridge collisions	High / Very high
2	Lock collisions	High
3	Weir collisions	Medium
4	Barrier collisions	Medium

This study used a method similar to the Monitor Nautische Veiligheid, which relied on accidents reported in the SOS-database and expert's insights. Scores ranging from 0 to 100,000 were used to evaluate the three categories previously mentioned.

From 2009 to 2018, 528 collisions with bridges were reported. However, experts believe that the actual number is much higher. During bridge inspections, considerably more damage is found than reported; these are minor damages in most cases. Over the same period, there were 333 reports of collisions with locks, the second-largest group. Based on the damages found, it is expected that there is also a significant amount of under-registration for ship-infrastructure accidents.

The risk of fatal accidents in most ship-infrastructure accidents is relatively low, resulting in lower risk scores compared to ship-ship collisions. These scores are primarily based on economic losses or blockages

of the waterway, resulting in overall lower risk level when compared to the most critical risks presented in Table 2.3. In this case, the scores are assigned based on known economic and environmental damages.

When evaluating the actual risk scores, it is noteworthy that collision between commercial vessels and bridges, locks or weirs are ranked second, below collision involving commercial and recreational vessels. This underscored the importance of incorporating these incidents into the assessment of nautical safety (Hofmeijer, 2020).

2.2.3. Shipping accident registration in the SOS-database

Rijkswaterstaat manages the national shipping accidents database, called the SOS-database. A public version of this database is available where a part of the data is provided, the rest is restricted (Rijkswaterstaat, 2023b). The database contains information on shipping accidents and other incidents on water that occurred within the jurisdiction of the Netherlands. In this section, the recording of accidents in the database, the database format, and the quality and reliability of reported incidents is discussed. To illustrate this evaluation, three incidents are reviewed, one involving a collision with a buoy and two instances of groundings.

Format of the data entry

The database is filled with reports submitted by skippers, waterway authorities or the police. The accidents must all be reported using a specific form, known as the SOS-form. This document contains numerous questions, from the cause and (technical) circumstances to the damage and victims. To ensure consistency, Rijkswaterstaat published a guiding manual with explanatory notes on the form. This document contains explanations of definitions used in the form and codes that must be used to fill in the form (Rijkswaterstaat, 2018).

Codes are provided for almost every question on the form, covering various aspects, from the vessel types involved in the accidents to the nautical administrator of the waterway. Completing the form with all these specific codes is expected to result in a comprehensive and accurate overview of the accidents. Nonetheless, it should be noted that accurately filling out the form can be time-consuming, which could potentially demotivate skippers when submitting accident reports.

However, the accurate filing of incident reports and the identification of the correct method for doing so are noted as common challenges in literature. An example is presented by Ferroli et al. (2012), where a model of accident reporting from aviation is translated to a healthcare environment. The primary conclusion is that everybody has to corporate, and a shift in mindset is important, transitioning from direct blame to the objective description of the situation. An analytical approach is believed to reinforce the prevention of recurring mistakes in incident reporting.

Quality and reliability of the data

The extensive SOS form aims to gain complete oversight of an accident. Within this section, an assessment on the reliability of three accident reports will be conducted. These incidents were selected because their dates align with the available AIS data use din this research.

One incident, documented in the public SOS-Database in October 2019, involves a ship-object accident. Objects on the waterway near the reported location are buoys HD 44 and HD 46; the assumption is that these buoys should be involved in the accident.

The following plot in Figure 2.2 shows the location of the accident as registered in the database and the locations of the buoys. This shows a significant distance between the buoys and the accident data point, of 550 and 1340 meters, to buoy HD 44 and buoy HD 46, respectively.

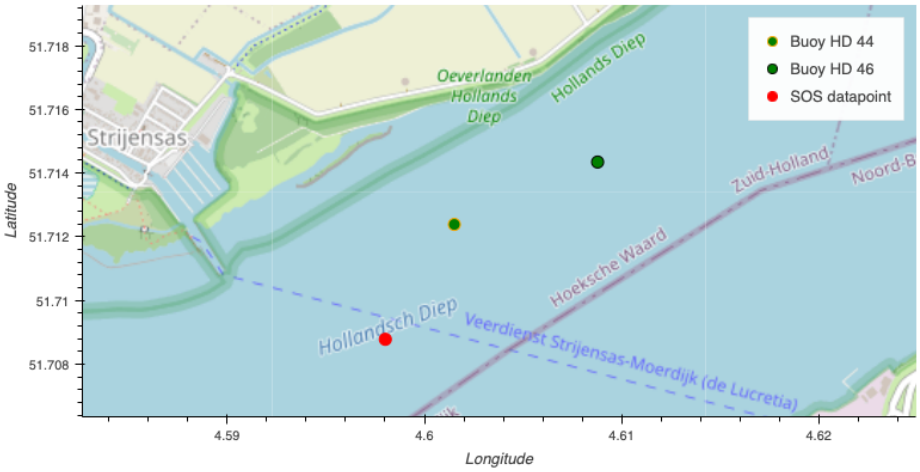


Figure 2.2: Plot of SOS data point and buoys Hollands Diep (Rijkswaterstaat, 2023b)

Buoys are not entirely stationary as they are anchored with a chain to the riverbed and can experience small displacement due to currents and water levels. However, this displacement is significantly smaller than the distance to the recorded point. This indicates that the incidents reported in the SOS database may not always be accurate.

To gain more insight into the quality, ship-infrastructure accidents are examined, in this case, collision with a bank. Unlike the buoys, the bank is a fixed point along the waterway, so the exact location of occurrence should be known.

Two examples are taken from the SOS-Database; the first accident was filed in October 2019 and plotted as SOS data point 1. The second, registered in July 2019, is plotted as SOS data point 2 in Figure 2.3. Both involve recreational vessels colliding with a bank, and the accident is marked as a significant shipping accident. These examples are chosen since the dates correspond with the available AIS data.

Figure 2.3: Plot of SOS data points Hollands Diep (Rijkswaterstaat, 2023b)

Since the riverbanks are fixed, it is expected that the accidents are registered near or directly in front of the banks. However, a closer look at the plot reveals that both data points are far from the bank. Data point 1 is situated 150 meters away from the nearest shore, while data point 2 is as far as 550 meters from the shore. These examples indicate potential issues with the reliability of location data in the database, as they lack a logical location description.

2.3. Analysis of reported accidents

In this section, a better understanding of the current status on nautical safety is explored based on the reported accidents in the SOS-database. Trends in the numbers of accidents and the frequency of reported near misses are examined.

2.3.1. Statistics on shipping accidents on Dutch inland waters

Between 2009 and 2018, input years for the latest Monitor Nautische Veiligheid a total of 10,768 shipping accidents were reported within Dutch inland waterways and seaports, excluding Dutch coastal waters (Hofmeijer, 2019). This resulted in an annual average of approximately 1100 accidents being registered. Among these incidents, 14% were classified as significant shipping accidents, based on the previously defined criteria. This corresponds to a range of 114 to 176 significant accidents reported each year. With the latest data on accidents, a peak of 188 significant accidents is revealed in 2022, as depicted in Figure 2.4.

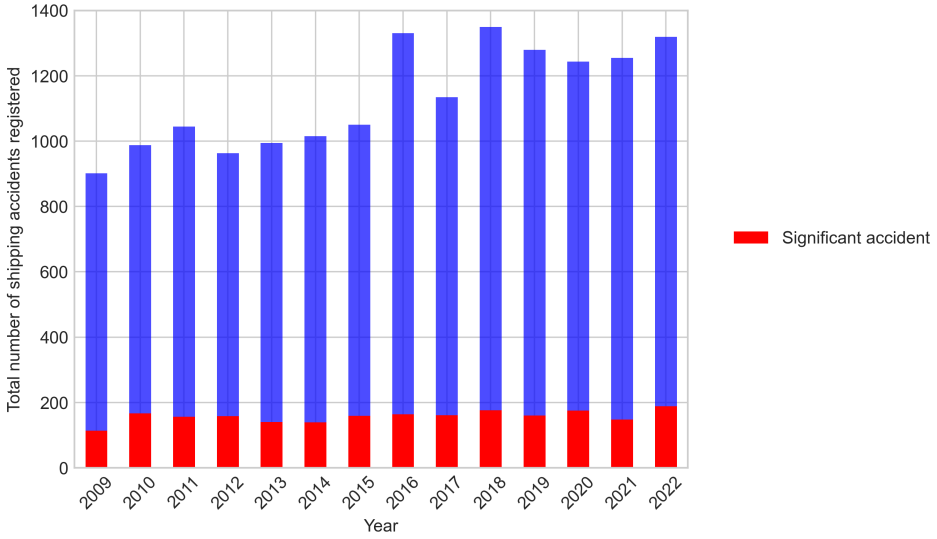


Figure 2.4: Total shipping accidents reported from 2009 until 2022 (Harbers, 2023; Hofmeijer, 2019)

A slight overall increase in registered accidents is recognised in Figure 2.4, with variations observed on annual basis. This trend is also visible in significant shipping accidents. The rise in registered accidents may be because the increased attention is given to record-keeping. For instance, Rijkswaterstaat is actively promoting a campaign to improve registration rates. Rijkswaterstaat has introduced campaigns such as 'Varen doe je Samen!' to increase safety. Part of the campaign is the Vaar Melder-app, where waterway users can report unsafe situations in an application to create a more complete view of the situation on inland waters, However, these locations are not included in the SOS database yet (Varen doe je Samen, 2017).

Another observation is that the increase in significant shipping accidents, although present, is relatively modest. Significant accidents are known to exhibit higher registration rates compared to non-significant accidents due to their greater impact. These incidents typically involve the participation of emergency services and the recording of a more extensive set of information (Hofmeijer, 2019).

From 2009 to 2018, no significant increase in traffic intensity of inland navigation was observed. In recreational navigation, there was a slight decrease in intensity. This information is derived from counts conducted at various locks and counting points along the waterways. As a result, it can be concluded that the increase in registered accidents does not result from an increase in vessel movements at these specific locations(Hofmeijer, 2019).

To obtain a complete overview, estimations are necessary to calculate the total number of accidents annually. It is estimated that around 31% to 35% of the accidents were registered between 2009 and 2012

(Movares, 2013b). Furthermore, the distribution of registration grade across the waterways is uneven, making it challenging to gain a complete understanding of the total number of accidents. Rijkswaterstaat suggests that the slight increase in registered accidents over the years may be attributed to awareness campaigns.

2.3.2. Near misses on Dutch inland waters

A near miss, as defined by Wright and Van Der Schaaf (2004), refers to an event in which unwanted consequences were prevented because a recovery occurred through the identification and correction of a failure, either planned or unplanned. Near misses can be of great importance in nautical safety, assuming the common cause hypothesis holds, which suggests that causal pathways leading to near misses are similar to those of actual accidents.

Near misses are categorised non-shipping accidents in the SOS-data from Rijkswaterstaat, and they are assigned their unique incident code for use on the SOS form. The information on near misses can be essential to give insight into dangerous locations. Examples of near misses reported in the SOS-database include instances where vessels sailed on the wrong side of the waterway, unsafe crossing or overtaking manoeuvres that narrowly avoided a collision.

In the SOS-data from 2009 to 2018, 9473 non-shipping accidents were reported, which corresponds to 47% of all unique reports. Approximately 10% of these non-shipping accidents are near misses. This places near misses as the third most frequent type of non-shipping accidents, following engine problems (the largest group by a significant margin) and the category of other issues. As illustrated in Figure 2.5, the distribution of registered near misses as a proportion of the total non-shipping accidents has increased over the years, ranging from 30 reported in 2009 to slightly over 200 in 2016. This trend aligns with the overall increase in the total number of non-shipping accidents.

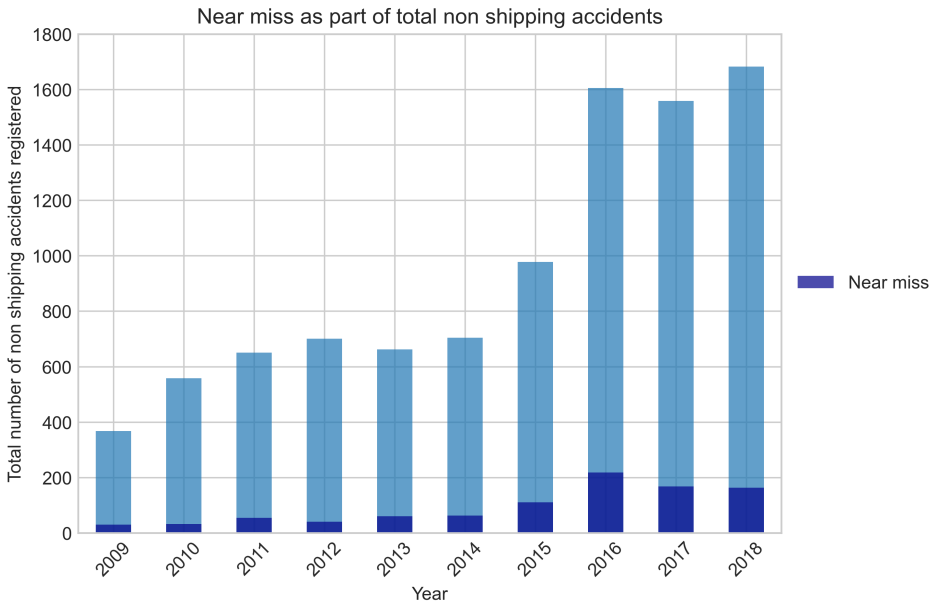


Figure 2.5: Total non-shipping accidents including near misses registered from 2009 until 2018 (Hofmeijer, 2019)

2.4. Limitations of nautical safety assessment

The quantitative method of risk determination described in Subsection 2.2.1 is a consistent method which gives an indication of the number of accidents that occurred and the related impact. Trends over the past years could be discovered but the analysis does not give a completely reliable picture, there are some drawbacks to mention:

- The main concerns have to do with the quality of the data. Not all accidents are reported

and recorded, resulting in under-registration in the SOS-database.it is expected that the missing registrations are mainly situations with minor consequences. Besides, the registration discipline of the employees filling the SOS-database is not optimal, resulting in accidents recorded under an incorrect risk or incomplete information as seen in Subsection 2.2.3;

- Furthermore, risk scores are automatically derived from the recorded data. If data is missing or incomplete, this may result in scores that do not correspond to the damage that actually occurred;
- Accidents with more serious effects are relatively rare in practice. With a low number of registrations, coincidence can play a significant role. If no significant accidents have occurred in recent years, it cannot be concluded that a shipping accident will not occur in the future at a specific location. The other way around, if a serious accident has happened, it does not necessarily mean that it will occur in the future. Therefore, specific locations cannot be definitely marked as safe or unsafe when accidents are not reported;
- The risk score for financial and environmental damage is estimated by Rijkswaterstaat based on the known descriptions. The actual damage usually becomes clear after registration of the shipping accident and no correction takes place afterwards. As a result, the severity of shipping accidents cannot always be properly estimated.
- Another disadvantage is when the studies are conducted, and the results are presented. The last Monitor Nautische Veiligheid was released in 2020 and is based on data until 2018, lagging two years behind. When measures are carried out based on the conclusions of such a report, it could be too late to prevent an accident in the intervening period as the intensity of inland shipping increases.
- Applying the method to a location with numerous minor accidents or near misses at one place can yield a relatively low score. In contrast, when an area experiences a single severe accident, in most cases well documented, it receives a high score. However, this approach tends to overlook all minor accidents. These little accidents can serve as early warnings, the leading indicators, of potentially more severe accidents in the future, so this should not be disregarded. Besides, accidents are not frequent but extreme, so you may not observe them in limited data sets.
- Finally, the guideline states that the accident has to be registered within three months after occurrence. In these months, valuable information may be lost simply because the skipper forgets the details of the accident.

3
Conceptual framework and background on methods

This chapter focuses on the development of the method to overcome the current limitations in the nautical safety assessment. To generate a more complete overview of the status of nautical safety, the limitations mentioned in Section 2.4 should be overcome.

As mentioned before, the current problem is characterised by two main aspects: incomplete accident data and insufficient data quality. Improving data quality requires precise and consistent incident descriptions. This can be achieved by requiring that incident reports be completed by skippers immediately after the event, rather than waiting for three months. Real-time reporting and continuous updates on the data base should be implemented to address reporting delays and capture delayed consequences.

To address the incomplete accident data, the analysis starts from vessel behaviour. This behaviour will be defined from AIS data, which is widely available on vessels navigating inland waters. Since 2016, the system has been required on all commercial vessels and recreational craft with a length of over 20 meters (Rijkswaterstaat, 2022). Any deviations from expected behaviour will be identified as unusual. Unusual behaviour may include false alarms, near misses, or actual accidents.

For this purpose, a clear understanding of vessel behaviour is needed. Definitions of vessel behaviour will serve as the basis of the analysis using AIS data. The following sections will provide an overview of the methodological steps and elaborate on the locations where the method will be implemented and tested.

No further recommendations on the actual method of scoring the accidents will be made in this report. Still, the focus will be on generating a more complete picture of the accidents and near misses.

3.1. Definitions of shipping behaviour

To start, a definition of the standard situation is needed to determine anomalies, which refer to something that deviates from what is standard, normal or expected. Definitions of normal shipping behaviour will be stated, split into three main categories based on the main accident categories defined by Rijkswaterstaat, since theses types of accidents are all covered in separate safety assessments. Ship-object and ship-infrastructure are combined in this study due to their similarities in behaviour.

3.1.1. Shipping behaviour in general

Vessels, in general, are expected to follow a smooth trajectory during their journey. This becomes more challenging on inland waters since there is less space to navigate, especially in busy waterways where many vessels interact.

The expected behaviour on a waterway can be defined with the following basic waterway rules as stated in the Binnenvaartpolitiereglement (2017). Vessels are expected to keep as much as possible on the waterway's starboard, right, side. Course and speed should be adjusted in time before prioritising another vessel. At this moment, it must be clear which course the vessel is on, and enough room must be given to each other to manoeuvre. Speed and course should be kept the same.

Vessels navigating on the incorrect side of the waterway could be marked as anomalous. Nevertheless, some exceptions need to be taken into account. For example, laden vessels travelling upstream in a river may intentionally select the inside of a bend to mitigate the impact of strong currents, as depicted in Figure 3.1. Moreover, vessels with fluctuating speeds or those making significant course changes during their journey can be of interest, as such behaviour is considered deviant. An example is given by Mestl et al. (2016) where extreme values of the rate of turn (ROT) where found in AIS logs prior to an recorded accident.

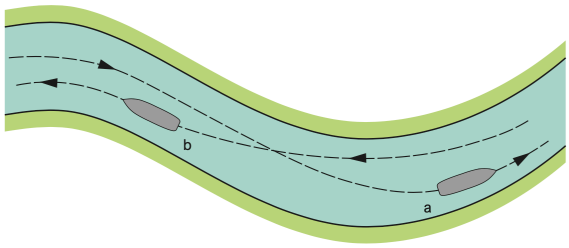


Figure 3.1: Course of vessel sailing downstream (a) and upstream (b) (Van Koningsveld et al., 2021)

Numerous exceptions to the definition above exist, influenced by factors such as vessel type, infrastructure interactions, and waterway dimensions variations. Additionally, manoeuvres such as overtaking, head-on encounters or crossing of vessels can significantly impact expected vessel behaviour. The anticipated behaviour for each vessel type and interaction with infrastructure will be explored in the upcoming subsections.

Behaviour per vessel type

Various types of vessels are expected to exhibit distinct behaviours when navigating waterways. These differences arise not only from variations in size, which affects manoeuvrability or speed but also from designated navigation areas for specific vessel categories, as discussed in Subsection 2.1.1. There are areas where inland vessels and recreational vessels are separated for safety reasons. A distinction between vessel types is made to define the expected behaviour of each specific type on the waterways. The same categories as those used in the Monitor Nautische Veiligheid by Hofmeijer (2020) are chosen:

- Inland vessels
- Recreational vessels
- Ferries
- Other type

In this classification, the 'Other types' category includes every vessel that does not fall within the inland, recreational or ferries categories, such as service vessels, police boats, fishing boats, and cruise ships. Unknown vessel types are also grouped within this category.

For each of these four types, expected characteristics have been defined. A simple notation consisting of a plus or minus symbol represents these characteristics. In this notation, a plus sign indicates an expectation of high values, while a minus sign means the opposite. For example, when it comes to acceleration, a small value is expected for inland vessels. Therefore, a minus sign is associated with this characteristic in the reference table, Table 3.1. When an inland vessel with a significant acceleration is encountered in the data, this deviation from the norm should be noted. The same principle applies to the expected speeds for each vessel type.

Moreover, a very low or nearly zero speed value over a longer period can also be a determinant factor. Stationary vessels are expected to be anchored or moored. The specific locations for anchoring or mooring vary among vessel types: recreational vessels are commonly found in marinas, while inland vessels are often found in larger ports.

Additionally, the presence of stationary vessels outside ports or designated anchorage areas may indicate potential grounding accidents. In the last place, there are certain areas where anchoring is prohibited, close to bridges, locks or in the middle of the waterway. This could lead to dangerous situations, and detecting vessels anchored in these restricted zones can prevent hazardous situations.

Furthermore, manoeuvrability is another characteristic that varies with the vessel type, often determined by the vessel size. Abrupt directional changes are more common for smaller vessels, such as speedboats. When a large vessel has an extreme course change, this can indicate a dangerous situation.

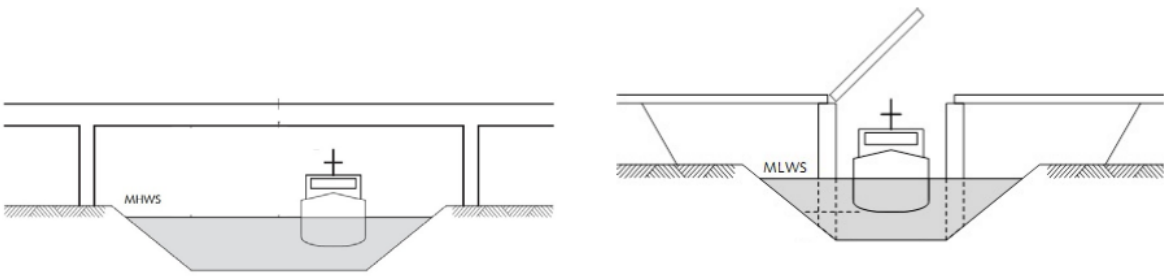
Table 3.1: Overview expected behaviour per vessel type

	Inland vessels	Recreational	Ferries	Other
Speed	-	+/-	+	+/-
Acceleration	-	+/-	+	+/-
Manoeuvrability	-	+	-	+/-

3.1.2. Interaction with infrastructure and objects

Defining the expected behaviour of vessels during interaction with infrastructure or objects is essential for identifying anomalous behaviour and assessing the potential risks of accidents of this nature. The primary identification criteria will be based on the relative position, speed and manoeuvres observed in close encounters with infrastructure or objects.

To minimise the risk of a collision, the relative position with respect to an object or part of the infrastructure should be sufficient. For instance, when navigating under a bridge, the vessel is expected to pass through the midpoint between the two supporting pillars and, if applicable, use the designated passage opening. A bridge with a movable part has a typically smaller passage width, as illustrated in Figure 3.2b. In the case of a bridge without pillars, the vessel typically passes without deviating from its course.



(a) Passage of fixed bridge (Rijkwaterstaat, 2020) **(b)** Passage of movable bridge (Rijkwaterstaat, 2020)

Figure 3.2: Typical bridge passage for fixed bridge and movable bridge

In an ideal scenario, a skipper will position his vessel at a significant distance from the bridge in the correct position and maintain a constant speed an heading during passage. Therefore, vessels that significantly accelerate, abruptly decelerate or heavily change course close to a bridge should be considered deviations from the norm and potentially anomalous.

In the case of a lock, vessels are expected to approach in a steady and controlled manner, gradually decelerate, and moor in the designated position while awaiting the next locking cycle.

Similar definitions and criteria for vessels interacting with objects such as buoys, sign poles or jetties can be established. These criteria are primarily based on expected passing positions, course adjustments or velocity adaptations. For instance, when approaching a buoy, guidelines dictate the side from which

a vessel is expected to pass. Any deviation from this prescribed course may indicate a potentially dangerous situation.

3.1.3. Interaction with other vessels

When vessels interact with each other, their behaviour often diverges from that of vessels navigating solo in rivers or canals. This altered behaviour is often related to water movement around the vessel, which can lead to suction effects. Typical manoeuvres like head-on encounters, overtaking or crossings, are typically investigated in vessel interaction studies. An example of this is the work by Montewka et al. (2010), who introduced a new approach for collision probability modelling during such vessel interactions. The subsequent sections will explore these manoeuvres to identify aspects that define important vessel behaviour in these scenarios.

Encountering

During encountering manoeuvre, two potentially dangerous situations arise. The first occurs when the two ships begin to feel each other's influence, causing them to push each other aside The vessels push each other aside due to the water movement around the bow. This can result in grounding on the riverbank, as depicted in situation a in Figure 3.3. The second situation arises when too much rudder is used to avoid grounding, causing the vessel to yaw toward the centre line. This introduces the risk of collision with a following ship or colliding with the opposite bank, as illustrated in situation b in Figure 3.3 (Van Koningsveld et al., 2021).

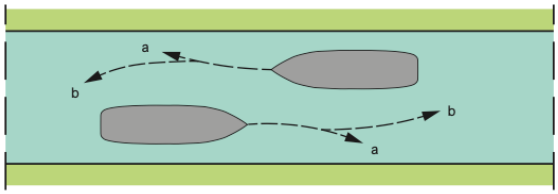


Figure 3.3: Dangerous situations during an encounter (Van Koningsveld et al., 2021)

Encountering manoeuvres are typically of short duration, therefore the impact of the forces exerted by the vessels on each other and their resulting movements are relatively mild. The relative distance between the vessels is crucial, as smaller distances increase the forces and associated movements. Furthermore, the angle between the vessels can indicate the effects of the encountering manoeuvre.

Overtaking

At the start of an overtaking manoeuvre, the bow of the overtaking vessel is drawn towards the stern of the ship being overtaken. As the ships sail alongside each other, they are pulled towards each other. If this attraction persists, the ships effectively merge into one and risk losing control. During the final part of the manoeuvre, the overtaking vessel must overcome the adverse water level gradient created by the other ship's water level depression. The same situation as in the beginning occurs, only this phase takes much longer (Van Koningsveld et al., 2021).

The relative speed is an important factor in overtaking manoeuvres. Insufficient relative speed means overtaking will take a very long time or even becomes impossible. Indicators in the vessel behaviour include relative speed and distance between the vessels. Instances of overtaking manoeuvres found in areas where this is prohibited indicate anomalous situations, they can be found based on the vessels paths. Likewise, unsuccessful overtaking attempts can signal potential danger.

Crossing

Vessels crossing each other on waterways can give rise to dangerous situations. Relatively many ship-ship collisions occur at bifurcations or waterway crossings, especially with high shipping intensity (Hofmeijer, 2020).

When vessels cross, the most important risk identification can be made by the relative distance between them. Furthermore, the angle at which ships cross a river or canal section can indicate potential risks.

In most cases, it is desirable to cross a waterway at an angle as close to 90 degrees as possible, to minimise the path.

In all these examples, the distance between the vessels is important. Various definitions have been used over the years to study these interactions. For instance, Berglund and Huttunen (2009) use the absolute distance between two vessels as the criterion to define near misses, while Goerlandt et al. (2012) apply the elliptical ship domain proposed by Fujii and Tanaka (1971). This domain represents the area around a ship that other vessels must avoid. Szlapczynski and Szlapczynska (2017) overviews various safety domain definitions used in previous research, including the safety criterion, method of determination, ship-related factors, situation and environment-related factors. This diversity highlights the various choices in defining safety regions related to distance.

Recent work on ship domains is executed by Baak (2023) proposing a method to determine ship domains within port areas. The method considers both situation-specific parameters like encounter type and the relative positioning of the vessels as well as ship-specific parameters like the length, width and velocity. By determining the relations of the parameters to the critical distance, the size of the domain is found, with the shape being recommended to be an ellipse. While primarily designed for port environments, the principles of the method can serve as valuable input for the vessel domain for investigating ship-ship interactions on inland waters.

3.2. Background and overview on methodological steps

This section provides a broad perspective on the steps involved in the methodology. These methodological steps are the foundation for a comprehensive analysis of vessel behaviour patterns using the AIS logs. A brief description of the main steps is provided below, while subsequent subsections will offer more detail and background on each step.

- **Feature engineering:** The key features describing the vessel behaviour are extracted from the AIS data in this step. These features are used to describe the behaviour patterns that are aimed to be analysed.
- **Dimension reduction with Uniform Manifold Approximation and Projection (UMAP):** UMAP is a dimension reduction technique that represents complex data in a lower-dimensional space while preserving essential characteristics. This helps simplify the analysis of vessel behaviour patterns.
- **Clustering using K-means:** K-means clustering is applied to the dimensionally reduced, unsupervised data. K-means helps to group similar vessel behaviour patterns into clusters, facilitating further analysis.

3.2.1. Feature engineering

Feature engineering is the process of creating and transforming input data into a format that is suitable for a machine-learning model. This process is an important part of building a machine-learning model, as the quality of the input data significantly affects the performance of the model (Patel, 2021). A feature is any measurable input that can be used in a predictive model. The input variable type defines different methods to create a feature, which will be discussed later.

The features used in this research aim to describe a vessel's behaviour on a waterway, as defined in Section 3.1 with the AIS data as input source. Next to the features generated based on general shipping behaviour, specific features for ship-infrastructure interactions are created. It is noteworthy that although behaviour definitions include details on ship-ship interaction, they are not further employed in this study.

Data types

As discussed before, how a feature is defined varies depending on the type of variable used. Stevens (1946) categorises variables into four measurement levels: nominal, ordinal, interval, and ratio. Each level has unique measurement properties and different allowed mathematical operations, as in Table 3.2. The framework guides in choosing appropriate statistical techniques and determining the level of analysis that can be applied to different variables.

Table 3.2: Level of measurement (Stevens, 1946)

Level	Measure property	Math operators	Example
Nominal	Classification	$=, \neq$	Vessel type
Ordinal	Comparison	$>, <$	Manoeuvrability scale
Interval	Difference	+, -	Direction measured in degrees from true north
Ratio	Magnitude	$\cdot$, /	Length

Nominal variables are categorical variables with no inherent order or numerical value. They represent qualitative data and can have multiple distinct categories. Examples include gender, colour or country. Nominal variables can only be classified into categories, and the only operation that can be performed on them is counting or frequency analysis. The AIS data contains nominal variables in vessel types, categorising vessels into categories like inland, recreational, fishing vessels and many more types.

Categorical data is also represented by ordinal variables, but an inherent order or ranking exists among the categories. A logical order is assigned to the categories, but the exact differences between them may not be uniformly meaningful. Examples include satisfaction rates and educational levels. The order or ranking in ordinal variables can be defined, but the exact differences between categories cannot be determined. In this context, an ordinal variable, such as vessel size categorised as small, medium, or large vessels, or manoeuvrability ranked from highly manoeuvrable to poorly manoeuvrable, can be considered.

The interval variables represent numerical data, and the differences between values are meaningful and consistent. The order is defined, and the intervals between values are equal. However, interval variables lack a meaningful zero point. Temperature measured in Celsius or Fahrenheit is an example of an interval variable. Mathematical operations like addition and subtraction can be performed on interval variables but cannot interpret ratios or calculate meaningful differences based on the zero point. Typical interval variables in AIS data are timestamps, time and date values representing specific moments when data was recorded. The course over ground (CoG) is another interval variable, the Vessel's direction of movement relative to true north, measured in degrees. The CoG represents a chronological order with consistent and meaningful intervals but lacks a meaningful zero point and does not support ratio comparisons.

The ratio scale consists of variables similar to the interval variables but with a significant zero point that enables the interpretation of ratios. The ratio variables have a defined order, equal intervals between values, and a true zero point. Examples include height, weight, or income. All mathematical operations can be performed with ratio variables, including addition, subtraction, multiplication, division, and calculating meaningful ratios. Ratio variables in AIS data are the longitude and latitude position, distance travelled and speed.

While an overview of variable types and their role in feature engineering is presented in this section, Chapter 5 will provide a more in-depth exploration of each individual feature, including the way of generating it from AIS data.

3.2.2. Dimension reduction with UMAP

The high-dimensional representation of vessel behaviour, resulting from the feature generation, makes data analysis complex. Therefore, an efficient method for data compression is needed to preserve all relevant information regarding the vessel behaviour. Therefore, dimension reduction is employed using Uniform Manifold Approximation and Projection (UMAP), a nonlinear dimension reduction technique by McInnes et al. (2018). All features are consolidated into a lower-dimensional embedding. For this study, a 2-dimensional embedding is chosen for ease of visual exploration.

In addition to UMAP, several other algorithms for dimension reduction are available, such as t-distributed Stochastic Neighbour Embedding (t-SNE) and Principal Component Analysis (PCA) (Jollife & Cadima, 2016; Van Der Maaten & Hinton, 2008). UMAP is selected due to its computational efficiency and better

scalability. UMAP is compared to both t-SNE and PCA, and shows faster computation and improved scalability, allowing the generation of high-quality embeddings for large data sets (McInnes et al., 2018).

Ashush et al. (2023) demonstrates the effective application of UMAP in an unsupervised machine learning-based approach for drone swarm characterisation and detection. This study extracts features from drone radio frequencies, followed by dimension reduction techniques, including Independent Component Analysis (ICA), PCA, t-SNE and UMAP. The findings indicate that a combination of ICA as a linear pre-dimension reduction technique and nonlinear UMAP for post-dimension reduction yields promising results (Ashush et al., 2023).

In general, dimension reduction techniques aim to represent the original data in a lower dimension while preserving the original data structure as much as possible. UMAP constructs a high-dimensional graph representation of the data and optimises a low-dimensional graph to be as structurally similar as possible.

This reduction condenses all the data into just two dimensions, suitable for plotting on a scatter plot to facilitate visual exploration. Theory behind UMAP suggests that, similarly to the original data, after applying the dimension reduction, data points close share similar shipping behaviour. This allows similar navigating patterns to be more easily recognised within the data set.

The following part offers further insight into how UMAP works, the algorithm consists of two primary phases: graph construction for the high-dimensional space, and optimisation of the low-dimensional graph layout (Sainburg et al., 2021; Wang, Huang, et al., 2021).

Understanding UMAP: key principles simplified

UMAP makes use of simple combinatorial building blocks known as simplices. Geometrically, a simplex is a simple way to build a k-dimensional object formed by taking the convex hull of k+1 points. Figure 3.4 illustrates lower-dimensional simplices, from a 0-simplex representing a single point to a 3-simplex, which is a tetrahedron consisting of four 2-simplices as faces (Leland McInnes, 2023).

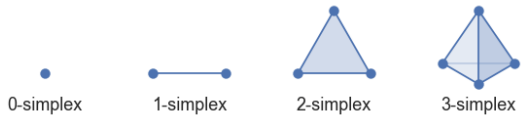
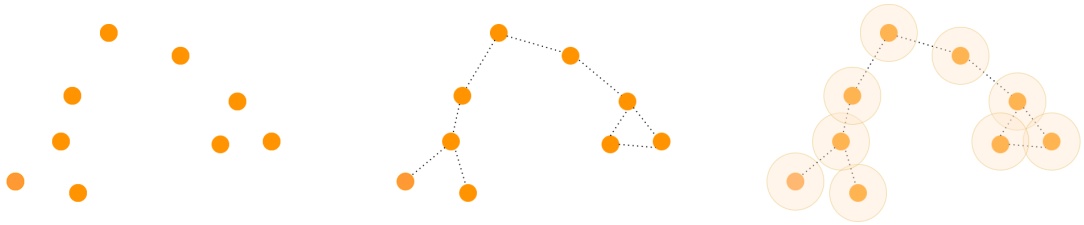


Figure 3.4: Example simplices (Leland McInnes, 2023)

To approximate the shape of the data, these simplices are utilised by UMAP, creating 1,2, or higher order simplices among the data points to approximate the topology. Example data is presented in Figure 3.5a, while Figure 3.5b shows the connections represented by these simplices in the data.



(a) Example data **(b)** Linked example data **(c)** Example data with fixed radius

Figure 3.5: UMAP example data 1

To establish these connections, the UMAP algorithm extends a radius around each point, and connections are created where their radius intersect, as shown in Figure 3.5c. However, it is possible that not all points fall within this radius, resulting in some unconnected points as depicted in Figure 3.6a. This

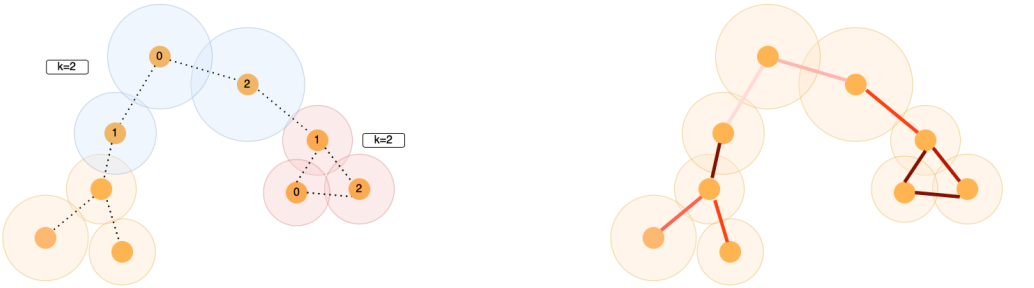
usually happens in low-density regions, whereas high-density regions have many neighbouring points, leading to numerous connections.

(a) Example data with unconnected points

(b) Example data with variable radius

Figure 3.6: UMAP example data 2

To overcome this issue, a variable radius is employed by UMAP, ensuring that all data points are connected, as seen in Figure 3.6b. The radius is larger in low-density regions and smaller in high-density regions. UMAP estimates density using a proxy. The density is estimated higher when the kth nearest neighbour is close. In this context, the term 'kth nearest neighbour' refers to the k closest data points to a given point. When k=2, this means the two closest points are used. As illustrated in Figure 3.7a, results in identifying regions with varying densities, where higher density is represented in red, and lower density is represented in blue. The choice of 'k' is one of the UMAP hyperparameters, a parameter whose value is used to control the learning process in machine learning. A large 'k' preserves the global structure, while a small 'k' reduces the radius, and the local structure is more preserved.



(a) UMAP example data density regions k=2

(b) UMAP example data with connection weights

Figure 3.7: UMAP example data 3

The next step involves assigning weights to all connections, denoting the connection probability. Points that are farther away receive lower weights and have lower connection probabilities. In Figure 3.7b, the colours represent the weighted connections, with darker colours indicating stronger connections. The constructed high-dimensional graph can now be projected to lower dimensions.

The graph projection can be compared to the high-dimensional graph where connections function like springs, with stronger springs indicating higher probabilities. As a result, points linked by high-weighted edges are more likely to remain in close proximity within the lower dimensional space.

3.2.3. Clustering using K-means

After reducing the data to two dimensions, various methods can be used to group the unlabelled data. In this case, clustering by K-means by Pedregosa et al. (2011) is applied. The clustering aims to convert the two-dimensional data into several clusters, where similar vessel patterns will be grouped together (Jin & Han, 2017).

The K-means algorithm clusters data by attempting to separate samples in 'n' groups of equal variance, minimising a criterion known as inertia or within-cluster-sum-of-squares. This algorithm requires the number of clusters to be initially specified. It is known for its scalability to large data sets and has been widely applied across many different fields (Pedregosa et al., 2011).

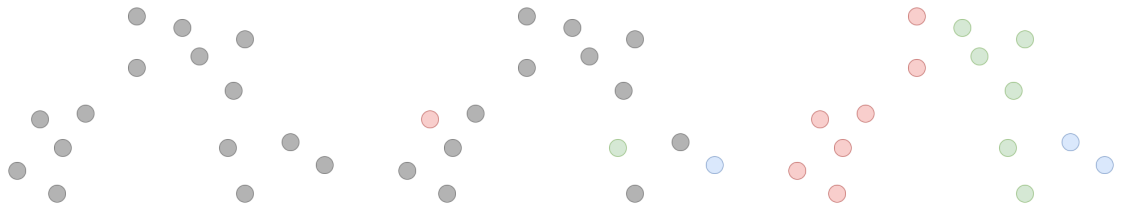
UMAP combined with K-means clustering is a widely used approach in various fields, as demonstrated by several studies in the literature. For instance, the study by Hozumi et al. (2021) a large data set of Coronavirus mutations undergoes dimension reduction using UMAP, followed by K-means clustering. Their results show that UMAP outperforms other dimension reduction techniques when combined with K-means with large data sets.

Clustering techniques are also used in shipping studies. In the work of Daranda and Dzemyda (2020), Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is applied for anomaly detection in vessel turning points, using longitude and latitude values. This method identifies anomalies as points lying outside the normal traffic flow. More recently, Widyantara et al. (2023) have focused on clustering similar vessel movement patterns based on AIS data. Their approach involves using the Longest Common Subsequence to determine the similarity between vessel paths, followed by Multi-Dimensional Scaling to reduce the 3D data to 2D and a clustering with DBSCAN. While this approach effectively groups vessels navigating in different directions in the Lombok Strait, it primarily relies on the distance between trajectories, which may be less suitable for anomaly detection in inland waters where vessels tend to sail in closer proximity.

In this research the vessel behaviour is defined using multiple features, making UMAP by McInnes et al. (2018) in combination with K-means a suitable option. This choice is driven by the capability to handle large data sets and scalability. The upcoming section gives a simple explanation of the K-means clustering algorithm along with an introduction to various evaluation methods.

Explanation of the K-means clustering algorithm

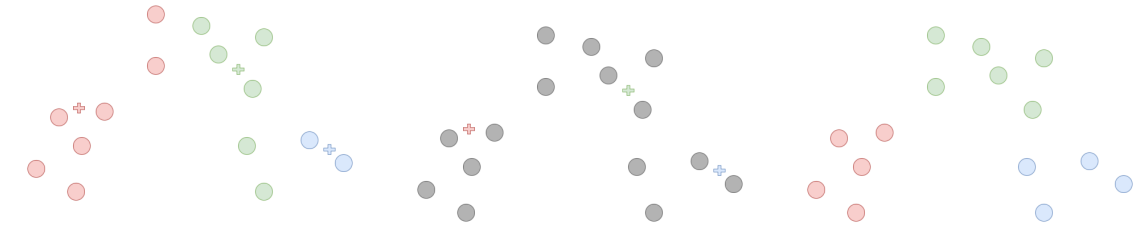
K-means clustering is explained with a simple example of 15 randomly chosen data points in Figure 3.8a. First, the desired number of clusters, 'k', is selected. In this case, three clusters are aimed for, and three starting points are randomly chosen from the data points, represented by the red, blue and green colours in Figure 3.8b. Next, each point is assigned to the closest cluster centroid, resulting in Figure 3.8c.



(a) Example data **(b)** Initial cluster centres **(c)** Initial clustering result

Figure 3.8: K-means example 1

Subsequently, the centroid of the new clusters is computed, indicated by the cross in the corresponding colour, illustrated in Figure 3.9a. Then, the points are once again assigned to the closest new centroid in Figure 3.9b. Figure 3.9c displays the final result. This process of determining the new centroid of the cluster and assigning points to the clusters continues until one of the stopping criteria is reached. These are reaching the maximum number of iterations, centroids of newly formed clusters not changing, or points remaining in the same cluster.



(a) Updated centroids represented by + symbols

(b) New cluster assigning

(c) Final clustering result

Figure 3.9: K-means example 2

The 'k' value in K-means clustering is a crucial hyperparameter that determines the number of clusters to be formed in the data set. Finding the optimal 'k' value can be challenging, as a small value can result in under-clustered data, and a large value can cause over-clustering. The following section will discuss clustering performance and a method for determining a suitable initial number of clusters.

Determine number of clusters 'k'

While there is currently no established method for precisely determining the optimal 'k' value, there exist techniques for making estimations. In this instance, the elbow method is used where, for a range of 'k' values, the within-cluster-sum-of-squares (WCSS) is computed and plotted on the y-axis against the number of clusters on the x-axis as demonstrated in Figure 3.10 (Rao, 1969).

WCSS quantifies the square average distance between all data points within a cluster and the respective cluster centroid. In the process the Euclidean distance between a point and the centroid to which it is assigned is measured. This is iterated over all points within a cluster, followed by summation across all clusters, as demonstrated in Equation 3.1.

$$WCSS = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{3.1}$$

Where:

- x_i is a data point belonging to the cluster C_k
- μ_k is the mean value of the points assigned to the cluster C_k

The point at which the plot shows a bend, often referred to as the 'elbow' is generally considered to indicate an appropriate number of clusters. An example of such a plot is presented in Figure 3.10. In this specific example, after 6 to 10 clusters, the decrease in WCSS is minimal, suggestion that the optimal value for 'k' lies in this range (Cui, 2020).

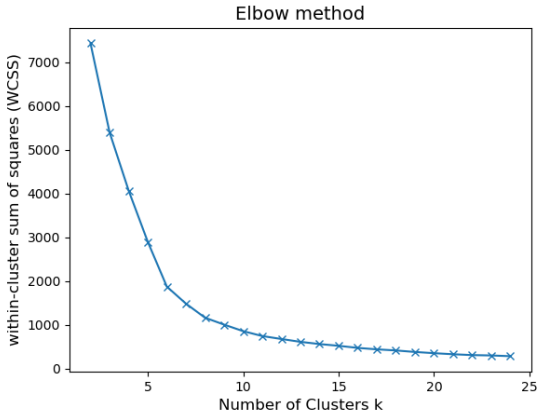


Figure 3.10: Example plot elbow method

It is worth noting that the method is not always straightforward and distinctive, as the presence of a distinct elbow in the plot is not guaranteed. Additionally, visually identifying the elbow point is subjective, different individuals may interpret the plot differently. However, the kneed Python package based on the paper by Satopää et al. (2011) offers assistance in determining this point. This algorithm defines the elbow as the location with the maximum curvature on the line. An illustration demonstrating this concept is presented in Figure 3.11, including the calculated elbow point.

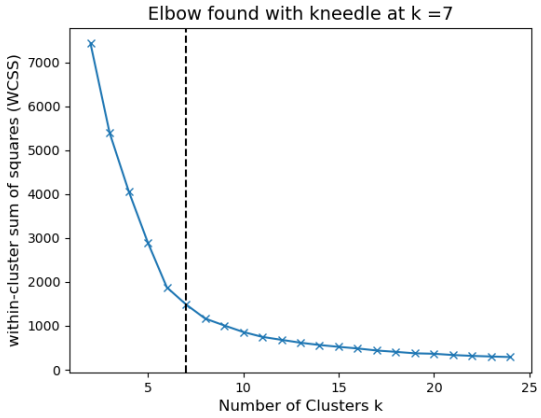


Figure 3.11: Example plot elbow method with calculated elbow point

The upcoming part will introduce additional scoring metrics to evaluate the clustering results, providing valuable insights into clustering performance.

Clustering performance evaluation

Given that the data is unlabelled, determining the optimal number of clusters is a challenging task. Evaluating the quality of clustering results relies on distance-based metrics. In this context, several scoring methods are introduced, including the Silhouette Coefficient by Rousseeuw (1987), Calinski-Harabasz Index by Caliński and Harabasz (1974), and Davies-Bouldin Index by Davies and Bouldin (1979). Detailed descriptions of these metrics can be found in Table 3.3 (Halkidi, 2001; Pedregosa et al., 2011).

Table 3.3: Clustering performance and score descriptions

Score	Description
Silhouette Coefficient	The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters.
Calinski-Harabasz Index	The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.
Davies-Bouldin Index	A lower Davies-Bouldin index indicates more effective separation between the clusters

The application of the elbow method, followed by an examination of the scoring metrics presented in Table 3.3, provides valuable insights into the clustering performance and helps in determining the optimal number of clusters.

After the clustering step, the naming of the generated clusters will be addressed in Chapter 5, where a more detailed and practical exploration of the methodology will be provided. With the overview and background of the methods presented in this section, an appropriate location to develop and apply the method is necessary and will be discussed in Section 3.3.

3.3. Location of interest to set up model and test method

The next step in developing the method, as described in Section 3.2, involves the selection of a location for setup and testing. To define this location, considerations were made regarding specific types of accidents, the vessels involved, and the top ten risks from Section 2.2 as defining characteristics of potential interesting locations.

Two locations, Hollands Diep and the IJ, were chosen after evaluating various options for model setup. While these choices might not be optimal, they hold significance due to their alignment with specific waterway characteristics, outlined in Table 3.4. These characteristics are derived from top risks, considering factors such as vessel types and their influence on vessel behaviour, which are influenced by traffic intensity, infrastructure, and waterway geometry. The IJ is illustrated as the red line in the northern region of Figure 3.12, while Hollands Diep is situated in the southern region. In the following subsections, detailed descriptions of both locations and an further evaluation of their characteristics is provided.

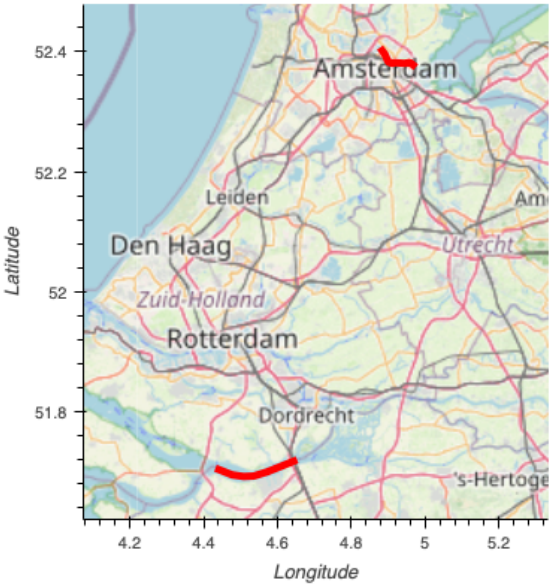


Figure 3.12: Location of the study cases: the IJ in the north and Hollands Diep in the south

Table 3.4: Waterway characteristics Hollands Diep and the IJ

	Hollands Diep	The IJ
Interaction recreational-commercial	+	++
Recreational vessels	+	++
Ferries	-	+
Inland vessels	++	+
Traffic intensity	++	++
CEMT-class	VIc	VIb
Corridor	Rotterdam-Germany	Amsterdam-Rhine
Infrastructure	Bridge / Locks	Bridge / Locks
Registration rate	+	+
Geometry waterway	+	++

3.3.1. Hollands Diep

Hollands Diep is a river in South Holland, extending from the point where the Amer and Nieuwe Merwede merge to the Haringvlietbrug, where it transitions into the Haringvliet (Rijkswaterstaat, n.d.-a). This river is part of the Rotterdam-Antwerp corridor. A visual representation of the river section used in this study is provided in Figure 3.13, further details on the river's dimensions are found in Table 3.5.

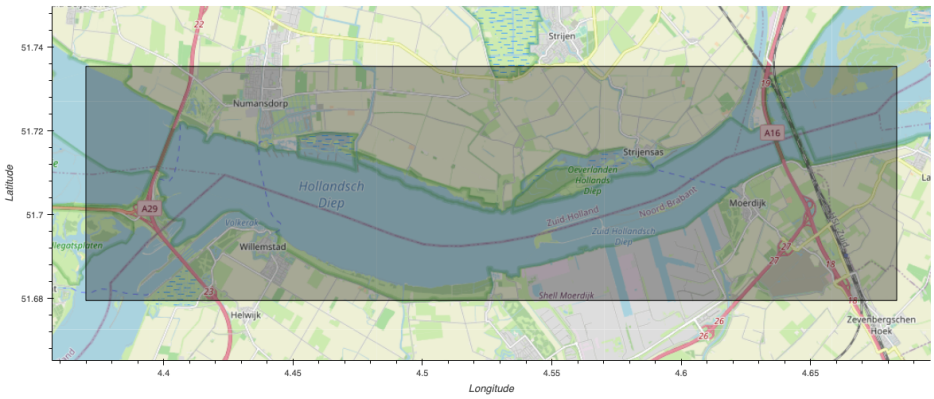


Figure 3.13: Overview Hollands Diep data area

Hollands Diep holds a classification of CEMT-VIc, which allows for the navigation of a push-barge unit with 3x2 barges, corresponding to a vessel with a total length of 270 meters and a width of 22,8 meters (Rijkwaterstaat, 2020). Moreover, the Port of Moerdijk is situated along Hollands Diep, indicating a substantial presence of large inland vessels navigating the river. Additionally, the waterway sees significant recreational activity, with several smaller marinas located along its banks, the largest of which is situated in Willemstad. Consequentially, interactions between inland vessels and recreational vessels will occur.

Table 3.4 indicates a minus symbol concerning the ferries operating on the Hollands Diep. Two small ferry lines are in operation, for pedestrians and bikers, but they only run a couple times a day during the summer months. One route connects Moerdijk and Strijen in the east, while the other links Willemstad and Numansdorp in the west (Vereniging vrienden van veerponten, 2023). The traffic intensity is classified as high, as determined by data from the Monitor Nautische Veiligheid (Hofmeijer, 2019).

Regarding the geometry of the waterway, it is generally quite wide, with few exceptional features. One notable aspect in terms of geometry is the corner to the Dordtsche Kil which lies close to the Moerdijkbrug. This corner is recognised as a challenging area. Rijkswaterstaat has implemented modifications at this location in the past (Varen doe je Samen!, 2016). Furthermore, in terms of infrastructure, the Volkeraksluizen are situated to the west, ranking as Europe's largest and busiest inland locking complex (Rijkswaterstaat, n.d.-c). To the east, the Moerdijkbrug spans the Hollands Diep, as depicted in Figure 3.14.

Figure 3.14: Moerdijkbrug at Hollands Diep (Rijkswaterstaat, n.d.-a)

Three bridges span Hollands Diep: one is part of the A16 motorway, and the other two are railway bridges, connecting South Holland and Brabant. These bridges all provide a passage width of 100 meters between their pillars and include ten openings, each designated for a specific direction of navigation. With the exception of four openings, the openings situated near the banks, are reserved for recreational vessels, allowing them to cross the bridge in both directions, a complete overview of the

passage directions and dimensions per span is found in Figure D.1 (Varen doe je Samen!, 2018).

Table 3.5: Waterway dimensions Hollands Diep and the IJ (Rijkswaterstaat, n.d.-a, n.d.-b)

	Hollands Diep	The IJ
Length	21 km	12 km
Width	905 to 2070 m	150 to 600 m
Depth	-8,00 to -6,00 NAP	-11,00 KP

3.3.2. The IJ

The IJ, flowing through North Holland from the Markermeer crossing Amsterdam before reaching the Noordzeekanaal, is divided into two sections: the Buiten-IJ to the east of the Oranjesluizen and the Afgesloten- or Binnen-IJ to the west (Rijkswaterstaat, n.d.-b). This waterway is part of the Amsterdam-Rhine corridor. Figure 3.15 highlights the study area represented by the grey-shaded region. Additional details on river dimensions are stated in Table 3.5.

Figure 3.15: Overview the IJ data area

The Binnen-IJ part is classified as CEMT VI-b, permitting the navigation of push-barge units up to four barges. East of the Oranjesluizen, in the Buiten-IJ, the classification is VI-a, corresponding to push-barges units of two barges. With the Port of Amsterdam close by and the direct link to the Amsterdam-Rijnkanaal, the IJ experiences substantial inland vessel traffic. Additionally, the IJ attracts numerous recreational vessels originating from the Amsterdam canals or coming from the Markermeer heading towards the North Sea.

The IJ host nine ferry lines, mainly linking central Amsterdam to various parts of the city. While the frequency of service varies among these lines, the overall presence of ferries contributes to enhanced waterway activity. Figure D.2 provides an overview of the ferry routes. Additionally, river cruises and other passenger vessels frequently navigate the water of the IJ, resulting in regular interactions between recreational and commercial vessels.

In terms of geometry, the IJ has some special challenges. Specifically, there is a relatively confined area in front of Amsterdam Central Station where many vessels pass through. Numerous smaller waterways from the city centre flow into the IJ, and the presence of tight corners and smaller harbours further complicate navigation. In regard to the infrastructure, particular interest lies in the eastern part, where the Oranjesluizen and Schellingwouderbrug are located, shown in Figure 3.16.

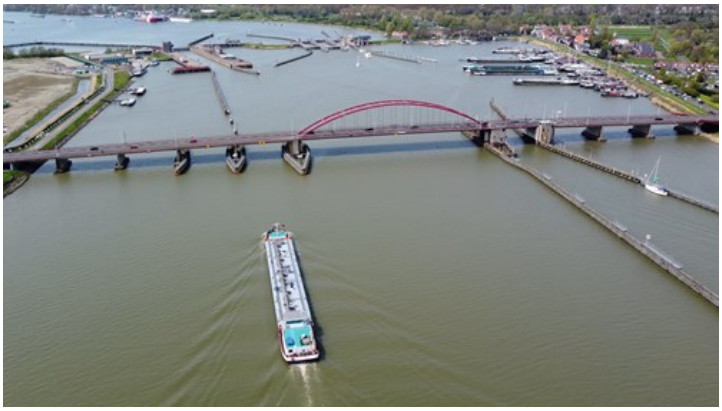


Figure 3.16: Schellingwouderbrug and Oranjesluizen at the IJ (Watersportverbond, 2021)

The Oranjesluizen complex consists of four locks, with one large lock situated to the south, visible on the left side in Figure 3.16, and three smaller locks located in the north. Furthermore, the bridge has six navigable passages of variable size. The middle span has a width of 105 meters and permits vessels to navigate in both directions. To the right in the image, the movable section of the bridge is visible, offering a passage width of 18 meters (Waterkaart Live, 2023a). Detailed information about the passage rules is added in Figure D.3 in the appendix.

4
Available AIS data and its characteristics

The input data for defining trips and their corresponding behaviour during the trips is derived from AIS logs. Understanding the data type and background is essential for utilising this data effectively.

This chapter will explain how the AIS system works and which ships must have these systems on board. Furthermore, the parameters and characteristics of the data set used in this project are explained, followed by an analysis and exploration of the data. The last part of the chapter will threat other input data sources, for the geographical data.

4.1. General introduction to AIS

Automatic Identification System (AIS) is a system that enhances the safety of shipping. AIS aims to enable good communication between skippers, other vessels, and traffic stations. AIS is a fast, accurate and practical way of exchanging information. The AIS device consists of a transponder with GPS (Global Positioning System) and VHF (Very High Frequency) systems. The system automatically transmits radio waves at regular intervals. These radio waves carry location, speed and ship data related to the voyage. AIS devices automatically receive all information broadcast by other AIS devices on other ships and ashore within the transmission range (Bureau Telematica Binnenvaart, 2009).

The signal incorporates both dynamic and static information, with some static data requiring manual submission. Unfortunately, this manual submission is often associated with unreliability as random, or no information may be entered by the skipper, given the absence of a verification process. In contrast, the dynamic component of the signal, containing position and time data, is automatically derived from the ship's GPS system. This dynamic data is generally considered more reliable although it may still be subject to outliers due to environmental conditions, and the quality of the GPS coverage varies by location. The update frequency of dynamic information depends on the ship's speed, with shorter intervals when the speed over ground (sog) is higher, and vice versa. (Van Koningsveld et al., 2021).

With AIS, vessels are visible for traffic management and other ships over several kilometres, depending on location, antenna height and weather conditions. This allows skippers extra time to prepare for manoeuvres like overtaking, head-on passing or crossing safely, especially of great importance for locations with challenging geometry like bends and junctions or at places where the infrastructure on or around the waterway is of influence. It is often difficult to see other vessels in time at these locations. Knowing the position of approaching vessels makes it easier for a skipper to plan and make decisions independently. Besides, the waterway can be managed more efficiently (Rijkswaterstaat, 2022).

AIS has been mandatory on international seagoing vessels since 2004, while the use on inland waterways followed much later. Since 1 January 2016, AIS has been required on waters covered by the Binnenvaartpolitiereglement. Furthermore, vessels navigating the Rhine must have a mandatory electronic navigational chart, Inland-ECDIS (Electronic Chart and Display Information System). The obligation applies to all commercial vessels of CEMT class I and above and recreational vessels longer than 20 meters (Rijkswaterstaat, 2022). In addition to the recreational vessels for which it is mandatory to have AIS, some smaller boats have an AIS system on board for safety reasons on larger inland waters or because they like to track their trips.

The AIS system should continuously operate when ships are sailing or at anchor. If the skipper believes that the operation of AIS might compromise the safety or security of the ship or where security incidents are imminent, the AIS may be switched off (International Maritime Organization, 2015).

4.2. AIS data characteristics

Data used in this project originates from a data set made available to the TU Delft for research by Rijkswaterstaat. The collection contains anonymised AIS logs for four months of 2019; January, April, July, and October. An example of the dataset is given with a description of the parameters followed by an analysis of the data used.

4.2.1. Properties in AIS data

AIS logs are lists of consecutive positions and timestamps and contain additional information such as the IMO identification number, the ship name, the ship type, the ship's basic dimensions, draught, etc. Not all parameters are relevant for this study, an example of the data present in the AIS logs is given in Figure 4.1, the table contains 5 entries of the AIS log of 'test ship-1051'.

newtimestamp	shipname	vesseltypeERI	cog	latitude	longitude	sog	speed	length	width	draughtInland
2019-01-05 00:43:49	testschip-1051	8021	15.1	51.730145	4.628122	10.0	5.176234	85.0	10.0	1.4
2019-01-05 00:43:59	testschip-1051	8021	15.1	51.730598	4.628312	10.0	5.218988	85.0	10.0	1.4
2019-01-05 00:44:09	testschip-1051	8021	15.2	51.731041	4.628513	10.0	5.115036	85.0	10.0	1.4
2019-01-05 00:44:19	testschip-1051	8021	15.1	51.731491	4.628705	9.9	5.181279	85.0	10.0	1.4
2019-01-05 00:44:31	testschip-1051	8021	15.0	51.731941	4.628898	9.9	4.319140	85.0	10.0	1.4

Figure 4.1: AIS data frame example of 'test-ship-1051'

Timestamp

The timestamp contains a log's specific date and time in the AIS data, with an accuracy of seconds. The timestamp depends on the vessel's speed; a larger speed results in a smaller interval between the logs, up to a second. In Figure 4.1 the time difference, Δt, is around 10 seconds, which is used in combination with the distance between two consecutive logs to compute the vessel's speed.

Ship name

The ship name uniquely identifies the specific ship in the AIS data. This is part of the static information transmitted, which means that this will not change after the installation of the system. The data set used in this study is anonymised by Rijkswaterstaat and provided to the TU Delft for research purposes. This means the real ship name is changed to a standardised name, in this example, 'test ship-1051', with this identification, the privacy of the skipper is guaranteed, but it is still possible to identify a sailing pattern of a specific ship.

VesseltypeERI

Another static parameter is the VesseltypeERI, this is an international way of logging the vessel type. ERI stands for Electronic Reporting International. The VessletypeERI is a four-digit code, where the first digit indicates the type of navigation, '8' is specific for inland navigation and '1' for other types of vessels. The following two digits indicate a vessel or convoy and the last indicates the subdivision. In the given example data 8021 corresponds to a motor tanker, liquid cargo, type N. Pleasure crafts over 20 meters, mandatory to have an AIS on board are represented by the code 1850 (The European Commission, 2019). The complete overview of the codes and corresponding vessel types is found in Appendix A.

Course over ground (CoG)

The course over ground is the actual direction of a vessel, between two points, with respect to the surface of the earth. This can be different from the heading of the vessel, which is the direction where

the nose is pointed. Differences between the heading and CoG may occur due to external effects such as wind, tide and currents. This number is automatically updated from the vessel's main position sensor connected to the AIS if the sensor is able to calculate. The number has the unit degrees, relative to the north which is equal to zero degrees.

Latitude and longitude

For each timestamp, the latitude and longitude of the vessel are logged, representing the geographic location. Between two timestamps, the distance and direction can be determined based on the latitude and longitude position.

Speed over ground (sog)

The AIS data contains the speed over the ground (sog), which indicates the sailing speed of a vessel relative to the ground, expressed in knots. Besides the sog transmitted the speed of a vessel can be computed based on geographical information and time. Dividing the distance between two successive points over the time between these successive logs results in the speed in m/s, this remains relative to the ground and the results are stored in the column 'speed'.

Vessel dimensions

The dimensions of the vessel are part of the static data transmitted, consisting of the length, width and draught. The example shows the 'draughtInland' which is the draught of a vessel on inland waters. The AIS data also contains a column with 'DraughtMarine', indicating the draught on large open waters in this case disregarded since the focus area of this project consists of inland waterways. The dimension values are in most cases entered manually by the skipper, which results in missing or incorrect values.

4.2.2. Data exploration and analysis

With the general properties known, the subsequent step involves a more in-depth examination of the specific details of the data, starting with Table 4.1, which contains general information about the size of the area, number of vessels and trips.

Table 4.1: General details of the AIS data sets

	Hollands Diep	The IJ
Area	134,73 km^2	36,26 km^2
Number of logs	72.574.457	244.769.858
After removal of duplicates	41.663.806	170.080.541
Percentage of duplicates	43 %	31 %
Number of unique vessels	6848	7148

Data quality

Understanding the data is necessary for meaningful insights and consecutive step. Exploring the data often reveals missing values, outliers and inconsistencies or duplicates, which determine the data quality, which is not always the best according to Iphar et al. (2020). The issues in data transmission range from intrinsic weakness within the AIS system to errors in the messages, falsified data, and signal spoofing (Iphar et al., 2020).

Over time, AIs data quality has improved, according to Zhang et al. (2015), from mediocre in its early implementation years to quite reliable in the current days. Further quality increase is possible by using improved antenna installations in the AIS systems.

Duplicate logs in the data impact generated trajectories, instances of identical timestamps with different locations result in incorrect trajectories. In this case, removing duplicates reduces the data on the Hollands Diep by 43% and 31% at the IJ. Looking into the missing data, in most cases, static information is encountered, such as the vessel dimensions or the position of the AIS transmitter relative to the vessel. These are mostly manually entered, as discussed in Section 4.1.

However, missing values are not limited to manually entered data. The rate of turn (ROT), a dynamic property, often contains missing or questionable values. Only 5% of the values are non-zero, and extreme values of 720 and -720 deg/min, are prevalent, which are extremely large and, in the case of larger inland vessels, nearly impossible. A faulty value for the ROT is not uncommon in AIS logs according literature, the study by Felski and Jaskolski (2013) shows the absence of credible ROT values, especially at a low speed or stationary.

Data preparation steps are taken to overcome the duplicates and missing or incorrect values; Section 5.1 addresses all actions on the data set to ensure the accuracy and reliability of the subsequent analysis.

Comparison of vessel types between locations

Based on the analysis in section 3.3, differences in vessel types between the two locations are expected. The total count of specific vessel types is plotted using the parameter vesseltypeERI, offering a complete insight into the various vessel types within the four months' data. Figure 4.2 displays the top 20 vessel types in the Hollands Diep data set, while Figure 4.3 presents the same for the IJ section.

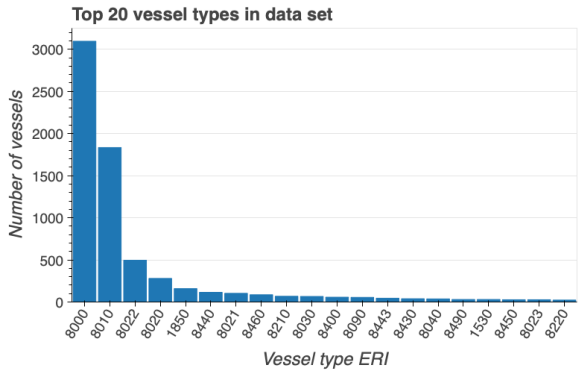


Figure 4.2: Top 20 vessel types in data set, Hollands Diep

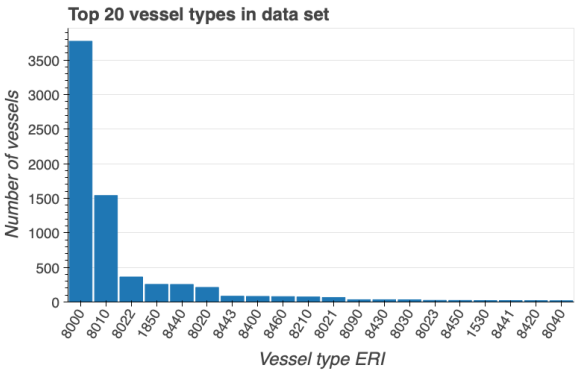


Figure 4.3: Top 20 vessel types in data set, the IJ

In both locations, the largest group consists of vessel type 8000, which represents an unknown vessel type. Following, the second and third largest groups correspond to motor freighters and motor tankers, respectively. Recreational vessels rank fifth in the Hollands Diep data set and fourth in the IJ data set, aligning with the expected higher presence of recreational vessels in the IJ region compared to the Hollands Diep section. More details on vesseltypeERI numbers and type descriptions can be found in Appendix A.

Furthermore, vessels are categorised into three groups, inland, recreational and other types, based on the vessel type code, with unknown vessel types falling under the other category. The categorisation is made since different behaviour is expected for different vessel types. The results for both locations are depicted in Figure 4.4 and Figure 4.5.

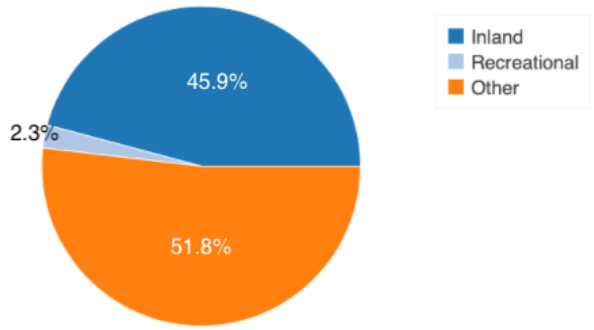


Figure 4.4: Distribution of vessel categories, Hollands Diep

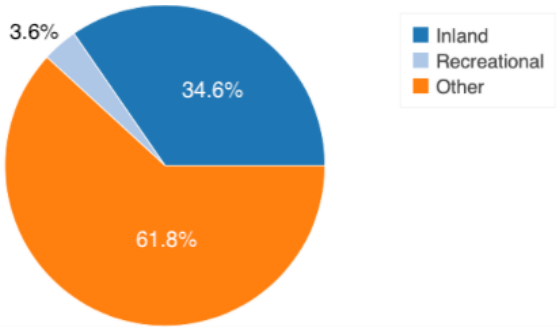


Figure 4.5: Distribution of vessel categories, the IJ

In both cases, the other category is the largest, primarily due to the substantial presence of unknown

vessel types in the data sets. Notably, in the IJ section, the recreational category slightly outweighs that of Hollands Diep, while the portion of inland vessels is smaller in the IJ compared to the Hollands Diep section.

For a better understanding of the vessel types navigating these two distinct locations, examining vessel length can provide valuable information. This property is often manually entered, leading to potential errors or missing data. The plots in Figure 4.6 and Figure 4.7 showcase vessel lengths in 20-meter bins, excluding the missing values.

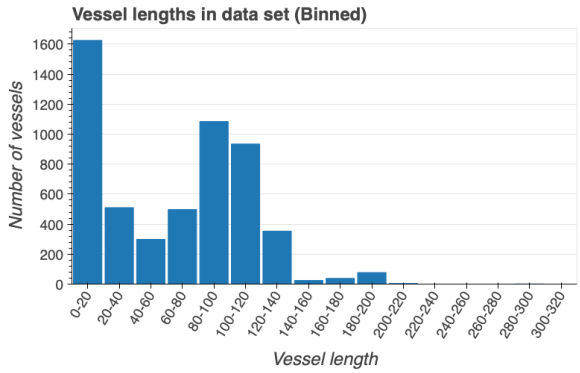


Figure 4.6: Distribution of vessel length, Hollands Diep

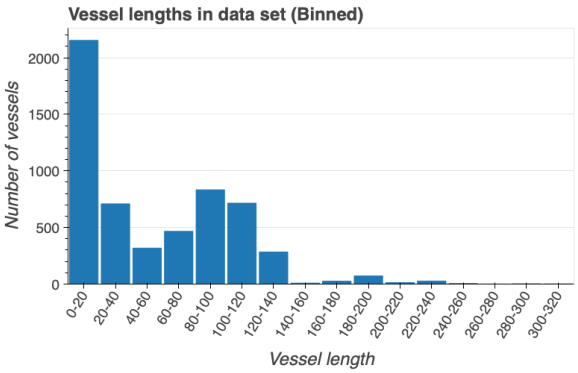


Figure 4.7: Distribution of vessel length, the IJ

In both cases, the group within the 0 to 20-meter range comprises the largest number of vessels. Additionally, a notable difference exists within the 80 to 140-meter range, with a greater concentration of vessels in the Hollands Diep region of this size, corresponding to a higher proportion of inland vessels. Conversely, in the IJ section, the largest vessels, reaching sizes of up to 320 meters according to the AIS data, are observed

4.3. Geographical data

Apart from the ship details in the AIS logs, this thesis relies on geographical information. Precise location information regarding bridges and buoys is essential for calculating vessel distances relative to bridge pillars or buoys. Furthermore, coordinates are necessary to clip the trips inside a specific region of interest, for example the close proximity of a bridge. Ship-infrastructure interaction features are defined in a clipped section of the trip, like speed in a range of 400 meters around the bridge.

The location data is obtained from OpenStreetMap, a collaborative world map created by contributors worldwide, available under the open license, CC BY-SA 2.0 (OpenStreetMap, 2023). This map is enriched by local knowledge, verified through various means like aerial imagery, GPS devices, and field maps, offering details from country borders and waterway sections to signs, all with corresponding coordinates.

Further geographical data, in the form of polygons used for clipping trajectories, is retrieved from GeoJSON.io (2023). This is a quick, simple tool for creating, viewing, and sharing maps. Geojson.io is named after GeoJSON, an open source spatial data format.

5 Implementation of method to detect anomalous vessel behaviour

This chapter aims to provide a detailed explanation of the practical implementation of the methodological steps introduced in Chapter 3. The primary focus lies in detecting unusual vessel behaviour concerning ship-infrastructure interaction. To achieve this objective, the definitions of vessel behaviour, methodological concepts, and location selection initially introduced in Chapter 3 are built upon.

The focus now shifts to implementing the complete method for detecting anomalous shipping behaviour. Beginning with the AIS data, the preparation steps for the data are explained. Individual vessel trips will be generated, incorporating information such as vessel characteristics, position, direction, speed and various other properties that define key aspects of shipping behaviour, referred to as 'features'. Additionally, software tools extract relevant characteristics from the time series data for the individual trips as additional features. With all vessel trips defined based on these features, the next step involves dimension reduction to represent individual trip behaviour in a lower dimension, known as 'embedding', while preserving essential underlying characteristics. The reduction is executed by applying Uniform Manifold Approximation and Projection (UMAP). After the dimension reduction, a visual representation of all behaviour characteristics can be made. Subsequently, K-means clustering is used to group similar behaviour patterns into specific clusters, facilitating classification and naming the clusters. The approach is summarised in the workflow diagram presented in Figure 5.1.

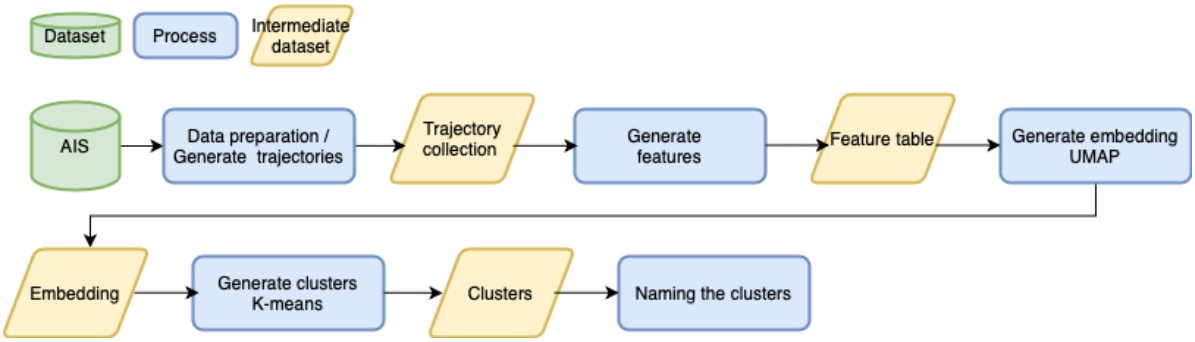


Figure 5.1: Workflow general method

5.1. Data preparation

The AIS data characteristics are elaborated on in Section 4.2. To be able to use the raw AIS data logs as input for the model, preparation steps have to be taken. These preparation steps aim to generate a smooth trajectory path per vessel in the area of interest. Therefore duplicates, outliers and location filtering is used, an overview of the workflow is presented in Figure 5.2.

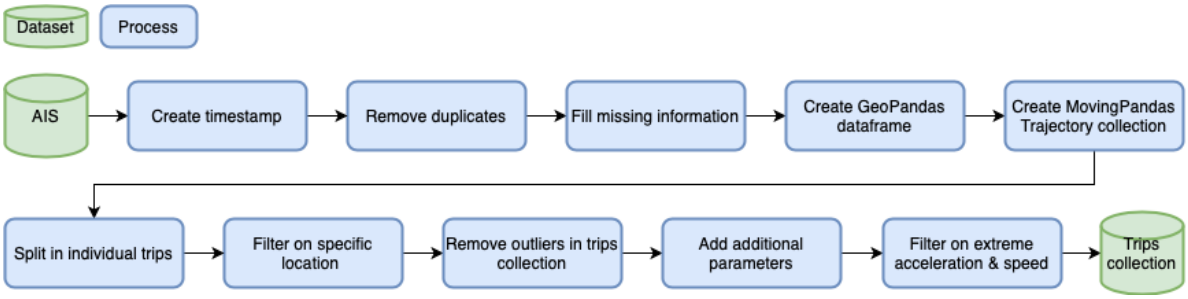


Figure 5.2: Workflow data preparation

Create timestamp

The first step in this process is creating a timestamp for all logs in the same format and units, ensuring data consistency for further usage in the processing and analysis. The timestamp, combined with geospatial data, generates the trajectories, representing trips of individual vessels. Over time every consecutive geometry point is linked to generate a path, therefore the format of every entry should be in the same form.

Duplicates removal

Duplicates in the dataset are filtered out to improve data quality and computational efficiency. Duplicate records can introduce abrupt changes in vessel positions and movements when constructing trajectories, which can impact calculated velocities and accelerations. Duplicates are identified based on their later occurrence in time, location and ship name, the first appearance of the record in time is kept, similar to the approach of Chen et al. (2020).

Fill missing information

Missing vessel information is added to the data set after the duplicates are sorted out. In some AIS logs, not all vessel characteristics are present, mainly information on the type, length, width and position of the AIS transmitter relative to the vessel. In most cases, only the first log of the specific vessel contains this information, which will be lost when the vessel's complete trajectory is split into individual trips. To overcome this lack of information, the known time fixed dimension values per vessel are copied to every AIS log of the specific vessel.

Create GeoPandas dataframe

The complemented Pandas data frame is geospatial encoded into a GeoPandas data frame, as the GeoPandas package can handle geospatial operations. A new column, 'geometry', is added to the data set, which can store geospatial data types, a point with the latitude and longitude coordinates per log, which allows accurate distance calculations or intersections with a specific region. Besides, a Coordinate Reference System (CRS) is added to the geometry, which defines the spatial reference for the data, making it possible to work with data in different coordinate systems and analysis tools.

Create MovingPandas trajectory collection

The data frame with the geometry points is converted into a trajectory for each unique vessel in the data set with the Python package MovingPandas by Graser (2019), which focuses on working with and analysing movement data. A trajectory consist of all points of the geometry column linked in time, in this case the latitude and longitude coordinates. A minimum length of 1000 meters is set as the required length for a single trajectory, smaller trajectories are disregarded because the number of data points becomes very small, where the effect of outliers will have a relative large impact in the analysis. The trajectories per vessel are split into individual trips since a single vessel can sail more than once through the area of interest over the period, leaving the area and entering again will be seen as a new trip.

Filter trips on location

After generating the individual trips, a filter on a specific location can determine which trips are relevant in a particular case. For example, when interacting with a bridge is investigated, trips in close approximation to the bridge will be kept in the analysis since vessels not close by the bridge by definition

will not be near misses. The minimal interaction distance with an object can be specified to increase the chance of finding a near miss or accident in the data set. Clipping a portion of complete trajectories is considered a more effective method than filtering of data points in a specific area, clipping results in more consistent trajectory paths.

Outlier removal

Once the group of trips in the area of interest is complete, single trajectories are examined on potential outliers in the data. This step aims to assess irregular data points within each trajectory. In general, the geographic position is logged quite well by the system, but some errors occur, which will affect calculated velocities or other trajectory properties. Two subsequent data points, i and i+1, are evaluated along a complete trajectory with the following process, as described in a Python notebook by van der Werff (2021):

The distance ($\Delta 1$) between two subsequent data points, i and $i+1$, is calculated using the Euclidean distance formula:

$$\Delta 1 = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$$

Followed by calculating the distance from point i to the midpoint between i and $i+1$ ($\Delta 2$):

$$\Delta 2 = \sqrt{(x_i - M_{i_x})^2 + (y_i - M_{i_y})^2}$$

$\Delta 1$ and $\Delta 2$ are compared with the following criteria:

- If $\Delta 1 > \Delta 2$, the data point (x_i, y_i) is considered valid and is retained.
- If $\Delta 1 \leq \Delta 2$, the data point (x_i, y_i) is identified as an outlier and is removed from the data set.

This process helps identify and remove outliers in the trajectory data, ensuring the quality of the data set.

Add additional parameters and filter

The MovingPandas package has several standard functions that calculate trajectory-based properties, like the distance between two consecutive points, speed, acceleration and the vessel's direction. These will serve as input for the next step of feature generation. Furthermore, the speed and acceleration found with these functions help identify the last outliers in the trajectory data. The vessel speed is limited to 30 m/s, allowing police patrol boats and recreational vessels with high speeds to remain in the dataset. Data points with speeds exceeding 30 m/s are assumed to be outliers due to the extremely high speed. Similarly, an acceleration range of -3.5 to 3.5 m/s^2 is used for outlier detection.

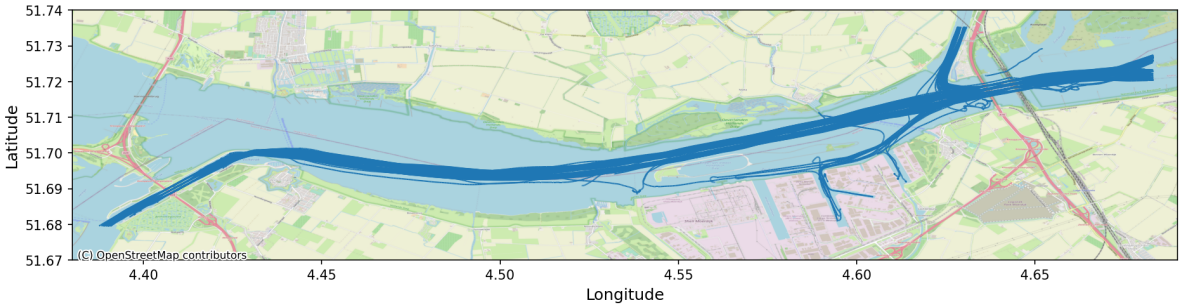


Figure 5.3: Example trajectories after data preparation, Hollands Diep

5.2. Feature engineering

The vessel behaviour, as described in Section 3.1 can now be translated into concrete features extracted from the prepared AIS data.

The subsequent sections explore general features, features defining ship-infrastructure interaction and features extracted with tsfresh.

5.2.1. General features defining vessel behaviour

The features related to general vessel behaviour are described in this section. The following are addressed under general features:

- Vessel type features
- Speed-related features
- Acceleration-related features
- Direction-related features
- Manoeuvre-related features

Vessel type features

For different types of vessels, different behaviour is expected. Therefore, a distinction is made between recreational, inland and other based on the ERI code in the AIS data. A complete list of the vesseltypeERI codes assign to inland vessels can be found in Table A.2.

Categorical encoding is used to create the vessel type feature, relying on the variable type, each vessel either belongs to a category or not. Employing binary encoding, a binary output of 1 or 0 is generated. For instance, a vessel associated with an ERI code designating inland vessels receives a 1 for the vessel type inland feature, automatically resulting in a 0 for the recreational vessel type feature.

This process ensures that each vessel is categorised or not within specific types, allowing for a clear binary distinction based on their characteristics or designated categories.

Speed-related features

Speed-related features are derived from the calculated speed using the MovingPandas package, which considers the distance and time interval between two consecutive data points in the AIS logs to calculate the speed.

Speed is a rational variable, allowing for the comparison of various aspects of speed, such as maximum, minimum, median and standard deviation, which define vessel behaviour.

Acceleration-related features

The acceleration-related features are comparable to the speed-related. Acceleration is calculated by the AIS logs, and various statistics, including maxima, minima, median and standard deviation, are extracted.

Direction-related features

Direction-related features are generated from the course over ground (CoG) variable in the AIS data. The features include minimum, maximum, standard deviation and median, which provide insight into overall directions and their distribution during the trips

Manoeuvring-related features

Insight into manoeuvre-related aspects is based on the rate of turn (ROT) and CoG difference over the course of the trips. The CoG difference is calculated based on the consecutive AIS logs, and features are extracted from the calculated value.

The ROT is one of the logged variables in the AIS data but is regarded as unreliable in this data set, as explained in Subsection 4.2.2. In the study conducted by Mestl et al. (2016), the ROT value just prior to a collision is examined, resulting in the observation of peaks before the collision. The ROT value is derived from the subsequent calculated heading values of the vessel and the time difference between two consecutive points, expressed as:

$$ROT = \frac{\Delta heading}{\Delta time} \quad [Deg/min] \tag{5.1}$$

In the research, the observed peaks in the ROT do not exceed 200 deg/min. Therefore, in the calculation, a boundary value of 250 and -250 is applied, with values exceeding this range assumed to be outliers. Furthermore, the ROT value at low speed or for stationary vessels are not calculated, because Felski and Jaskolski (2013) shows the absence of credible ROT values, especially at a low speed or stationary. The following features are based on the ROT: the maximum, minimum, median and standard deviation.

Additionally, insight into vessel manoeuvrability is provided by the length-to-beam ratio, which, in this case, is calculated by dividing the vessel's length by its width as provided in the AIS data:

$$L/B\ ratio = \frac{lenght}{width} \quad [-] \tag{5.2}$$

A lower L/B ratio indicates higher manoeuvrability. For example, tugboats have a low L/B ratio of 2.5 to 3, while container vessels tend to have much higher L/B ratios, resulting in good course directional stability. A numerical simulation study on standard manoeuvres for vessels of different dimensions suggests better turning performance for vessels with a lower L/B ratio, often in combination with multiple other ship dimensions, which play an significant role (Pérez & Clemente, 2007). In this context, the length-to-beam ratio is considered to be a good indicative feature for manoeuvrability. A complete overview of the general features defined is found in Table E.1.

5.2.2. Features defining ship-infrastructure interaction

The features for ship-infrastructure should describe the vessel's behaviour concerning a stationary point, the infrastructure, or an object in a specific case.

- Minimal distance feature
- Nearby object features

Minimal distance feature

The interaction is primarily defined by the minimal distance with respect to an object or infrastructure. The distance is computed by measuring the perpendicular distance between the vessel's trajectory path and a point object from the shapely package. The trajectories are generated from the AIS data points, the position of the transmitter, the actual vessel dimensions are not taken into account.

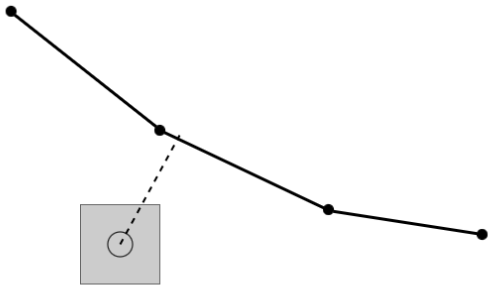


Figure 5.4: Example minimal distance trajectory to object, path represented by solid line between data points and minimal distance indicated by the dashed line

An example is given in Figure 5.4, where the trajectory path between data points is shown as a solid black line. The square represents a bridge pillar, with the circle as a shapely point object marking the specific location. The dashed line in the figure denotes the minimal distance between the trajectory path and the point.

Nearby object features

As indicated in Subsection 3.1.2, a skipper usually positions their vessel well in advance of any interaction with an object or infrastructure. Therefore, changes in their trajectory near the object can provide insight into last-minute changes or potentially dangerous situations. The trajectories are clipped in

an area close to the object to determine the actions before interaction. Two areas are defined: one extending 200 meters in front of and behind the object and another twice as large, covering 400 meters, displayed in Figure 5.5.

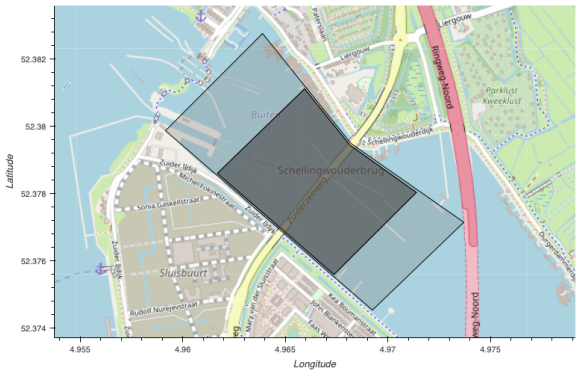


Figure 5.5: Areas around Schellingwouderbrug to define nearby object features, dark part for 200 meters and lighter part for 400 meters

Several features are extracted from these clipped trajectory paths, similar to those in the general features. Parameters like the minimum, maximum, median and standard deviation for speed, acceleration, CoG difference and ROT are determined for the 200 and 400-meter paths. All features related to the ship-infrastructure interaction are presented in Table E.2.

5.2.3. Features extracted with tsfresh

Additionally to the manually created features, features can be extracted by software. In this case, the python package tsfresh by Christ et al. (2019) is used. This package automatically calculates a wide range of time series characteristics, referred to as features, from the AIS logs.

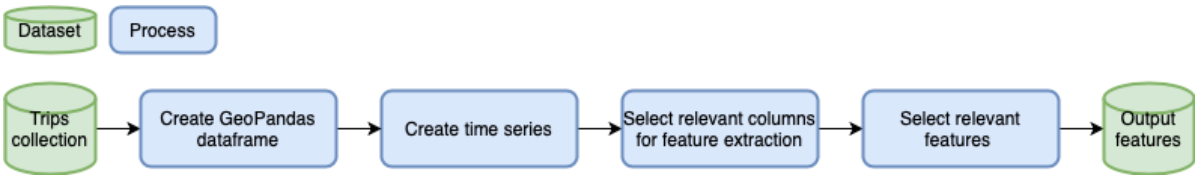


Figure 5.6: Workflow tsfresh feature extraction

Process

- During the data preprocessing stages, the final output consists of a Movingpandas trajectory collection containing all individual trips. Tsfresh, however, requires a different input format, namely a time series data frame. Therefore, the trajectory collection is geospatially encoded back to a GeoPandas data frame.
- With the input for tsfresh prepared, the relevant columns need to be determined, as not all columns contain variables suitable for feature extraction. For instance, the vessel type column remains constant over time.
- In this case, four columns of the data frame are selected for feature extraction: speed, acceleration, CoG, and ROT. These four input columns result in a total of 3132 columns generated by tsfresh, sorted by every trajectory. Thus, a set of 3132 individual features is generated for the time series of each input trajectory, explained in more depth in the following section.
- The subsequent step is identifying which generated features are relevant to consider in the coming steps. The main features extracted are already named in the general features section: the maximum, minimum, median, standard deviation, and quantiles.

Selected features

The features extracted with tsfresh contain several statistics, including the 0.1 and 0.9 quantiles of the input values and features that track the number of peaks within the time series. Monitoring the number

of peaks throughout the time series is of interest, as peaks in speed, acceleration, and ROT suggest fluctuating values, indicating potential anomalous behaviour. Spikes or irregularities in these metrics might signify abrupt or unusual changes in the vessel's trajectory, drawing attention to instances that deviate from expected or standard behaviour.

Peaks are found as follows, tsfresh scans through the time series and counts the number of peaks, where each peak is characterised by the following condition: A value is recognised as a peak if it surpasses its 'n' neighbouring data points to both the left and right within the time series. In other words, a peak is a data point that stands out prominently in comparison to its immediate surroundings. Various 'n' values such as 1, 3, 5, 10 and 50 are used in this research (Christ et al., 2019). An example illustrates this concept with the following 'time series' x:

$$x = [3, 0, 0, 4, 0, 0, 13]$$

4 is a peak of support 1 and 3 because in the sub sequences:

$$[0, 4, 0]$$

$$[0, 0, 4, 0, 0]$$

4 is still the highest value. However, 4 is not a peak of support 3 because 13 is the 3rd neighbour to the right of 4 and is bigger than 4. In this research the support values of 1, 3, 5, 10 and, 50 are used, as outlined in Table E.3, where all features extracted with tsfresh are displayed.

5.3. Generate embedding using UMAP

After feature generation process detailed in Section 5.2, all vessel trips are characterised by the same set of features, found in the tables in Appendix E. This approach results in a matrix where individual rows represent unique vessel trips and the columns contain the features of vessel behaviour. An example of a part of this feature table is presented in Figure 5.7, displaying five unique trips, in this case seven features are visible.

	acceleration__median	acceleration__standard_deviation	acceleration__maximum	acceleration__minimum	speed_near_bridge_min_400	speed_near_bridge_std_200	speed_near_bridge_std_400
0	-0.000373	0.041355	0.148813	-0.073204	3.325548	0.408971	0.285895
1	-0.000180	0.050178	0.144424	-0.148102	2.978321	0.394603	0.265778
2	0.001595	0.048428	0.124576	-0.101008	2.900958	0.357627	0.311158
3	-0.000431	0.096781	0.330608	-0.188851	3.000917	0.668455	0.591327
4	-0.001743	0.682086	2.793904	-0.841787	3.123493	0.986828	1.002060

Figure 5.7: Example feature table for 5 vessel trips, columns represent the features

Using the feature table, each trip can be compared based on all unique features. For instance, the standard deviation of the speed 400 meters around the bridge can be examined, as demonstrated in the last column of the table in Figure 5.7. While comparing all individual trip characteristics can provide insights into vessel behaviour, this process is time-consuming and lacks a clear linkage between the features.

Therefore, the Uniform Manifold Approximation and Projection (UMAP) algorithm, as introduced and explained in Subsection 3.2.2, is used to create a two-dimensional representation of all features in the table. The process is visualised in Figure 5.8

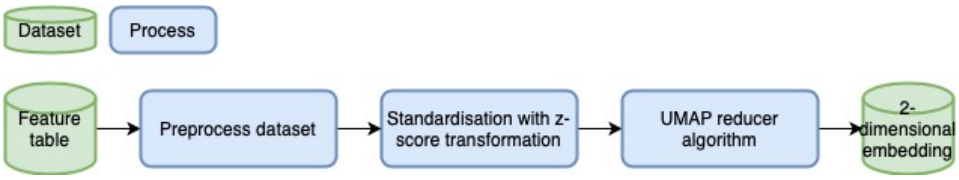


Figure 5.8: Workflow to generate 2-dimensional embedding with UMAP

Process

- Before the algorithm can be applied, some preprocessing is required. The feature data is extracted from the table and needs to be cleaned and standardised. The cleaning involves the removal of NaN values, and replacement with zero values, as some NaN values may result from the feature engineering step.
- Next, the standardisation process is conducted, where the z-score transformation is applied. This transformation involves subtracting the mean from each data point and dividing by the standard deviation. It ensures that all the features have the same scale (mean of 0 and standard deviation of 1). This standardisation is useful for making the features more directly comparable and can be especially important in many machine-learning algorithms.
- Subsequently, the UMAP reducer algorithm is applied, resulting in two dimensions, Embedding_0 and Embedding_1.

A visualisation of the embedding result is presented in Figure 5.9. In this case, input data from 1395 trips on the IJ near the Schellingwouderbrug were used, the trips are represented by 80 features. Each point in the scatter plot represents an individual trip, with Embedding_0 on the x-axis and Embedding_1 on the y-axis. These two dimensions indicate the two-dimensional representation of all features combined. Both Embedding_0 and Embedding_1 do not have any physical quantities.

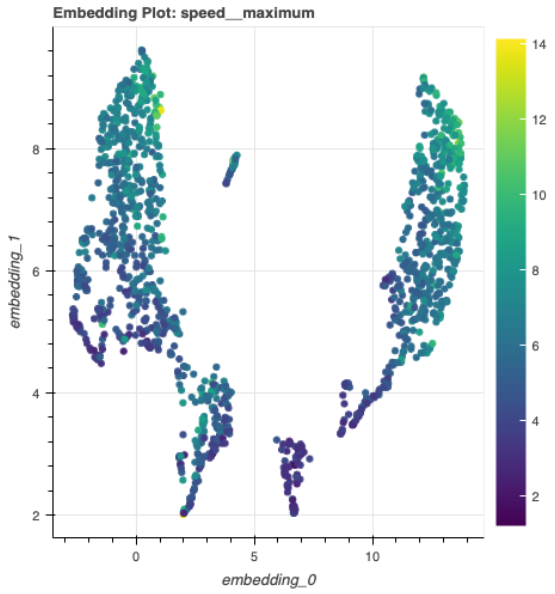


Figure 5.9: Embedding example, colour bar indicating maximum vessel speed over the trajectory [m/s]

A colour scale is applied based on one of the underlying features: the vessel's maximum speed during the trip. Here, a wide range is observed from around 2 m/s to a vessel with a speed of 14 m/s. By examining the distribution of the colours over the points, areas with a higher or lower speed can be identified. Further explanation of the shapes and patters found in the embedding are possible by plotting underlying characteristics and the trajectories of the vessel trips.

5.4. Clustering with K-means

The output of the UMAP dimension reduction is two dimensional embedding, the subsequent step is to cluster this outcome to group similar behaviour patterns into specific clusters, enabling naming of the clusters.

5.4.1. K-means clustering

A two-dimensional embedding is obtained as the output of the UMAP dimension reduction process. The subsequent step is to cluster this outcome to group similar behaviour patterns into specific clusters for identification. The K-means clustering method is applied, and its basic principles are explained in

Subsection 3.2.3. The objective is to group data points, in this case, the trips with underlying features, into 'k' clusters, with each cluster containing data points that are more similar to each other than to points in other clusters.

The output of UMAP is a 2-dimensional array of x and y values, serving as the input of the K-means algorithm. Additionally, hyperparameters such as 'k' the number of clusters, the method for initialising cluster centroids, and the number of times the algorithm will run with different initial centroids to find the best clustering results must be defined. The initialising method is k-means++ from Pedregosa et al. (2011), which selects one random centroid and places the others on the maximum squared distance to push the centroids as far as possible from one another (Bahmani et al., 2012). The number of times the algorithm runs is set to ten.

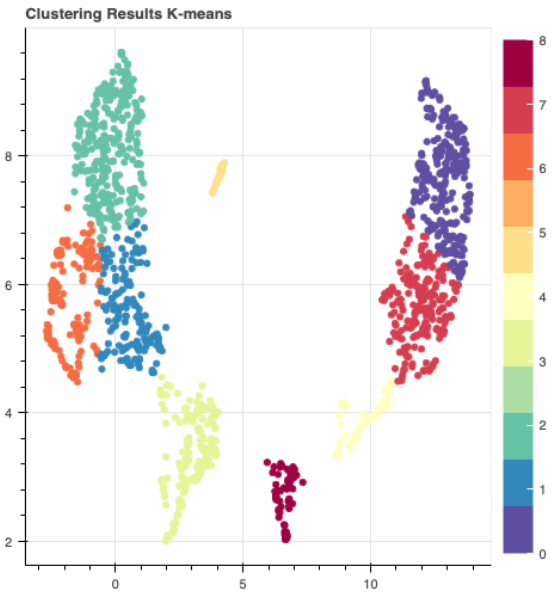


Figure 5.10: Clustering example, colour coded per cluster

In Figure 5.10, the clustering is applied to the example embedding generated in Section 5.3 , resulting in nine clusters based on the elbow method, Table 5.1 presents the scores on clustering results.

Table 5.1: Clustering performance indicators with scores

Silhouette Coefficient	Calinski-Harabasz Index	Davies-Bouldin Index
0.52	7671	0.60

The silhouette score is bounded between -1 and 1, the closer to 1, the better the clustering; a score of 0.52 is seen as an appropriate score. For the Calinski-Harabasz index, the higher the score means denser and more separated clusters. For the Davies-Bouldin Index, zero is the lowest possible score, values closer to zero indicate better clustering; 0.6 is found for nine clusters.

5.4.2. Naming and understanding the clustering

The following step involves understanding the clustering results by examining the characteristics of each cluster to understand what distinguishes one cluster from another. Since the input data is unlabelled, this is achieved by comparing the underlying data. The two primary approaches involve visually exploring the trajectory plots per cluster and inspecting the underlying data of individual features.

Visual examination of trajectory plots per cluster

Visualising the trajectory plots of all trips within a single cluster can provide insight into the main route patterns. The general expectation is that points within the cluster exhibit similar behaviour, so

the patters should exhibit similarities as well. These plots may reveal varying navigating directions across clusters, specific paths unique to each cluster, or a lack of distinct paths, resulting in clusters with vessel paths in all directions, covering the entire waterway.

Examining underlying individual features over the clusters

Since the underlying feature information is retained through the previous steps of dimension reduction and clustering, it can be reintegrated into the clusters. For instance, examining vessel types to determine if vessels of the same type are consistently grouped in the same clusters.

Furthermore, distributions of speed, accelerations can be used to categorise clusters as fast or slow. Changes in Rate of Turn (ROT) or Course over Ground (CoG) values might indicate groups of vessels making turns. Visualisations of individual features will be generated to identify their distribution across different clusters. The effects of individual features are, beforehand, unknown, making all underlying data potentially valuable for understanding the clustering results.

Part II

Results

6
Results

In this chapter, the results obtained by the proposed methodology will be presented. The behaviour characteristics, which were defined in Chapter 3 and subsequently converted into the features described in Chapter 5, are employed for the detection of anomalies in the vessel trajectories generated from the AIS data sets. The method is applied to both Hollands Diep and the IJ regions.

The examination of clustering results emphasises the distribution of vessel types among clusters and their corresponding navigation routes. The study focuses on ship-infrastructure interactions, particularly the interaction with bridges. Two distinct bridges are discussed: the Moerdijkbrug, situated in the Hollands Diep region, and the Schellingwouderbrug in the IJ area. As described in section 3.3, these two bridges have different structures, but the approach to identifying vessel behaviour remains the same. The most important feature is the minimal distance of the vessel with respect to the structure. Furthermore, velocity, acceleration and the rate of turn near the bridge will be examined. Two sections are defined, one covering a 200-meter distance around the bridge and the other extending to 400 meters.

6.1. Hollands Diep

The trajectories on the Hollands Diep were clipped near the Moerdijkbrug to exclude the behaviour further upstream and downstream that might affect the clustering but are unrelated to the interaction with the bridge, visualised in Figure 6.1.

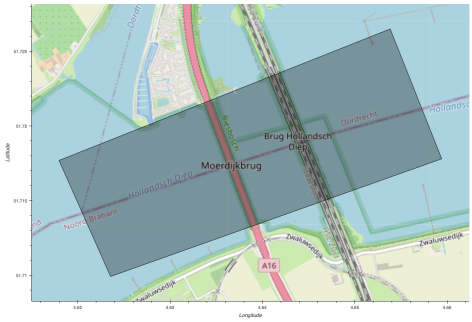


Figure 6.1: Area near Moerdijkbrug where trajectories have been clipped

The initial data set contains 2322 individual vessel trips passing the Moerdijkbrug. Further details on trip lengths and vessel types are provided in Subsection F.1.1.

The dataset undergoes dimension reduction using UMAP, followed by clustering via K-means into 11 clusters. The selection of 11 clusters is based on the elbow method, which is visualised in Figure 6.2. After a number of cluster of 11, there is little significant change in the Within-Cluster Sum of Squares (WCSS). Table 6.1 provides insight into the additional indexes used to evaluate the clustering results. For instance, the silhouette coefficient, with a score of 0.49, is considered a positive indicator, suggesting well separated clusters. Further plots illustrating the relationship between the number of clusters and the evaluation metrics can be found in Figure F.3 in Appendix F.

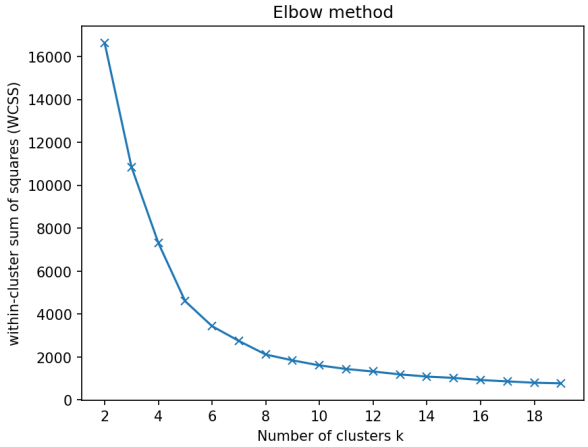


Figure 6.2: Elbow plot clustering to determine number of clusters 'k', Hollands Diep

Table 6.1: Clustering performance indicators with scores for 11 clusters, Hollands Diep

Method	Score
silhouette Coefficient	0.49
Calinski-Harabasz Index	18239
Davies-Bouldin Index	0.67

The clustered trips can be visualised in a scatter plot, each point on the plot in Figure 6.3 represents a single trip, with all relevant features as the underlying data. The points are color-coded to represent one of the eleven clusters and the horizontal and vertical axis do not have a physical quantity.

Before diving into the individual trip details, the clusters are examined to gain a more general insight into the generated clusters. It is anticipated that vessels of the same type will generally exhibit similar behaviour, and the trips displaying similar patterns are expected to be grouped together. Figure 6.4 illustrates the distribution of the vessel types across the generated clusters.

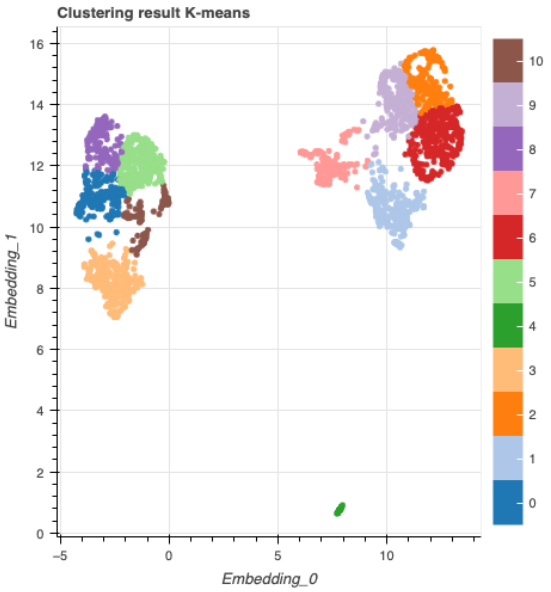


Figure 6.3: Scatter plot K-means clustering at Hollands Diep, colour coded per cluster

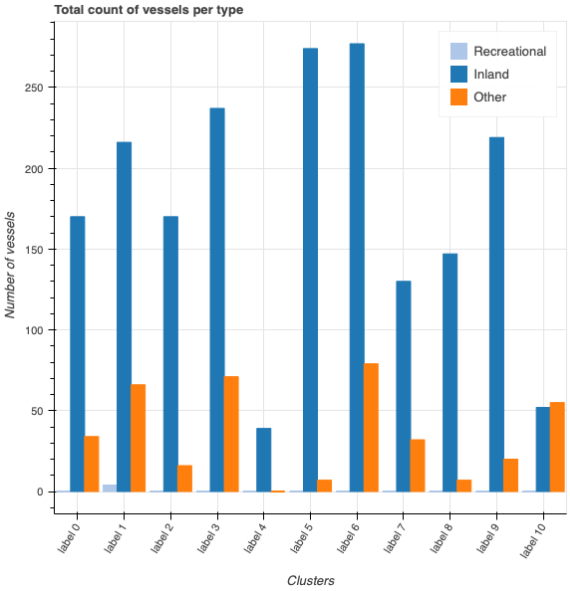


Figure 6.4: Distribution of vessel types across the clusters, Hollands Diep

For instance, all recreational vessel trips, four in total, are found exclusively in the cluster with label 1. However, a distinct separation is not visible for inland vessels and other types. An exception is the cluster with label 4, a relatively small cluster compared to the others, which only includes inland vessels. Such a small cluster with a specific vessel type can be interesting and is investigated in more depth.

Visual exploration of trajectories per cluster

Plotting all individual trips within each cluster enables comparison of the vessel patterns and gives insight into the different groups. In Figure 6.5, the trips navigating under the Moerdijkbrug, where the

pillars are visualised by the black squares, are plotted for all clusters. Furthermore, the green square represents the start point, and the red circle represents the endpoint of the trip.

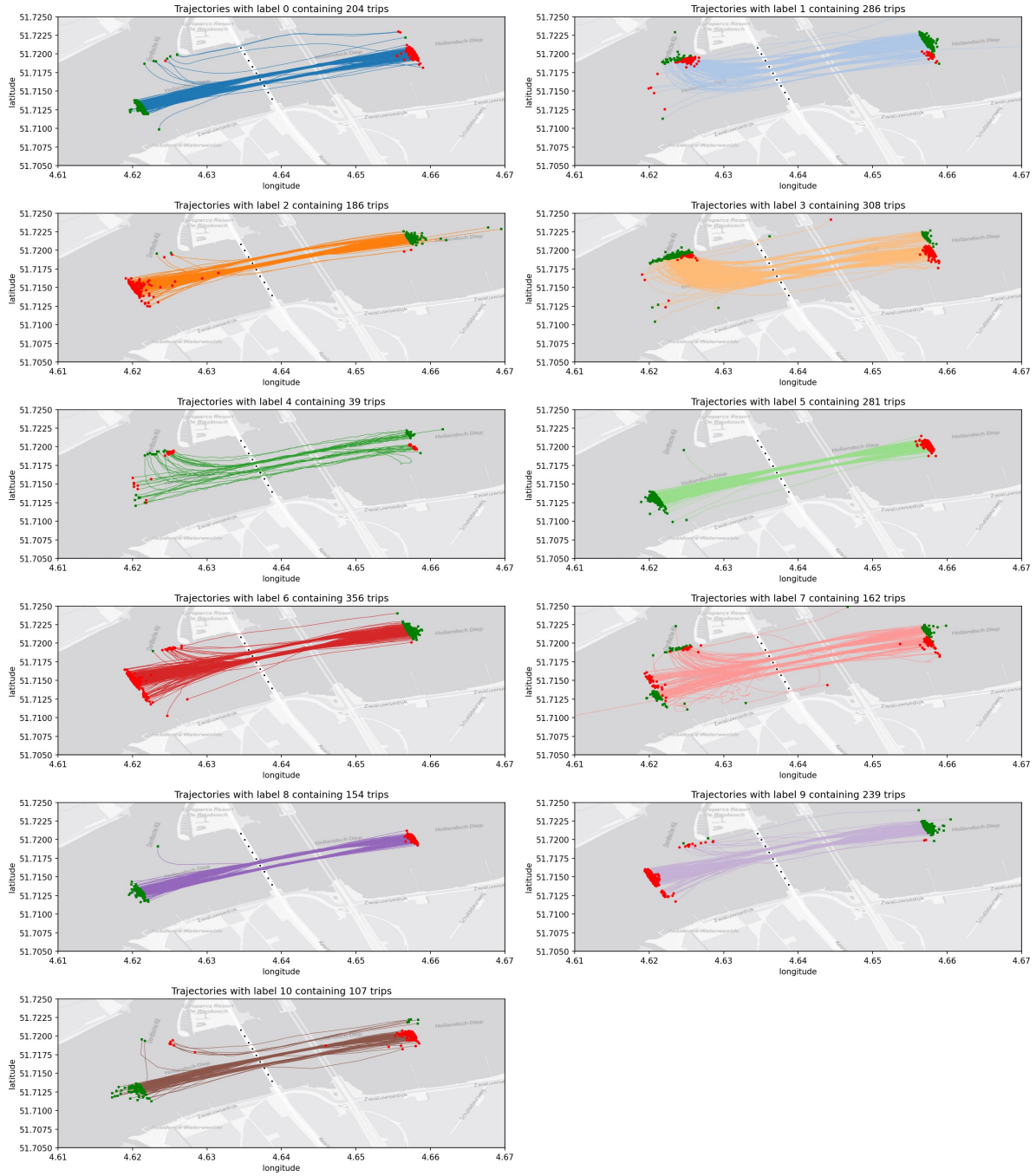


Figure 6.5: Trajectory plots per cluster, Hollands Diep

Generally, two main patterns can be noted over the clusters: vessels sailing from left to right crossing the bridge through the southern openings and vessels navigating the other direction, passing via the northern openings. Exceptions from this are clusters with labels 1, 3, 4 and 7, with no distinct consequent route. Observing the route of these trips, many vessels are turning toward or from the Dordtsche Kil.

The split in two main route patterns found by plotting the trips per cluster can be recognised on the scatter plot in Figure 6.3. Here, two main groups are clearly visible, one on the left and the other on the right side, where the vessels in the clusters on the left navigate from west to east and clusters on

the right side the other way around, indicated in Figure 6.6 by blue and orange, respectively.

More in-depth towards the clusters on the left side of the plot, on the top, cluster 8 is found and cluster 3 on the bottom, where cluster 8 has more straight trips of vessels following the Hollands Diep, and in cluster 3 the vessels make a turn from the Dordtsche Kil towards the Hollands Diep. A similar distinction is seen with the clusters on the right side of the plot. Furthermore, clusters 4 and 7 are found in the middle of the scatter plot, indicating the mixed patterns coloured purple.

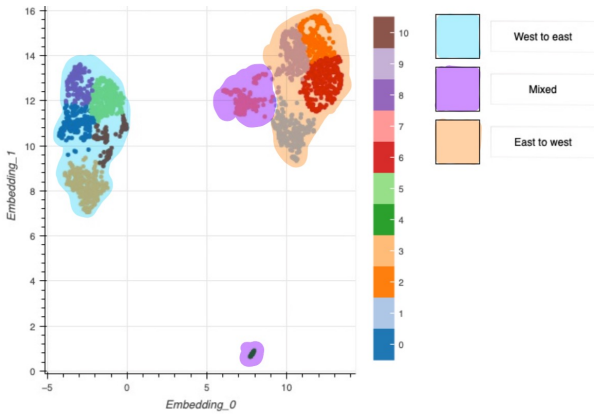


Figure 6.6: Identification of main vessel direction over the clusters, Hollands Diep

Further notable in the scatter plot is the position of cluster 4 relative to the other clusters, namely, at the bottom of the plot. Indicating different characteristics compared to other trips. With a closer look at the routes in figure Figure 6.7, a zigzag pattern is visible, deviating from the patterns of other vessels.

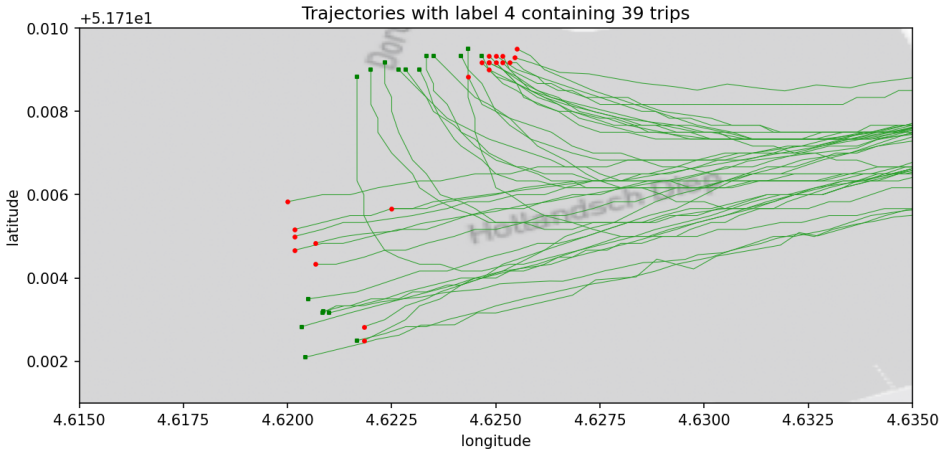


Figure 6.7: Zoom on trips in cluster 4, with zigzag patterns, Hollands Diep

The last cluster, which stands out based on the patterns and location in the embedding, cluster 7, is shown in Figure 6.8. Within these vessel paths, there is a significant amount of activity on the south side of the river compared to the other clusters. Vessels appear to remain stationary at the west side for a brief period before passing the bridge. This can be seen as dangerous behaviour since this is not an official mooring location. Conversely, the east side serves as a temporary anchorage area.

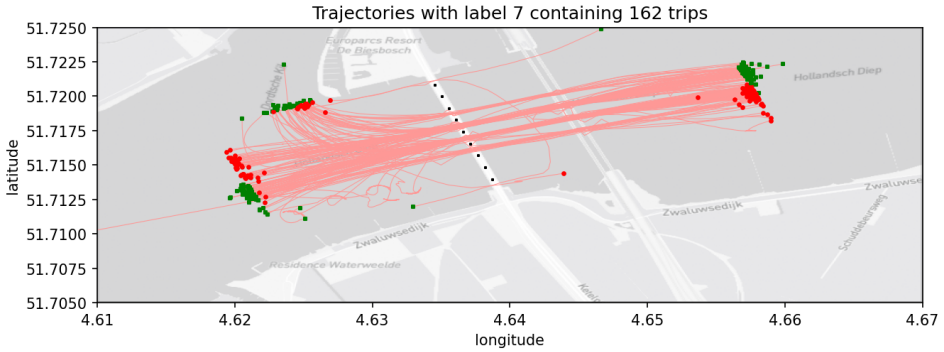


Figure 6.8: Trips in cluster 7, Numerous deviating patterns in cluster, Hollands Diep

Additionally, an interesting path crossing the river can be observed between the two bridges, as depicted in Figure 6.9. The colour bar on the figure represents the vessel's speed in m/s. Such a manoeuvre could be considered anomalous and potentially dangerous. However, after investigating the vessel type and dimensions, it was found that this path was made by a police patrol boat, making the manoeuvre acceptable.

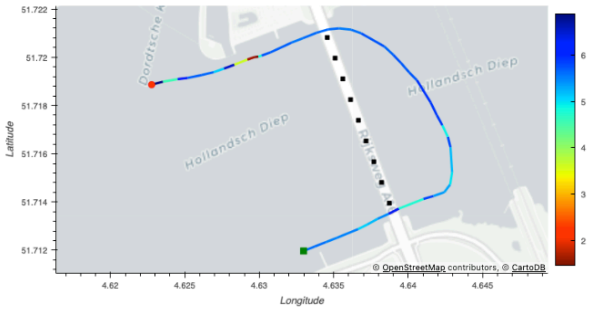


Figure 6.9: Trajectory plot trip 1502: path perpendicular to river between bridges, colour bar indicating vessel speed [m/s], Hollands Diep

Individual feature exploration

With the patterns and clusters visually examined, the underlying features of all trips will now be investigated. Figure 6.11 visualises the minimal distance relative to the bridge pillars. In this representation, all individual data points from Figure 6.10 are plotted in their corresponding cluster, represented on the horizontal axis, providing a concise overview of a single feature and its distribution across the clusters. The specification of the selected feature is depicted on the y-axis, in this case the minimal distance with respect to one of the bridge pillars in meters.

Most of the distance values fall within the 20 to 50 meters range, which aligns with expectations given that the distance between two pillars is 100 meters, and vessels typically navigate through the centre of this opening.

However, a couple of points stand out in this figure with a small relative distance, one in particular: the black circled dot within cluster 7. In this case, the distance from the AIS transmitter to one of the bridge pillars measures only 3 meters. This close encounter between the vessel and the bridge suggest a potential near miss. This data point is also highlighted in the scatter plot. As a result, this specific trip will be investigated in more depth.

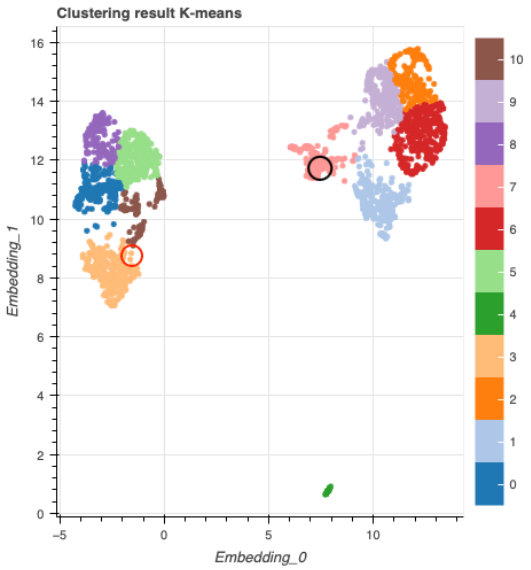


Figure 6.10: Scatter plot K-means clustering trips, trip 2236 indicated by black circle, trips 1429 marked by the red circle, Hollands Diep

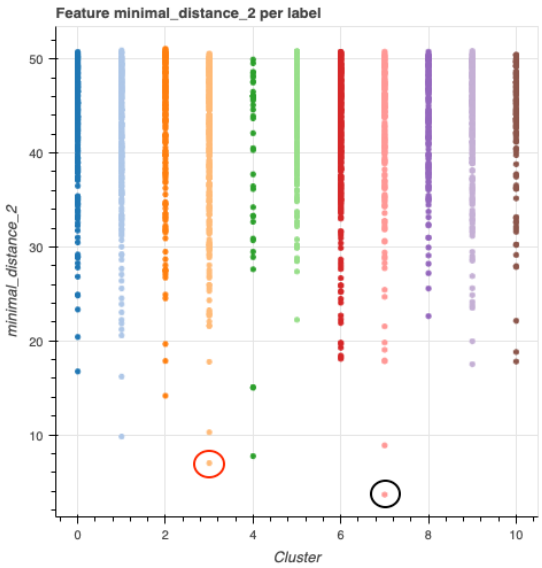


Figure 6.11: Jitter plot per cluster minimal distance to Moerdijkbrug, trip 2236 indicated by black circle, trip 1429 marked by the red circle, Hollands Diep

Figure 6.12 displays the trajectory of trip 2236 from the data set, which has been marked suspicious based on the minimal distance. In this plot, the vessel's journey starts at the green square and ends at the red circle, with the path colour indicating vessel speed. The underlying AIS data reveals that this trip was made by an vessel with an ERI code of 8440, a passenger ship or cruise ship. A remarkable pattern around the bridge is visible, where a real close encounter with the bridge pillar, indicated by the black squares can be seen. Given the vessel's dimensions, measuring 110 meters in length and 12 meters in width, such a pattern can be considered anomalous. Figure F.4 in the Appendix contains a closer look towards the path near the pillar.

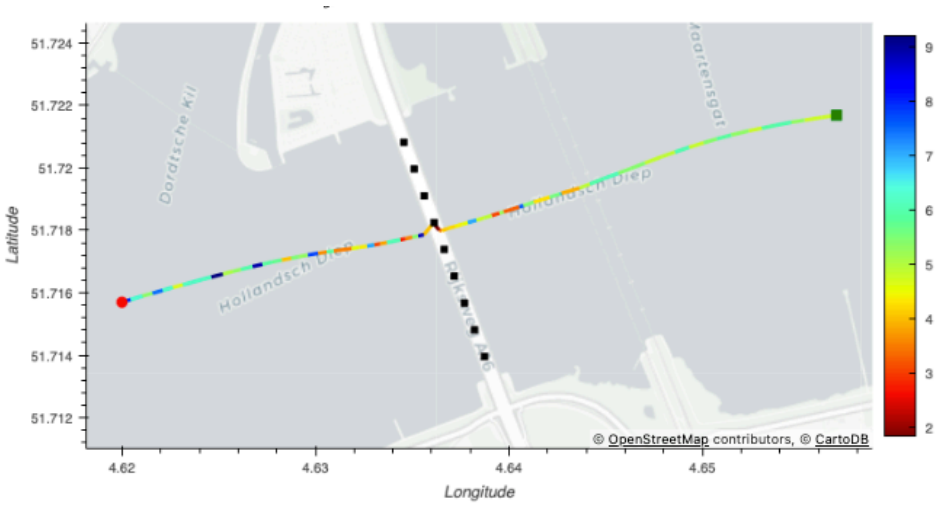


Figure 6.12: Trajectory plot trip 2236 near Moerdijkbrug, colour bar indicating vessel speed [m/s], Hollands Diep

In accordance with the concept that points closely situated in the embedding share similar behaviour, exploring these points in the scatter plot can be of interest. A zoom-in on trip 2236 has been executed, and the nearby points examined.

Four points in close proximity to trip 2236 within the clustering are found. A closer examination of the path plots indicates that the behaviour is not precisely alike. Among these paths, only one

displays a minor route deviation under the bridge, similar to trip 2236, and has an overall similar route. Comparing all four routes only their sailing direction is matching, navigating from east to west. Although the endpoints of the two trajectories differ. One turns toward the Dortsche Kil and the other bends more towards the south side of the river. An overview of the four trajectory plots can be found in Figure F.6.

Investigating another point with a small distance to the bridge pillar, the focus is placed on the point with a red circle within cluster 3, as depicted in both Figure 6.10 and Figure 6.11. A minimal distance of 6.5 meters to the pillar is found, indicating an extremely close encounter based on the available data. In Figure F.7, located in the appendix, a trajectory plot illustrates a path with relatively few data points close to the bridge. While this may create the appearance of the trajectory coming remarkably close to the pillar, it may not necessarily be the case while data points may be missing.

Thus far, exploration has focused solely on the minimal distance feature, while acceleration or changes in the rate of turn (ROT) close to the bridge were identified as interesting features to describe the interaction with infrastructure. Figure 6.13 and Figure 6.15 depict the maximum acceleration in m/s and the ROT in deg/min in the area 400 meters around the bridge, respectively.

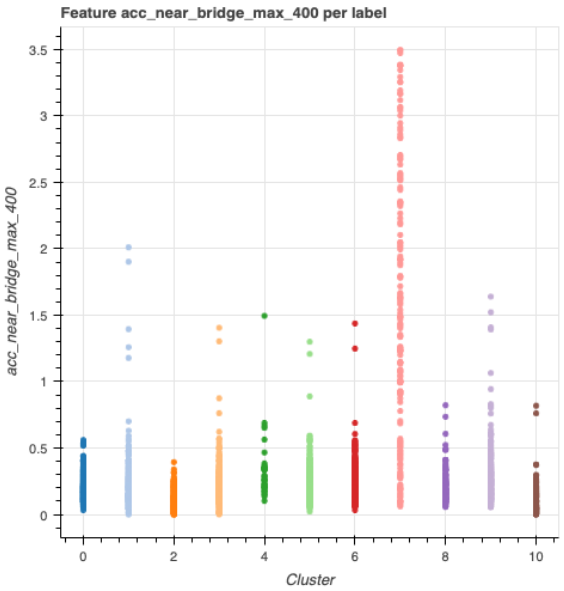


Figure 6.13: Jitter plot per cluster max acceleration [m/s^2] 400 meter around the Moerdijkbrug, Hollands Diep

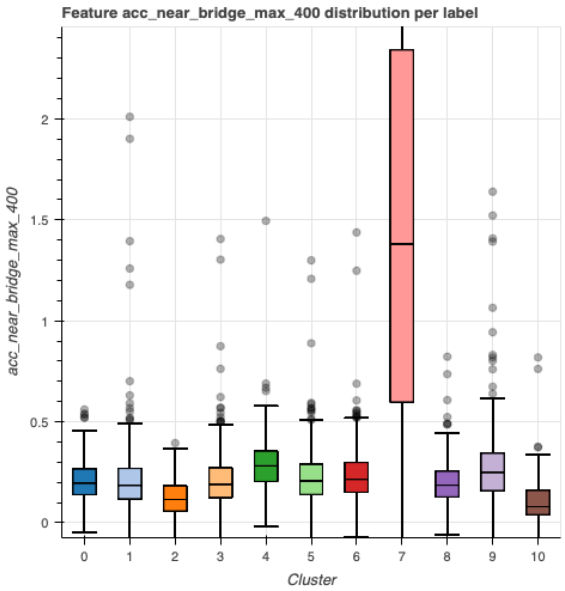


Figure 6.14: Box plot per cluster max acceleration [m/s^2] 400 meter around the Moerdijkbrug, Hollands Diep

The maximum acceleration plot illustrated in Figure 6.13 highlights distinctive values within cluster 7. This cluster consistently displays notably high maximum acceleration values across most vessels, setting it apart from the others. A box plot visual representation in Figure 6.14 supports this observation, indicating a significantly higher median acceleration in cluster 7 compared to the remaining clusters. Notably, despite this higher median, the distribution still follows a normal distribution pattern. This aligns with the visual exploration findings, where Cluster 7 stands out as well.

Within a 400-meter radius around the bridge, many vessel tracks exhibit high Rate of Turn (ROT) values. Nevertheless, clusters 2 and 10 maintain ROT values below 100 deg/min, in accordance with their trips following straight paths along the Hollands Diep. While it was expected that cluster 8, which also mainly contains vessels navigating straight routes along the Hollands Diep, would exhibit similar behaviour, their tracks show higher ROT values, as depicted in Figure 6.15.

Another cluster of significance is Cluster 4, where all trips consistently record ROT values of at least 150 deg/min. This cluster's characteristics align with our earlier observations during the visual analysis,

particularly the presence of vessels displaying zigzag patterns, which is in line with the findings related to ROT values.

The next feature under examination is the number of peaks in acceleration, where 'n' is set to 10, indicating that at least 10 data points before and after a given point have a lower acceleration. A jitter plot illustrating the distribution per cluster is given in Figure 6.16. Again, in cluster 7, several trajectories display a distinctive pattern with numerous peaks in the acceleration over their duration. Specifically, four trajectories are characterised by a minimum of 20 peaks in acceleration.

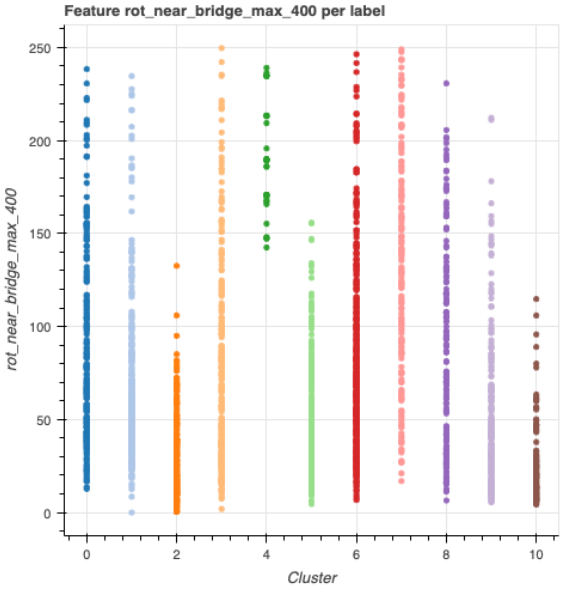


Figure 6.15: Jitter plot per cluster max Rate of Turn [deg/min] 400 meter around the Moerdijkbrug, Hollands Diep

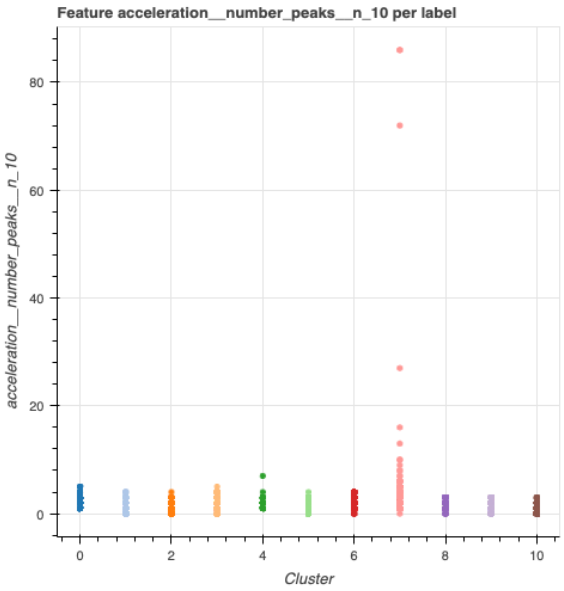


Figure 6.16: Jitter plot per cluster acceleration number of peaks for n=10, Hollands Diep

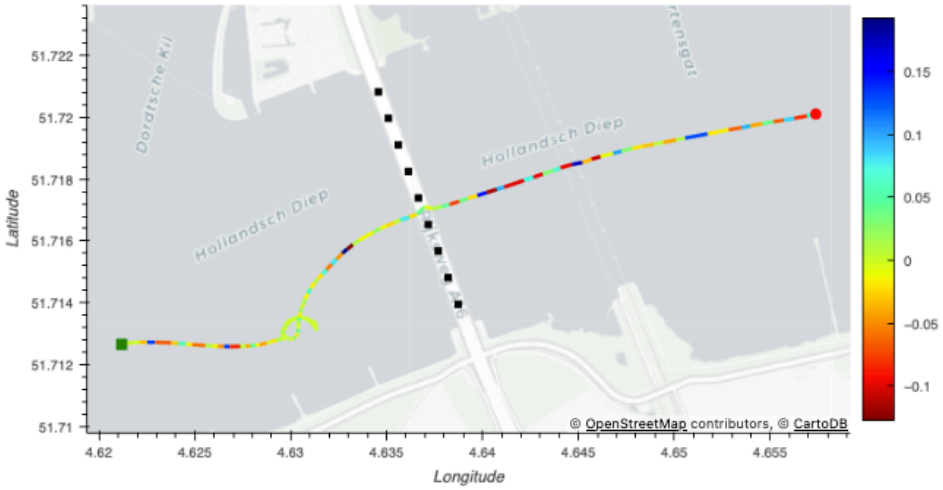


Figure 6.17: Trajectory plot 914 near Moerdijkbrug, colour bar indicating vessel acceleration [m/s^2], Hollands Diep

One noteworthy example is trip 914, where over 80 peaks in the acceleration were observed as plotted in Figure 6.17. However, examining the acceleration profile across the trajectory, no extreme values are detected. Instead, there must be fluctuations over time. Additionally, a nearly stationary pattern is observed at the west side close to the bridge.

A detailed analysis of trajectories with high numbers of peaks in acceleration reveals that all four of these trajectories cluster together, as depicted in Figure F.8. Further inspection of the patterns in these trajectories, shown in Figure F.9, reveals a consistent stationary pattern near the west side of the bridge. This observation suggests that similar behavioural patterns tend to group together in this context.

A more in depth analysis of the number of data points within these four trajectories, it is found that three of them consist of at least 1600 data points, and the last one contains 700 data points. This explains the higher number of peaks found, as an increase in the number of data points is associated with a greater likelihood of encountering more peaks. Comparing this to the rest of the data set, the average trip length is found to be 62 data points, with one of the four trips featuring the highest number of data points found within the entire data set.

6.2. The IJ lock complex included in area

The developed method is applied to the IJ region to investigate the interaction with the Schellingwouderbrug. Trajectories in the IJ section are clipped to focus on paths near the Schellingwouderbrug. Approximately one kilometre upstream and downstream is included, as depicted in Figure 6.18.

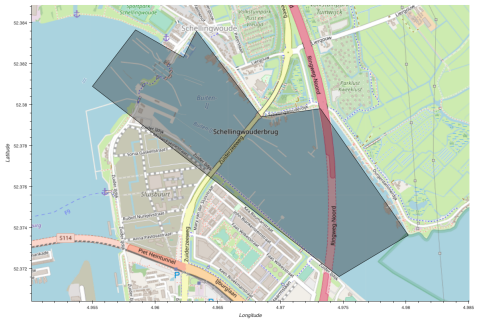


Figure 6.18: Area near Schellingwouderbrug where trajectories have been clipped, lock complex included in the north

In this scenario, 2248 trips are generated in the area and are subsequently clustered into 13 groups based on the elbow method. Further details on trip lengths and vessel types are provided in Subsection F.2.1. The elbow method figure and other clustering scores over the number of clusters can be found in Figure 6.19 and Figure F.12. The scores in Table 6.2 all indicate a weaker clustering compared to the application on the Hollands Diep.

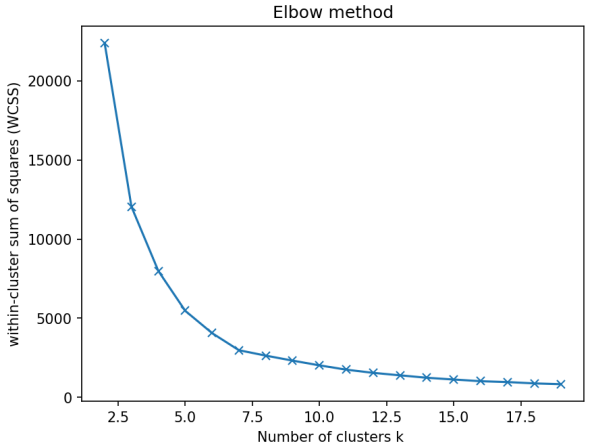


Figure 6.19: Elbow plot clustering to determine number of clusters 'k', the IJ lock complex included

Table 6.2: Clustering performance indicators with scores for 13 clusters, the IJ lock complex included

Method	Score
silhouette Coefficient	0.44
Calinski-Harabasz Index	4888
Davies-Bouldin Index	0.72

The scatter plot displaying all individual trips is shown in Figure 6.20, which will be examined by looking at the trajectory plots per cluster and individual underlying features of all trips.

The distribution of vessel types, as depicted in Figure 6.21, is initially explored, with recreational vessels, with recreational vessels found almost exclusively in cluster 8. An exception is found in cluster 11, where a single recreational vessel is found. Additionally, clusters 1, 9 and 10 mainly contain vessels classified as 'other' types, including unknown types.

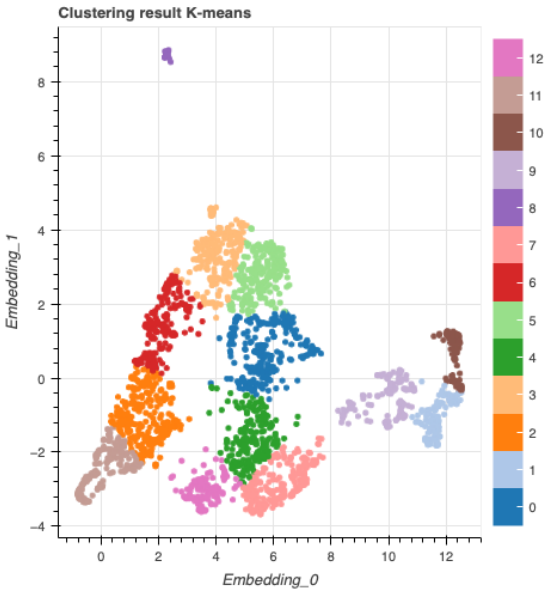


Figure 6.20: Scatter plot K-means clustering, colour coded per cluster, the IJ lock complex included

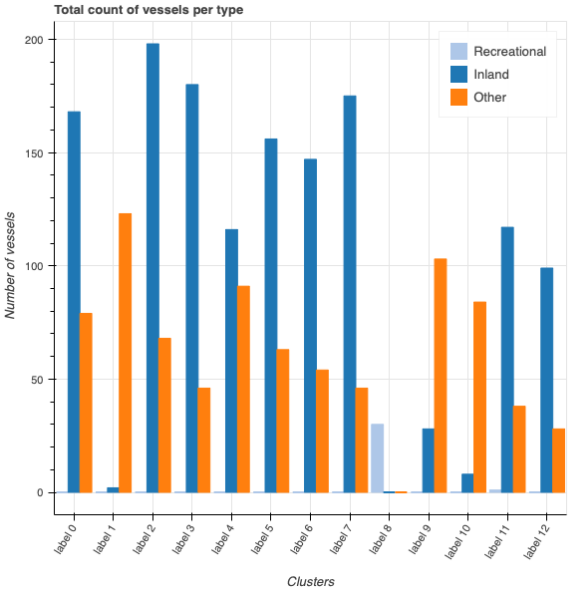


Figure 6.21: Distribution of vessel types across the clusters, the IJ lock complex included

Further investigation of the lone recreational vessel outside the primary 'recreational' cluster reveals that trip 1459 corresponds to this point. The trajectory for this trip is plotted in Figure 6.22, depicting a journey from north to south and a recognisable stop during the trip. Evaluating the surrounding trips in the embedding, shown in Figure F.14 and trajectories plotted in Figure F.15 in Appendix F, reveals similar stopping points. These vessels are identified as motor freighters or motor tankers based on their vessel types.

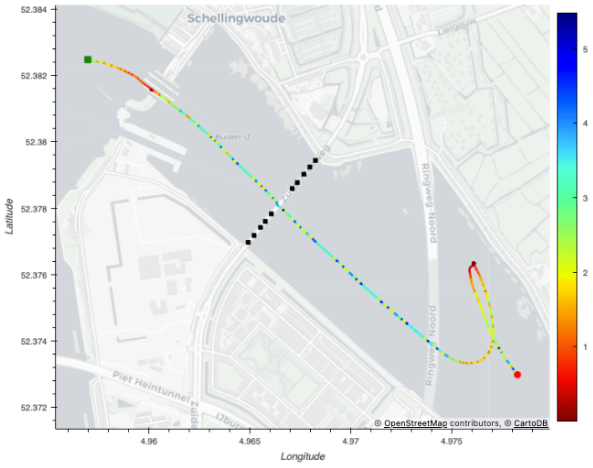


Figure 6.22: Trajectory plot of trip 1459 recreational vessel outside recreational cluster, colour bar indicating vessel speed [m/s], the IJ lock complex included

Visual exploration of trajectories per cluster

Transitioning to visually exploring the trajectories per cluster, the trips are plotted with each cluster represented by a distinct colour similar to the scatter plot, as shown in Figure 6.23.

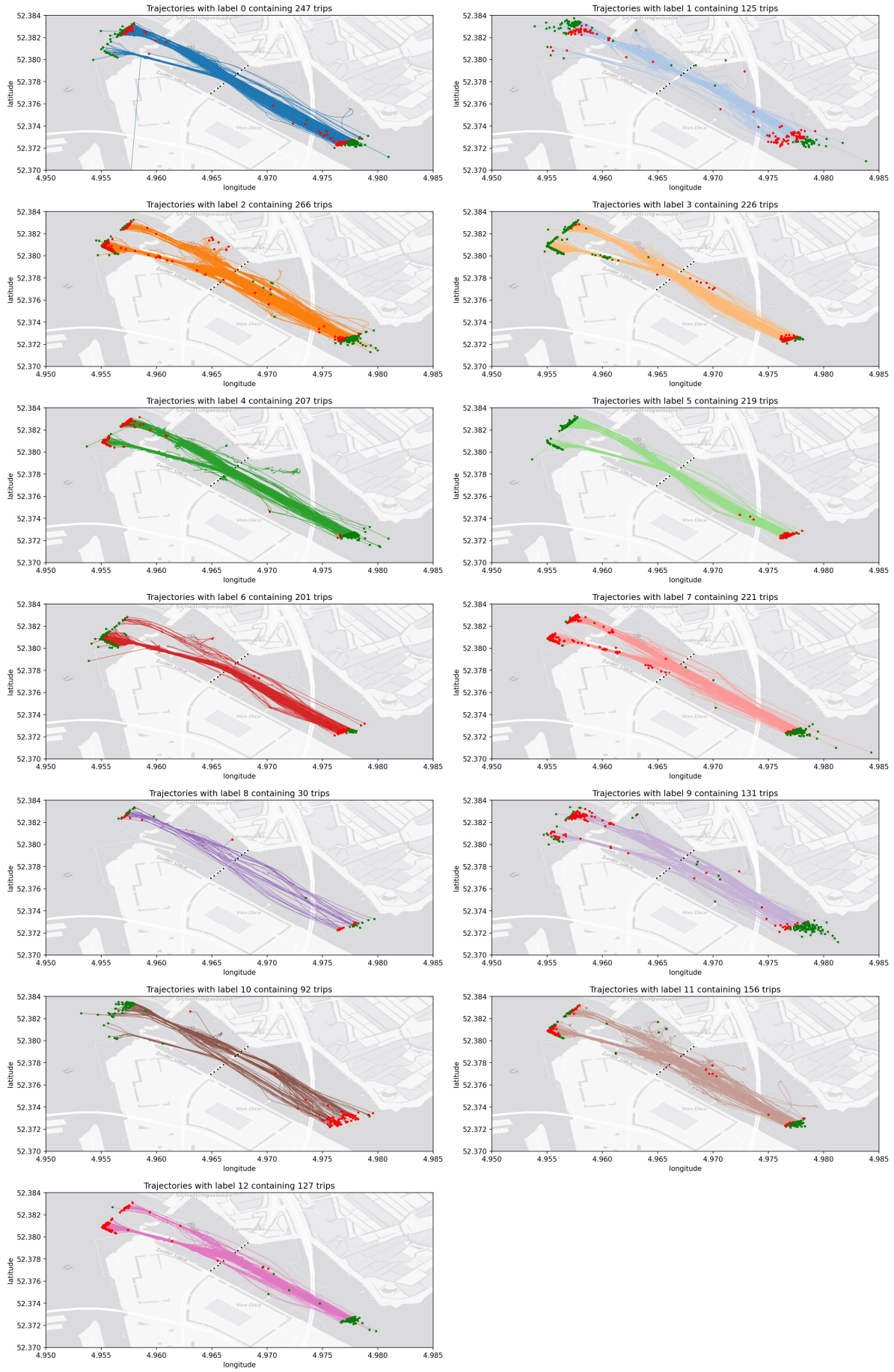


Figure 6.23: Trajectory plots per cluster, the IJ lock complex included

At first, examining the sailing direction, it is observed that only two clusters labelled 10 and 12, exhibit uniform directions, with cluster 10 navigating from north to south and cluster 12 following the opposite route. These clusters contain relatively few trips (92 and 127, respectively). In contrast, the other clusters display mixed sailing patterns, with the clusters at the top of the embedding tending to have routes from north-to-south, while those at the bottom navigate in the opposite direction. Clusters 0, 1, 2, and 9, located in the middle of the embedding, exhibit an even greater variety of directions, similar to the recreational cluster, which is positioned far away from the others. Sailing directions are indicated in Figure 6.24 where blue is used for north to south, purple for mixed directions, and orange for vessels mainly navigating from south to north.

Distinct routes for crossing the bridge can be identified, with the two primary locations being at the fixed part with a large span in the middle and the movable part located between the pillars just north of the wide span. This reflection on the clustering reveals that clusters 1, 9 and 10 mainly use the northern movable part of the bridge, while other clusters have the large span as their primary path. Almost all vessels in cluster 1 navigate through the movable part, as indicated orange in Figure 6.25. The passage underneath the main span is coloured blue and mixed purple.

Furthermore, variations in clusters with more or less straight paths are observed. Some clusters exhibit more endpoints or wiggling paths, such as clusters labelled 1, 2, 4, 7, 9, 11, and 12. Projecting these observations onto the embedding plot results in a lower cluster displaying more deviating paths, indicated by orange, and clusters at the top of the embedding with more smooth paths, indicated in blue, as well as clusters displaying mixed patterns, indicated purple, in Figure 6.26.

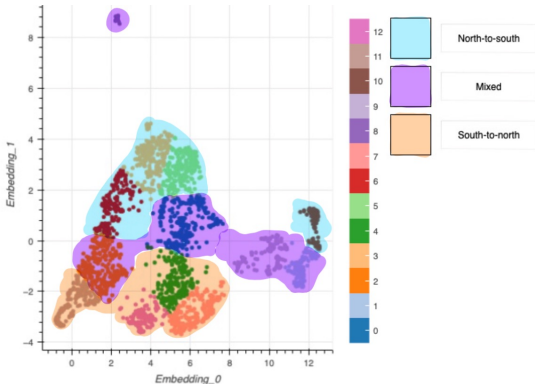


Figure 6.24: Identification of main vessel direction over the clusters, the IJ lock complex included

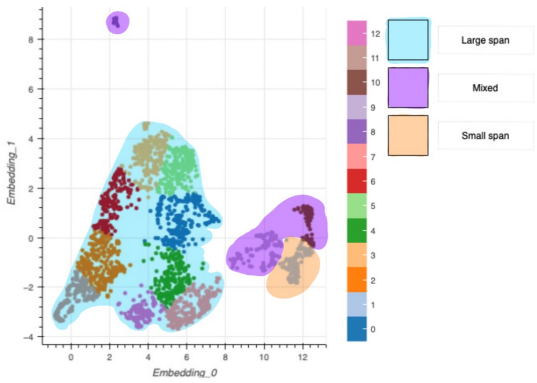


Figure 6.25: Identification of main bridge opening used over the clusters, the IJ lock complex included

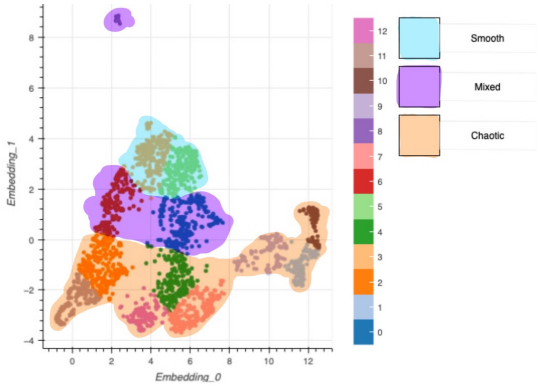


Figure 6.26: Identification of main path style over the clusters, the IJ lock complex included

Another noteworthy distinction appears between clusters 3 and 5, both of which expose similar direction paths while remaining separated in the clustering. Cluster 3 contains stopping points near the lock complex and in front of the bridge, potentially influencing the clustering outcome. It would be interesting to investigate the effect of the lock complex by excluding this part of the vessel paths and assessing the impact on the results. Given that this case focuses on the interaction with the bridge, the lock complex is not a primary area of interest.

In the scatter plot in Figure 6.20 the position of cluster 8 stands out relative to the other clusters, distinctly at the top of the plot, indicating different characteristics compared to other trips. The deviation is further supported by the vessel type distribution depicted in Figure 6.21, where cluster 8 predominantly contain recreational vessels. In contrast to the distinctive zigzag patterns observed in the outlying cluster around the Moerdijkbrug (illustrated in Figure 6.7), no such irregular patterns are evident in the outlying cluster around the IJ, as seen in Figure 6.27.

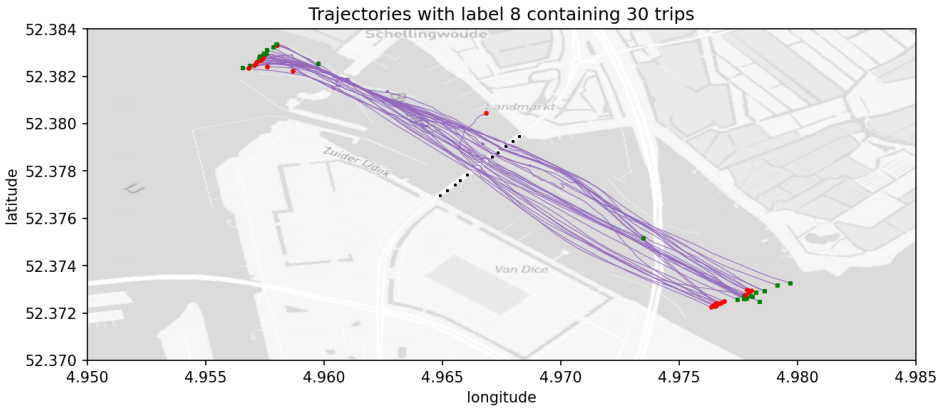


Figure 6.27: Trips in cluster 8, only recreational, the IJ lock complex included

With the main clustering explored based on the trajectory plots per cluster, the individual features will be inspected.

Individual feature exploration

Since the interaction with the bridge is the main objective in this section, the first feature explored is the minimal distance in meters with respect to the pillars, as depicted in Figure 6.29.

A completely different pattern is found compared to the minimal distances found at the Moerdijkbrug. In this case, many trips appear to have a close encounter with the bridge, resulting in a widely spread distribution of distances to the bridge across the clusters. Cluster 1 stands out, containing trips with most distances ranging from almost zero to 20 meters, with a couple exceptions. Reviewing the trajectory paths, this corresponds to many vessels utilising the movable part of the bridge, which spans 18 meters, resulting in a minimal distance of a maximum of 9 meters. Similarly, clusters predominantly using the wider span can be identified.

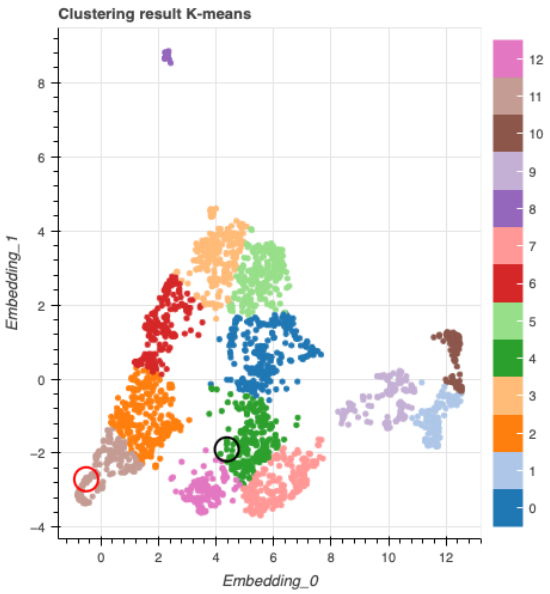


Figure 6.28: Scatter plot K-means clustering trips, trip 390 indicated by black circle, trip 2218 highlighted by red circle, the IJ lock complex included

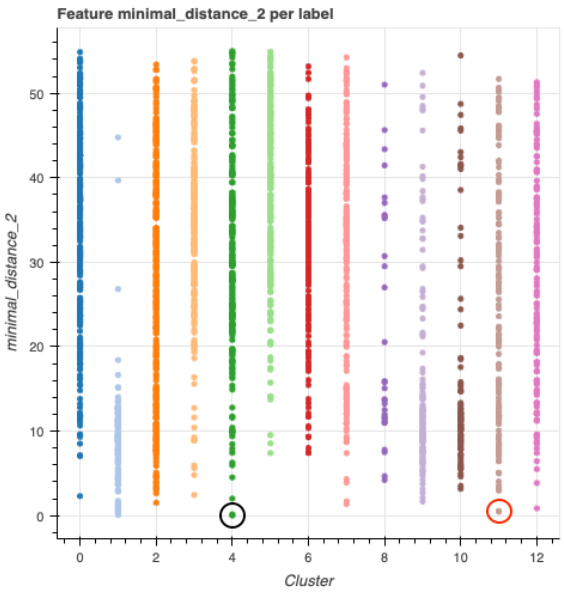


Figure 6.29: Jitter plot per cluster minimal distance [m] to Schellingwouderbrug, trip 2218 highlighted by red circle, the IJ lock complex included

Subsequently, all points with a minimal distance of less than two meters are examined by plotting their individual trajectory. A total of 20 trajectories meet this threshold value, with eleven of them belonging to cluster 1. The remaining nine are distributed among six other clusters. Closer examination of the individual trips in cluster 1, a recurring observation is noted: in almost all instances, the vessel type is unknown, and when dimensions are known, it typically is a small vessel.

This trend is also evident when examining the length-to-beam (L/B) ratio across the clusters, illustrated in Figure F.16. Numerous ratios are marked as zero due to the unavailability of vessel dimensions. Cluster 1 consistently displays a low L/B ratio, which, combined with the route via the movable part of the Schellingwouderbrug, suggests that this group may largely consist of recreational sailing vessels. This waterway section is part of the 'Staande Mastroute', a continuous shipping through the Netherlands, suitable for sailing and motorboats with a mast height or superstructure exceeding 6 meters. Moreover, this matches with the idea that should be a higher presence of recreational vessels in this area. Upon further examination of the L/B ratios, clusters 9 and 10 stand out as well, with a notable division in higher and lower values for the L/B ratios. This could indicate the presence of recreational vessels. Referring to the vessel types in the clusters in Figure F.17, it is evident that many unknown vessel types are found in these clusters. However the distinction is less clear compared to cluster 1.

Analysing the patterns in cluster 1 reveals that the small distances to the bridge are often attributed to a lack of data points near the bridge, leading to an apparently straight trajectory across a bridge pillar. For example, in Figure F.18 an encounter of 0.9 meters is observed. Another instance of a close encounter by a vessel is displayed in Figure F.19, featuring trip 1344 from cluster 2, with a minimal distance of 1.5 meters. In the trajectory plot, a small bump under the bridge is noticeable, resulting in the close encounter of this inland vessel.

In addition to these trajectories, two more are identified that exhibit a small minimal distance but do not show a concentration of data points close to the bridge or a single anomaly in the trajectory. These are trips 390 and trip 2218, which belong to clusters 4 and 11, respectively. Analysing their positions within the clusters, they are somewhat located on the edges, deviating from the typical patterns within the cluster. Both clusters display a higher frequency of irregular patterns near the bridge, as observed during the visual analysis. For trip 390, a minimal distance of 0 meters is found, and its trajectory plot is given in Figure 6.30 with a zoomed view to the bridge in Figure 6.31. A substantial number of data points are clustered just in front of the bridge. Examining the vessel type, it is identified as an inland

vessel measuring 110 meters in length and 12 meters in width, making this an even more interesting path

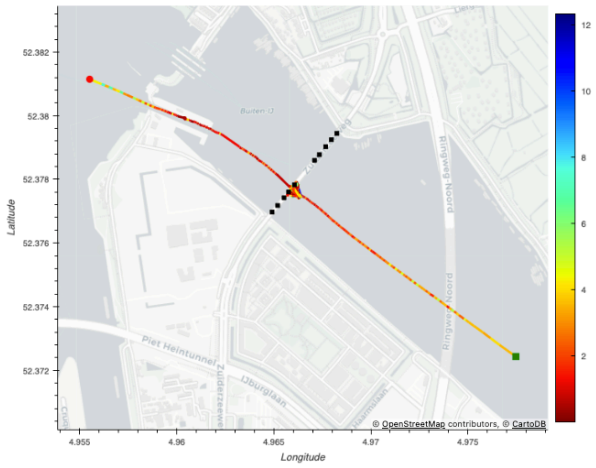


Figure 6.30: Trajectory plot of trip 390, 0 [m] from Schellingwouderbrug, colour bar indicating vessel speed [m/s], the IJ lock complex included

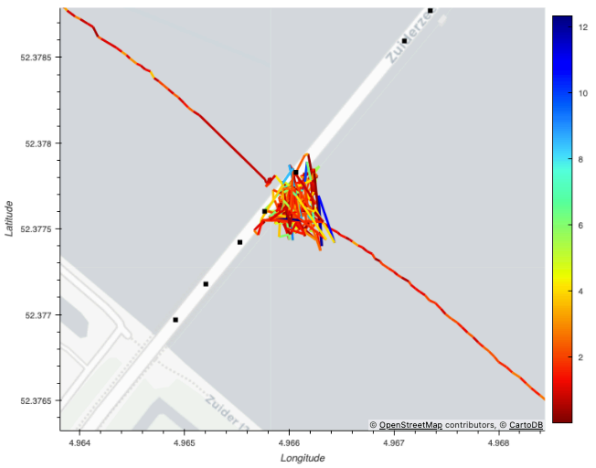


Figure 6.31: Zoom trajectory plot of trip 390, 0 [m] from Schellingwouderbrug, colour bar indicating vessel speed [m/s], the IJ lock complex included

Trip 2218 similarly has a very small minimal distance of 0.5 meters, as depicted in Figure F.20 and Figure F.21. Once again, this is an inland vessel measuring 110 meters in length and 12 meters in width and many data points are positioned precisely under the bridge, resulting in the minimal distance. Inspection of this bridge span, a guide structure is revealed, which may have been involved in a collision in both cases. However, nothing is registered on the moments corresponding to the AIS logs' timestamps in the public SOS-database, as found in Table C.2 in Appendix C.

Next, the trajectories close to trip 390 are analysed to determine if similar behaviour is found in neighbouring trajectories. In Figure F.22, seven trips in close proximity are identified, and the trajectory plots are visualised in Figure F.23. Among these plots, distinct patterns emerge, such as zigzag-like paths observed in trips 79, 81, 198 and 1510.

Furthermore, all trajectories exhibit minor route deviations near the bridge, except for trip 2086, which appears to follow a smoother trajectory. Notably, trip 79 is particularly distinctive, classified as a ferry based on the VesseltypeERI value in the data set. However, there is no corresponding ferry line according to the GVB Figure D.2, and a marina is located in the eastern part where a stop is made. This deviates from a typical ferry route within a 40-minute time frame.

When examining the maximum acceleration per trip within the 400-meter range around the Schellingwouderbrug in Figure 6.32, a complete different pattern is visible compared to the Moerdijk bridge case. Here, numerous trips exhibit considerably high maximum acceleration rates. Except for the clusters 1, 9 and 10, which all lie on the right side in the embedding in Figure 6.20. The distributions per cluster are visualised in Figure 6.33, showing large median values over the rest of the clusters, lying on the left side of the embedding.

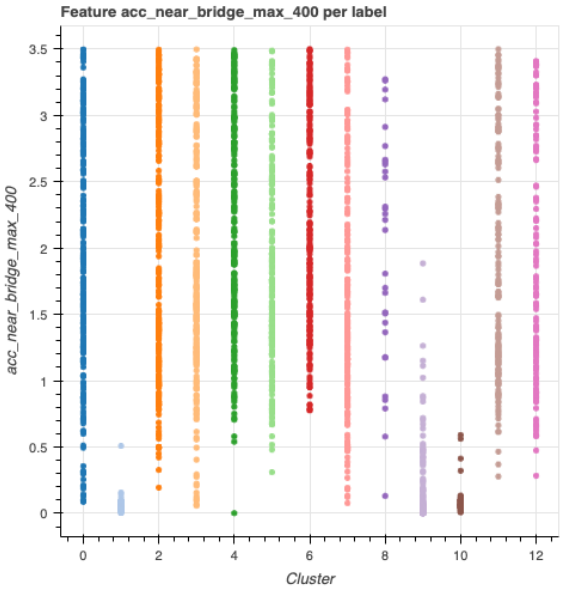


Figure 6.32: Jitter plot per cluster max acceleration [m/s^2] 400 meter around the Schellingwouderbrug, the IJ lock complex included

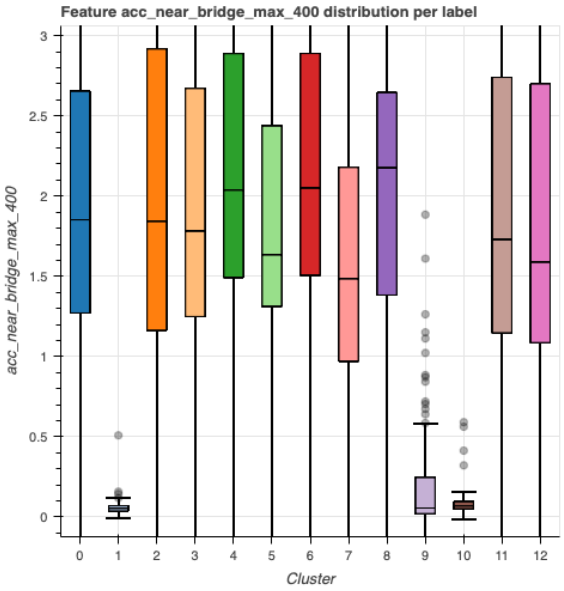


Figure 6.33: Box plot per cluster max acceleration [m/s^2] 400 meter around the Schellingwouderbrug, the IJ lock complex included

6.3. The IJ lock complex excluded in area

To investigate the potential influence of the Oranjesluizen lock complex in the north on clustering results, the trajectory clipping in this section excludes the specified area, as shown in Figure 6.34. This exclusion is applied to the same data set as before, resulting in 2222 trips, 26 less than when the area was included. As a result of the clipping, some trajectories might have become too small and were consequently filtered out. Further details on trip lengths and vessel types are provided in Subsection F.3.1.

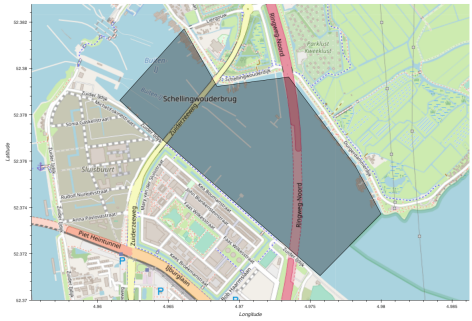


Figure 6.34: Area near Schellingwouderbrug where trajectories have been clipped, lock complex in the north excluded

The developed method is applied in the same way as for the trajectories with the lock complex, now the number of clusters is set to 8 based on the elbow plot in Figure 6.35. A significant difference is observed when comparing the scatter plot in Figure 6.36 to the plot in Figure 6.20. In the situation with the lock complex included, only one distinct cluster, the recreational cluster, is identified, with the other clusters much closer to each other. However, in the case with the lock complex excluded, the clusters are more widely spread and denser. A comparison of clustering scoring metrics in Table 6.3 indicates better scores across all indexes for this particular clustering.

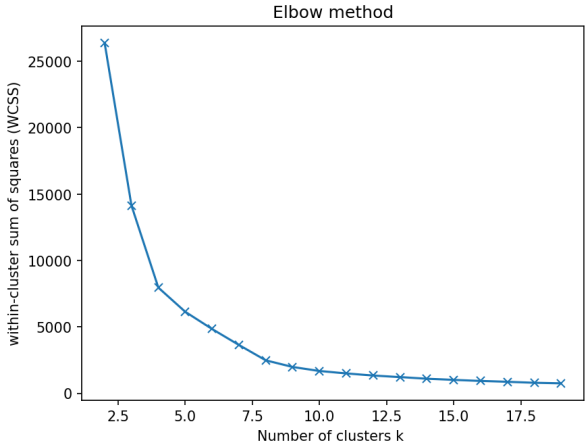


Figure 6.35: Elbow plot clustering to determine number of clusters 'k', the IJ lock complex excluded

Table 6.3: Clustering performance indicators with scores for 8 clusters, the IJ lock complex excluded

Method	Score
silhouette Coefficient	0.57
Calinski-Harabasz Index	9737
Davies-Bouldin Index	0.52

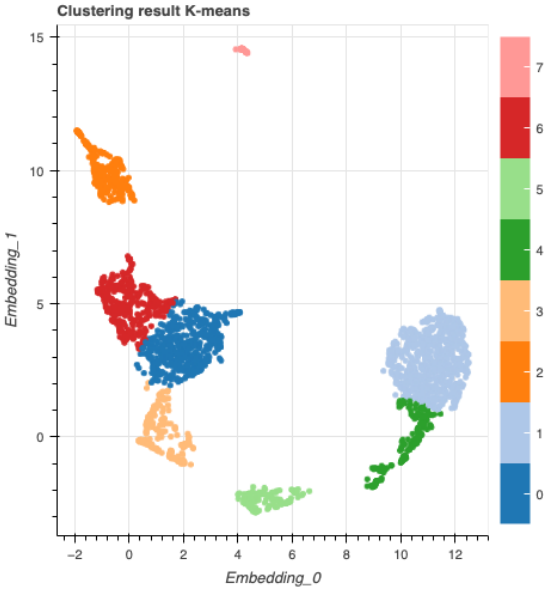


Figure 6.36: Scatter plot K-means clustering trips at Schellingwouderbrug, colour coded per cluster, the IJ lock complex excluded

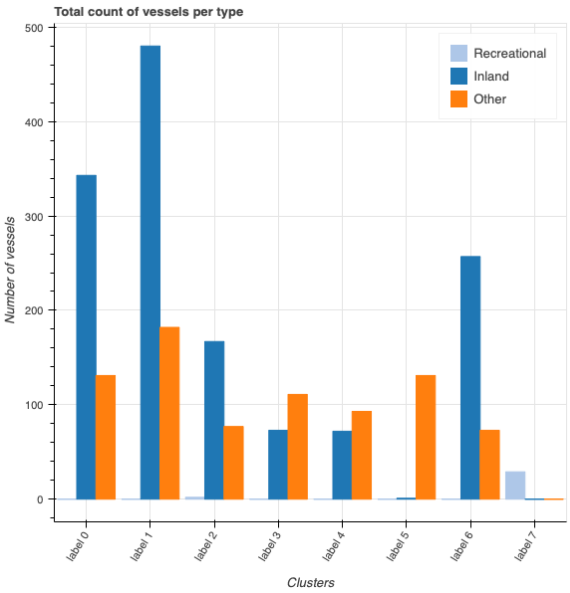


Figure 6.37: Distribution of vessel types across the clusters, the IJ lock complex excluded

Examining the distribution of vessel types over the clusters in Figure 6.37, two clusters where only one vessel type is included attract attention. The cluster with label 5 only contains vessels classified as 'other types,' while the cluster with label 7 includes solely recreational vessels. Their location in the embedding, as individually located clusters, stands out as well. These two clusters are individually located and distinct from the others. The remaining clusters exhibit a mix of other types and inland vessels. Additionally, the sizes of the clusters differ significantly, with clusters labelled 0 and 1 containing the most trips, 474 and 662, respectively.

Visual exploration of trajectories per cluster

To gain deeper insight into the meaning of all clusters, trajectory paths per cluster are plotted in Figure 6.38. Notable differences are observed in terms of the primary directions, bridge openings navigated, and overall trajectory patterns.

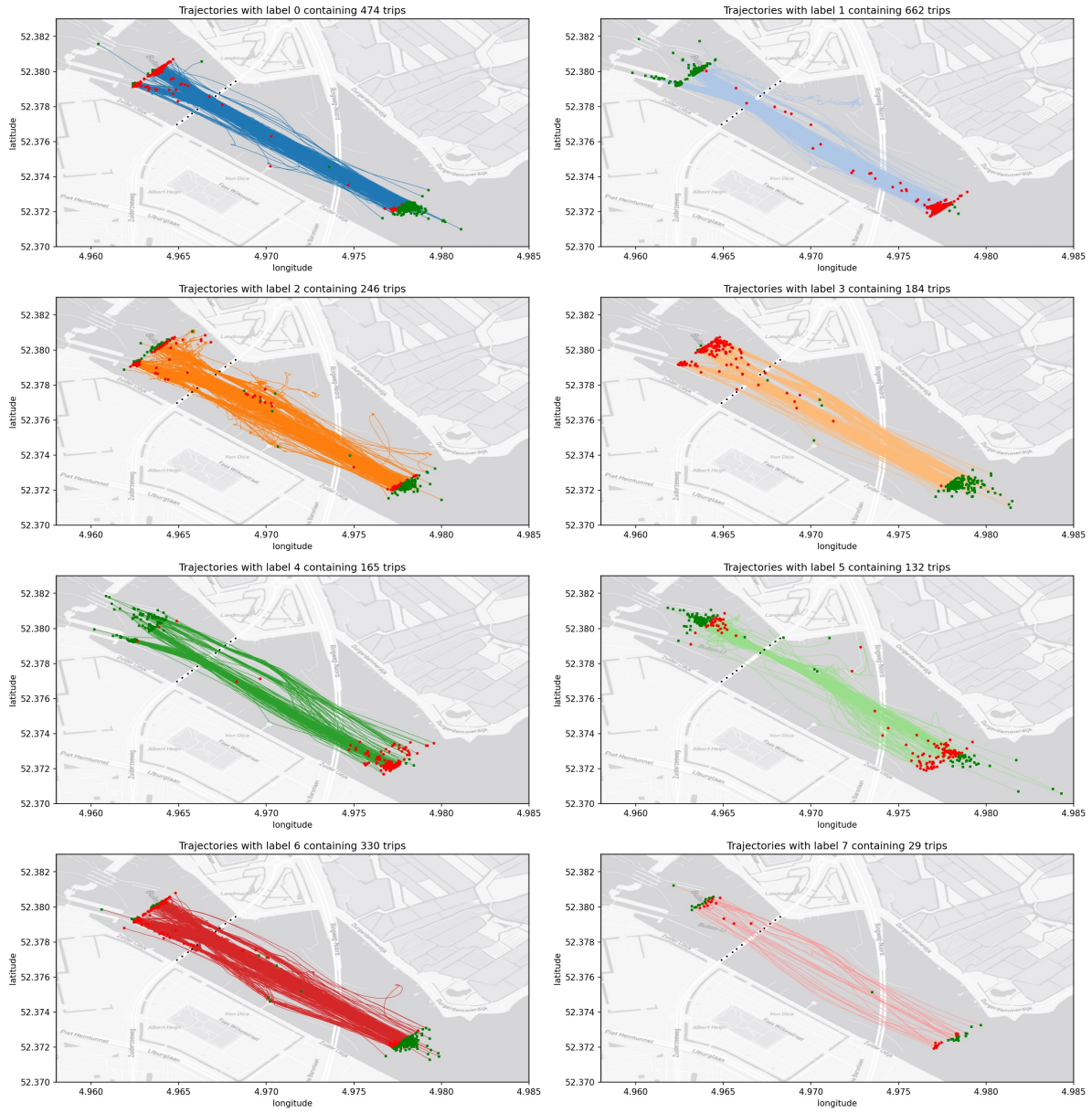


Figure 6.38: Trajectory plots per cluster, the IJ lock complex excluded

The north to south is prominent in clusters labelled 1 and 4, located on the right side in the embedding, coloured blue in Figure 6.39. Conversely, the south-to-north direction is found in for clusters labelled 0, 3 and 6, positioned on the left side of the embedding, and coloured orange. Clusters 5 and 7 exhibit a more mixed pattern, occupying the middle of the embedding. Cluster 2, located on the left side of the embedding, primarily follows a south-to-north direction, here identified as mixed an coloured purple.

Regarding the main bridge opening used across the clusters, they can be categorised into the large fixed span, the smaller movable bridge part, and a mixed pattern. Clusters 0 and 1 predominantly utilise the large span as their main route, while in cluster 5, the majority navigates through the smaller movable part. This cluster aligns with a previously identified group of sailing vessels based on their route and L/B ratios. In this case, the cluster is distinct from the others. Analysing the L/B ratio in Figure F.27, it confirms that this group is consistent with the one found earlier. For the remaining clusters, a mixed pattern is observed. Translating this finding, the small part is mainly used in the lower section, the large part in the middle, and the other sections exhibit a mixed pattern, the main bridge opening used is indicated in Figure 6.40

Furthermore, the embedding reveals a distinction between smooth trajectories and more chaotic patterns. Chaotic patterns are prevalent in clusters 2 and 6, while smoother trajectories are observed in 0, 1, 3, 4 and 7. Cluster 5 is identified as a mixture of smooth and chaotic patterns. The chaotic clusters are situated on the left side of the embedding, Figure 6.41 illustrates the main patterns found.

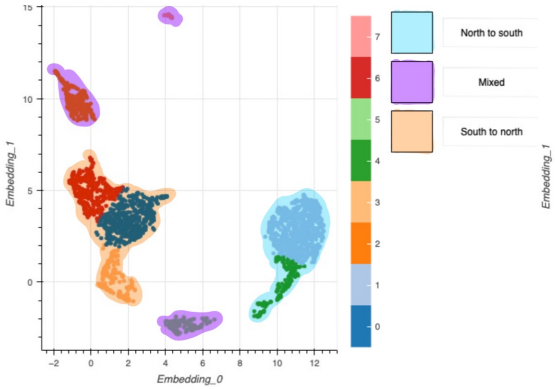


Figure 6.39: Identification of main vessel direction over the clusters, the IJ lock complex excluded

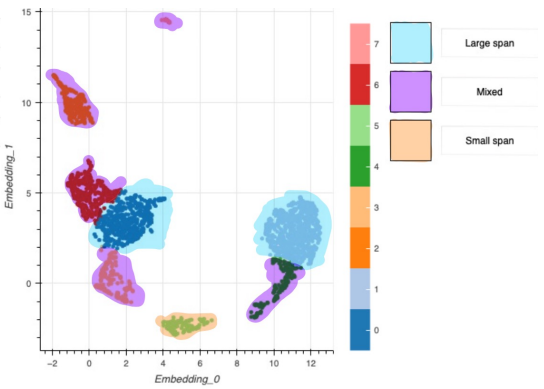


Figure 6.40: Identification of main bridge opening used over the clusters, the IJ lock complex excluded

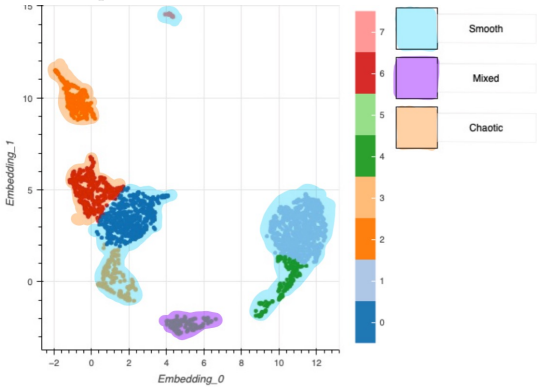


Figure 6.41: Identification of main path style over the clusters, the IJ lock complex excluded

In this case, a small cluster positioned in the top of the embedding once again comprises solely recreational vessels, the trajectories are visualised in Figure 6.42. The trips are identical to the case with the lock complex included, illustrated in Figure 6.27. Where in this case the cluster includes 29 trips instead of 30, probably due to clipping. This highlights the significant influence of the vessel type feature.

Figure 6.42: Trips in cluster 7, only recreational, the IJ lock complex excluded

Individual feature exploration

In the case with the lock complex included, two trips were identified with a really close encounter with the bridge pillar and numerous points before the bridge, indicating a collision. These trips were detected based on the minimal distance feature, which is visualised in Figure 6.44. In the figure, these two trips, labelled as 387 and 2191, are highlighted with black and red circles, respectively. Both trips belong to the clusters previously identified as having many deviating patterns. When examining their individual locations within the clusters, they are found at the edges of their clusters, indicating differences from the other trips within the same cluster. The trajectories corresponding to trip 387 and 2219 are the same as found in the case with the lock complex included.

Once more, one cluster stands out, cluster 5. This cluster notably contain a significant number of vessel paths in close proximity to the bridge, consistently aligning with the vessels using the movable part of the bridge.

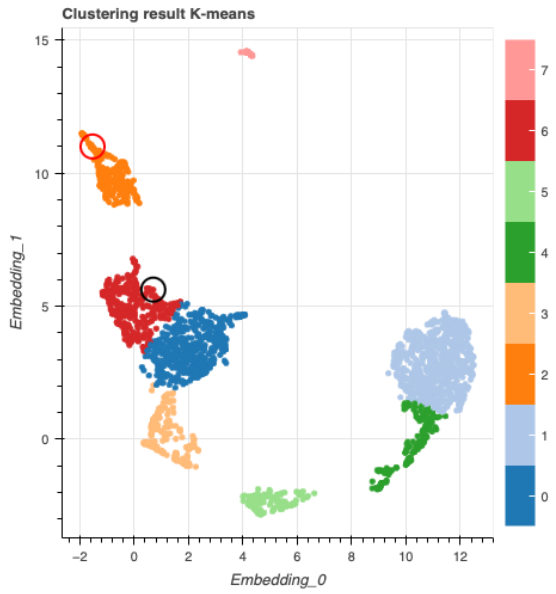


Figure 6.43: Scatter plot K-means clustering, trip 387 indicated by black circle, trip 2192 highlighted by red circle, the IJ lock complex excluded

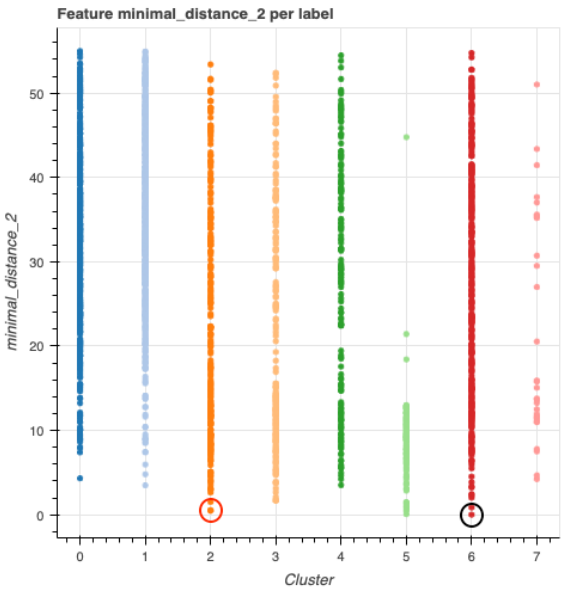


Figure 6.44: Jitter plot per cluster minimal distance [m] to Schellingwouderbrug,trip 387 indicated by black circle, trip 2192 highlighted by red circle, the IJ lock complex excluded

In Figure 6.45, the plot depicts the maximum acceleration within a 400-meter range around the Schelling-

wouderbrug per trajectory for all clusters. Cluster 5 exhibits notably low acceleration values, positioned at the lower end of the embedding seen in Figure 6.36. Additionally, clusters 3 and 4 display lower acceleration compared to the other clusters, a distinction that becomes more evident in the distributions shown in Figure 6.46. Notably, clusters 3 and 4 are also positioned on the lower side of the embedding.

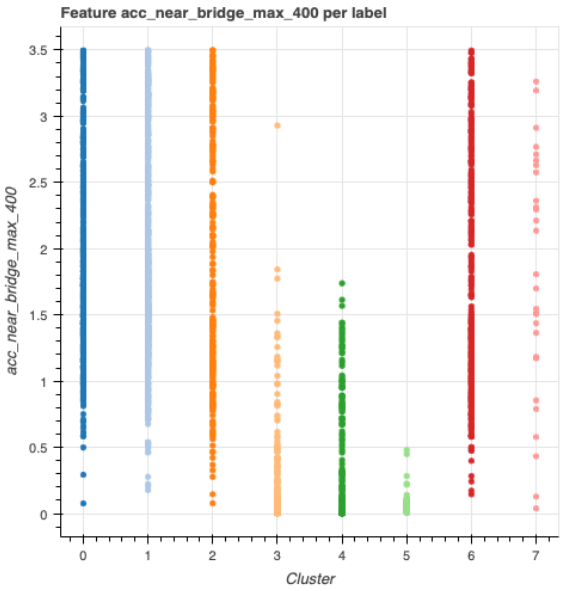


Figure 6.45: Jitter plot per cluster max acceleration [m/s^2] 400 meter around the Schellingwouderbrug, the IJ lock complex excluded

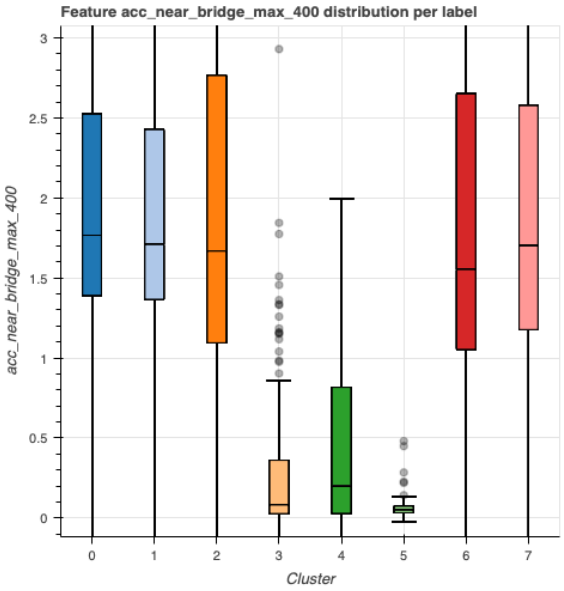


Figure 6.46: Box plot per cluster max acceleration [m/s^2] 400 meter around the Schellingwouderbrug, the IJ lock complex excluded

Part III

Discussion, Conclusion and Recommendations

7
Discussion

In this section, a discussion of the approach and results of this study is undertaken. Comparisons with relevant literature are drawn, and suggestions for improvement are provided. A distinction is made in reflecting on the general method of safety assessment, addressing challenges encountered in the data, and a view on the proposed method in this research.

7.1. Reflection on the general method of safety assessment

This study primarily focuses on the current method of assessing nautical safety on the Dutch inland waterways. In contrast, safety assessments are conducted across various domains, such as industry, healthcare, and different transportation modes. Studies in these domains learn from each other. For example, Marx and Slonim (2003), highlights how patient safety risk is evaluated in healthcare before the injury occurs, with an approach from the aviation and nuclear industry. The tool used, offers a away to find, prioritise, and reduce patient safety risk proactively. Various factors like equipment failures, human errors, risky behaviours and opportunities to recover are combined in fault trees, this approach places a distinct focus on the left side of the bow-tie model.

In the context of nautical safety assessment, recent research by Huang et al. (2023) reviewed maritime risk literature from 2002 to 2021. It analysed risk assessment methods, giving insight into the most used methods, highlighting a shift towards systematic approaches and the integration of artificial intelligence. However, many assessments, including the current method for assessing risks on Dutch inland waterways, rely on a retrospective approach. Under this approach, areas with a history of accidents receive high-risk scores, suggesting a closer examination of this situation or location. While improved accident reporting can offer a more complete view of nautical safety, it remains a backward-looking approach. Reports are often generated years after the events, offering hindsight and insights, yet not directly contributing to immediate navigational safety enhancements.

In contrast, Leveson (2015) illustrates how leading indicators in the aviation sector are employed to identify potential accidents before they occur. However, aviation and maritime contexts differ in terms of trajectory and behaviour. Vessels move slowly and in two dimensions, whereas planes can navigate in 3 dimensions and with much higher speeds (Wang, Yang, et al., 2021). Similar approaches to safety assessment can provide valuable insights. By leveraging real-time data, decisions can be made more directly, and the negative effect of lagging indicators can reduce.

This suggests the need to reevaluate the existing safety assessment method, especially regarding identifying near misses. One potential improvement involves the establishment of a threshold for the maximum number of incidents within a specific time frame. If this threshold is reached at any location, an immediate notification would trigger active actions by waterway managers. Under the current approach, the safety study report is awaited and measures are implemented years later. With this idea, further enhancement of the method proposed in this study, with a focus on the identification of near misses, should be made, primarily through an improved definition on near misses that will be addressed in the recommendations.

7.2. Challenges in the data

The outcome of this study is fundamentally influenced by the quality of AIS data, which is not always guaranteed (Iphar et al., 2020). To address this issue, several preprocessing steps were taken to clean

the data by removing outliers, the following parts will give some examples of issues encountered and potential measures to overcome this.

In the results, several trajectories exhibit deviant paths that appear incomplete or unrealistic. An example is provided in Figure 7.1 illustrating a vessel's path from the green square towards the red dot. According to the methodology, this was identified as an anomalous pattern due to the small distance to the bridge pillars. In reality, the vessel did not come close to the bridge pillar as suggested by the trajectory. It appears that the applied filter failed to identify this outlier.

Figure 7.1: Trajectory plot with unrealistic data point, colour bar indicating vessel speed [m/s], Hollands Diep

An unrealistic data point is visible, geographically distant from the vessel's track and an extremely high speed is found compared to the rest of the track. Such unrealistic data points, known as 'jumping', are not uncommon, as noted by Emmens et al. (2021), who discuss opportunities and limitations of AIS data based on literature. The example in Figure 7.1 may be attributed to GPS signal blockages, often caused by physical obstacles such as bridges. Which becomes especially tricky when accidents involving a bridge are investigated.

In addition to unrealistic data appoints, the absence of data points can lead to trajectories that may appear unrealistic. This situation was observed examining the minimal distance towards the Moerdijkbrug, where a trajectory with limited data points can result in paths appearing close to bridge pillars, which may not be accurately represent the vessels actual route. This discrepancy can occur due to points being filtered out or the AIS signal frequency being insufficient.

Additionally, vessel trajectories exhibiting a distinctive zigzag pattern are segregated into a separate cluster, distinctly different from the remaining trajectories. This appears consistent across multiple trips of the same vessel and is likely related to the onboard AIS system, as all trajectories of this vessel display the same zigzag pattern. Although this pattern deviates from the norm, these can not be indicated as incidents or near misses.

To address these data issues, various methods have been proposed to enhance the data quality. For instance, the work of Chen et al. (2020) involves the reconstruction of ship trajectories from AIS data through data quality control and prediction. Additionally, the field of robotics, as discussed by Ravankar et al. (2018), focuses on path data smoothening current methods and future challenges.

However, caution is necessary when extensively smoothing paths, particularly in the context of anomalous behaviour detection. The excessive smoothing may lead to eliminate of minor deviations in trajectories that serve as indicators of potentially dangerous situations in the proposed methodology of this study. Striking a balance in data improvement is essential, where path smoothening in a lot of instances focuses on eliminating lumps of points close to each other, to create a smooth path, in this case two accidents are identified as such partly based on this cloud of data points in front of the bridge,

which should remain in the path. It is suggested that irregularities in the paths should be retained to function as indicators because removing them at the front could result in fewer accidents or near misses being detected. Moreover, any anomalies that are filtered out at the front will not be found and it is better to evaluate trajectories afterwards.

Furthermore, it is worth noting that the limited number of accidents or near misses on waterways poses a substantial challenge in detecting these events. This means extensive data sets should be used. Although parallel computing with Dask by Dask Development Team (2016) has been applied to handle these large data sets, it is important to mention that certain challenges remain unsolved in executing the computation in parallel within this study. Parallel computing has been successfully applied to a portion of the data set, and suggestions for improvement are given in the recommendations. Moreover, the significance of extensive data becomes even more important when not only looked at accidents or near misses, but shift more toward the left in the bow-tie model by addressing abnormal or even normal behaviour.

However, significantly increasing the volume of data can impact dimension reduction and final clustering. While UMAP is recognised for handling large data sets, but this research does not cover the effects of increasing the data volume in this particular application.

7.3. View on proposed method

In Chapter 3, vessel behaviour is defined based on expected patterns of behaviour on waterways, categorised into general behaviour, behaviour during interaction with infrastructure or objects, and behaviour during interactions between vessels. Anomalies are defined as patterns or activities that deviate from this normal behaviour.

This method allows for the identification of clusters of vessels that exhibit similar behaviour on a broader scale. As a result, groups of vessels displaying deviating behaviour can be assigned for further in-depth analysis. The subsequent step involves transitioning from these anomalies to near misses since not all anomalies indicate near misses.

In the context of the bow-tie method, this research begins with behaviour which lies on the left side of the model, as it relates to cause associated to behaviour. Where the approach form the other side, starts with the actual accident. In the literature there is no consistent approach to the bow-tie method and both directions are used across safety studies (de Ruijter & Guldenmund, 2015). Starting from the behaviour assumptions are made regarding significant characteristics of vessel behaviour in accidents or collisions. This differs from starting from a real accident to define features, which requires well registered and documented accidents, and is not available in this case.

On the other hand, a study starting from an accident to identify behavioural actions is executed by Mestl et al. (2016) who suggest that the ROT just before a collision between two vessels is a good indicator. Applying the ROT as accident indicator in this study, it was not very effective in studying the interactions between a vessel and a bridge, as many high ROT values are found in the trajectories at the Hollands Diep and the IJ. This could be due to an increased need for steering in confined waterways than on open sea where these results were obtained.

Furthermore, in comparison to other studies, this research used a broader range of features. Many classification studies based on AIS data use static features like vessel dimensions, for example Wang, Yang, et al. (2021) employ random forest algorithms to classify ships according to the static information from AIS messages, and recommend to make more use of the dynamic part of the AIS data, which is incorporated in this study. The combination of both static and dynamic features appears to be effective. For instance, when analysing routes via the movable part of the Schellingwouderbrug in conjunction with a smaller length-to-breadth (L/B) ratio, it becomes possible to identify a group of recreational sailing boats.

In comparison to the approach taken by Zhou et al. (2019), this study uses a broader range of features

to define vessel behaviour. While Zhou et al. (2019) primarily relies on the ship position and SOG as behaviour features to cluster vessels and identify their type. Moreover, the work from Zhou et al. (2019) focuses on a straight section of a waterway at Hoek van Holland with fixed x-y coordinates to determine vessel position. In contrast, this study covers two distinct areas with more complex geometries, and in both instances, neat clustering results are found based on the scoring metrics.

However, an interesting finding is the influence of the waterway geometry on the clustering results. While it was initially assumed that infrastructure on or around the waterway would affect vessel behaviour, the analysis of the IJ section close to the Schellingwouderbrug reveals a significant difference in clustering results for the area with and without the lock complex included, visualised in Figure 6.20 and Figure 6.36, where much better separated clusters are found and for the part without he lock as well as the scoring metrics. In this case, the interaction with the bridge was the focal point for investigation. Apparently, the inclusion of the locks had a substantial impact on all features, resulting in a completely different clustering outcome. The domain used was found to have a significant influence on the results, and should be defined with the specific goal of investigation in mind.

Moreover, the impact of clipping trajectories near an object of interest. It appears that creating complete trajectory paths followed by clipping in a specific area yields better results than forming trajectory paths from a filtered set of data points from the complete data set, which resulted in incomplete trajectory paths. This makes it more challenging to compare individual paths and define concise features of these trips.

In the work by Yan et al. (2022) a similar division of vessel types as input features for a model is used as in this study. Defining five different types to construct geometry features, where only two main vessel types are included in this thesis, together with the L/B ratio feature to include vessel dimensions. Regarding behaviour features, the study by Yan et al. (2022) included speed and voyage duration. Where the speed features are divided into mean and standard deviation for high and low, in this report statistics are used in defining the features

The input data in the work by Yan et al. (2022) is satellite AIS data covering the global ocean, making the voyage duration likely a more interesting feature than inland shipping, where distances are typically smaller, but the incorporation can be interesting. The spreading in longitude and latitude is used as a feature next to the total voyage distance. In this study, no use is made of the total duration of a vessel trip or distance covered during a trip, as this was considered less important on the relatively short paths on inland waters than in the open ocean. Nevertheless, these features could positively impact the clustering of vessels, such as tugboats, which typically remain close to a port or recreational vessel which typically follow a distinct path to inland vessels.

Notably, an investigation into the trip length revealed that the number of data points along a trajectory appeared to have an impact, even though it is not directly incorporated as a feature. Trips with a larger number of data points were found to be correlated with the number of peaks in acceleration, and in the end exhibited similar patterns, often remaining stationary in front of the bridge for a certain duration. Further investigation is necessary to thoroughly examine the effects of individual features and to conduct a detailed analysis of each feature's impact. This will help reveal potentially deeper insights into the defined features and their underlying effect.

In this study, K-means is used, which is known to be a reliable clustering method. Nevertheless, there are other options available. However, determining the optimal number of clusters, denoted as 'k', remains a challenge. An inappropriately chosen initial number can result in a poor clustering outcome. The elbow method is used as the primary method to define the number of clusters. This method relies on visual assessment, and opinions in the literature vary regarding its effectiveness. In this case, scoring metrics are employed and show reasonable values for the clustering, which imply the elbow method works for this application.

Additionally, the kneed algorithm, mentioned in Subsection 3.2.3, is used once in this study to validate the visually chosen number of clusters. Implementing this algorithm universally could enhance confi-

dence in determining the optimal elbow point. In comparison to the calculated elbow point, the number of clusters chosen lies a bit further on the curve. Fining an optimal number of cluster remains challenging, for instance, while the silhouette coefficient is maximal for only three clusters in the Hollands Diep case in Section 6.1 the chosen number is 11 based on the elbow method, visualised in Figure 6.3. Choosing a smaller number, such as three clusters, results in a higher volume of trips within each cluster, making it less probable to identify clusters with similar behaviour or detect anomalous patterns. Thus, opting for a larger number is a deliberative choice, as the ideal number of clusters remains debatable.

8
Conclusion

This chapter summarises the results of this research, taking into account the limitations discussed in Chapter 7. The research questions defined in Chapter 1 are examined in terms of how they have been answered in this study.

To arrive at an answer to the main research question, four sub-questions were initially defined, each of which will be addressed separately:

Sub-question 1

> "What defines the current state of nautical safety on Dutch inland waterways, how is this measured, and what aspects could be enhanced for improved safety assessment?"

The state of nautical safety is assessed by the Monitor Nautische Veiligheid, which categorises various accident types, including ship-ships, ship-object and ship-infrastructure. The inputs for the Monitor Nautische Veiligheid, include recorded accidents and incidents in the SOS-database and expert opinions. The SOS-database, managed by Rijkswaterstaat, serves as the national shipping accident database, containing information on shipping accidents and other water-related incidents that have occurred within the Netherlands. Data for this database is provided through reports filed by the involved skipper or the waterway manager responsible for that section of the waterway.

From 2009 to 2022, an annual average of 1110 accidents were registered, with 14% categorised as significant accidents, meaning those with severe consequences (Hofmeijer, 2019). However, the data exhibits notable shortcomings, primarily arising from the low registration rate, which was estimated at just 35% in 2013 (Movares, 2013a). This estimation is based on observed damages on structures which do not correspond to recorded accidents. Additionally, data quality issues further complicate the problem. The manual data entry process is prone to mistakes and incompleteness. Some reported accident locations do not match the actual physical details, such as groundings documented as occurring meters away from riverbanks.

The current approach to safety assessment relies on historical yet incomplete data. In context of the bow-tie model on the right side, with a focus on the consequences. High-risk scores are assigned to regions with a history of accidents, suggesting a closer examination of this situation or location. While enhanced reporting of accidents can result in a more complete view of nautical safety, the approach remains retrospective. Reports are typically generated years after the incidents, providing hindsight and insights but not directly leading to enhanced navigational safety.

In conclusions, a transition toward a more proactive approach is recommended. While the main focus primarily addresses accidents on the right side of a bow-tie model, there is significant value in shifting to the actual prevention of the causes on the left side of the bow-tie. By defining normal behaviour, addressing deviations from this norm, including near misses and accidents, the shift from the lagging indicators towards more leading indicators can be facilitated.

Sub-question 2

"What is the definition of "normal" and "unusual" vessel behaviour, and how can near misses be defined in this context?"

Normal shipping behaviour is defined as the expected vessel behaviour on inland waterways, while anything that deviates from this is considered unusual behaviour. Definitions have been established for the three primary accident types recognised by Rijkswaterstaat: ship-ship, ship-object, and ship-infrastructure interactions.

These definitions are formulated based on various characteristics, including speed, acceleration, manoeuvrability and distances. Normal behaviour is characterised by minor deviations from intended courses, low acceleration rates and sufficient distances between vessels or between a vessel and an object. Conversely, unusual behaviour is defined as everything that deviates from normal. It is indicated by significant course changes, major changes in acceleration and speed, and close encounters with vessels, objects or infrastructure.

Within the specific context of ship-infrastructure interactions, vessels are expected to be positioned correctly at a considerable distance from the bridge and to follow a smooth trajectory while passing it. Therefore, any significant acceleration, abrupt deceleration, or drastic course changes close to a bridge should be considered deviations from normal behaviour and categorised as anomalous.

This approach helps identify vessel trajectories that deviate from the expected norms, categorising them as potentially dangerous. Within this category a distinction can be made between false alarms, real accidents, and near misses. Near misses are defined as just not accidents. For example, in the context of bridge interactions, near misses may be assigned when the minimal distance with respect to the bridge pillar is only slightly greater than the distance between the AIS transmitter and the vessel's outermost boundary.

Sub-question 3

"How can the transition from AIS data to vessel behaviour be made to identify anomalous behaviour, including near misses?"

The derivation of vessel behaviour from AIS data involves the analysis of available data to associate it with vessel behaviour characteristics such as speed, acceleration, and manoeuvres. In order to work with the AIS data, several data preprocessing steps are needed, including outlier detection and trip collection generation.

Vessel behaviour is represented through a set of features extracted from AIS logs and additional features computed during preprocessing or by supplementary software, tsfresh. These features contain all trip details and are processed through a dimension reduction technique, particularly UMAP followed by K-means clustering. This approach retains the underlying behaviour, preserving similarities in the original data.

Visual exploration of the trips per cluster reveals similarities among trajectories, including main directions and patterns. In-depth examination of these features provides insight into vessel behaviour, from interactions with bridges indicated by minimal distance features to speed or acceleration distributions along a vessel's trajectory. These analyses allow the identification of deviations from normal behaviour.

Numerous atypical patterns are observed, but actually designating one as near miss is complicated due to the many uncertainties arising from the vessel paths. Therefore, data of higher quality or the introduction of a more precise definition of a near miss is required. This could enhance the accuracy of identifying a pattern as near miss in these situations. In the context of ship-infrastructure, a suggestion on the definition is made. The definition should be based on a variable distance boundary around a vessel, dependent on the vessel type, size, and the specific circumstances of each case.

Sub-question 4

> "What is the added value of the proposed method for detecting anomalous vessel behaviour compared to the existing method, and how does it enhance insight into nautical safety?"

The proposed method is effective in generating insight in vessel behaviour per cluster and enables detecting anomalous vessel behaviour. When applied to the section near the Schellingwouderbrug, it successfully identifies two trips that could be classified as incidents. In these cases, the vessel trajectories pass the bridge, but a concentration of AIS data points is observed just in front of the bridge, resulting in a small minimal distance with respect to the bridge. The combination of numerous data points indicates that the vessel is stationary, and the close encounter suggests a potential accident. Comparing the locations and dates of these incidents with the data in the public SOS-database reveals that they do not correspond to any registered accidents. This means the SOS-database can be supplemented with these incidents found.

In this specific case, out of the 2248 trips generated, 20 were flagged as suspicious. Investigating these trips, two accidents as described above, are identified within this subset. This means that less than 1% of the initial data set was explored to uncover these incidents. This demonstrates the method's ability to efficiently point out where to look for anomalous vessel behaviour and accidents, thereby the potential value in enhancing insight in nautical safety. Nevertheless, it is worth noting that there is a possibility that some events may be overlooked among the remaining 2228 trips that were not individually investigated.

Research question

By answering the four sub-questions, the main research question of this study can be answered:

> "How can insight into **nautical safety** on **Dutch inland waterways** be improved by the detection of **anomalous ship behaviour** based on **AIS data**?"

The method defined and applied in this research involves the extraction of vessel behaviour from AIS data using feature engineering, followed by dimension reduction with UMAP and K-means clustering. This process enables the identification of similar trajectory clusters and anomalies, providing valuable insights into ship-infrastructure interactions.

The method's application to the Moerdijkbrug at the Hollands Diep and Schellingwouderbrug at the IJ yielded clusters displaying similar vessel behaviour, primarily in direction and routes. Numerous atypical patterns are observed which have been investigated in more depth, in both cases less than 1% of the data set. In this small portion of the input trajectories, two distinct patterns identified and classified as most likely accidents in the IJ case.

The distinct patterns identified by the proposed method could supplement the existing SOS-database used for nautical safety assessment. Although it may only partially address the issue of under-registration, it holds potential for enhancing insights. However, it is crucial to acknowledge that the current retrospective method is limited by the fact that reports are often written years after the incidents occur. To transition to a operational approach, the proposed method could be implemented with real-time data, setting a predefined threshold for the number of abnormal tracks or recorded incidents. When this threshold is reached, it can trigger immediate alerts to waterway managers. This shift from consequences on the right side of the bow-tie diagram more towards the left side represents a step toward improved safety management.

9 Recommendations

This chapter reflects on how this research should be applied, makes suggestions for future research on this topic and will contain recommendations for the field.

9.1. Possible applications of the methodology

The current methodology applied in this research enables the derivation of vessel behaviour from AIS data, wherein clusters of similar behaviour are identified, and subsequently, clusters displaying more deviant patterns are found. Based on the minimal distance feature and trajectory paths analysis, two unreported accidents were found in the Schellingwouderbrug case. Application of the method could be used directly to supplement the SOS-database regarding ship-infrastructure accidents, where certain challenges in the volume of input data, will be addressed in the future work section.

To be able to identify near misses with this approach a well-defined concept of near misses should be established. Anomalous patterns can be detected and subsequently categorised as false alarms, near misses, or actual accidents based on the trajectory path. In the context of ship-infrastructure, a near miss definition should be established, taking into account vessel type and dimensions. The current research uses the location of the AIS transmitter, neglecting the vessel's dimensions around this point, which holds significant importance. Literature over viewed by Szlapczynski and Szlapczynska (2017) provides a range of safety boundaries used in the maritime context, suggesting various possibilities. The Master's thesis by Baak (2023), who recommends a ellipse safety domain around a vessel based on research of ship-ship interactions in port areas, could serve as valuable input for the near miss domain in this context. By defining a safety domain around the vessel, a threshold value for the minimal distance can be set, which should differ based on the investigated situations. For example, when examining a small movable bridge opening, a lower threshold value should be applied, since closer encounters are anticipated compared to large spans.

Furthermore, the method could be applied for ship-object interactions, given that it essentially follows the same approach as for ship-infrastructure which is applied in-depth here. A potentially interesting work field is the North Sea. Due to the growing demand for sustainable energy, numerous wind parks are constructed off the Dutch coast, resulting in limited space for vessels. Given that the North Sea is already heavily trafficked, combined with the presence of more wind parks and the desire to maintain or improve safety levels, the coming years pose a significant challenge (MOSWOZ, 2022). The methodology may offer valuable insights into vessel behaviour during ship-turbine interactions, which could lead to recommendations on vessel routes through the current and new to build parks.

In this research the focus lies on inland waters, clusters with deviating trajectories are identified. The method could also be applied to coastal waters, where piracy and illegal fishing activities can be traced. It is known that fishing vessels have typical patterns during their fishing activities and are usually small vessels. These typical patterns in combination with the vessel size can be identified with this method, while the method proved to identify groups of recreational sailing vessels based on the L/B ratio and location on the waterway. Noteworthy, the vessels must have an active AIS system on board.

An immediate practical application of this methodology could involve assessing the impact of modifications made on a waterway. Often, various measures are implemented on waterways to enhance safety or regulate traffic flow. Determining the actual effects of these measures can be challenging, especially

when comparing situations before and after their implementation based solely on registered accidents, which are infrequent. Therefore the vessel behaviour patterns based on the AIS data could bridge this gap. Two data sets are necessary, one before the modification and one after. The effect of the measures should be visible in general vessel behaviour patterns, and validation of the desired or expected outcome should be possible.

Another potential application of this methodology lies in researching the interaction between inland vessels and recreational vessels, a high-risk scenario among the top ten risks according to the Monitor Nautische Veiligheid. Investigating the differences in behaviour patterns of inland vessels with and without recreational vessels present could offer valuable insights. A comparison can be made based on data on summer days, where many recreational vessel are present on the waterways, with AIS data of winter days where recreational vessels are absent could unveil differences. While it is generally assumed that the presence of recreational vessels influences the behaviour of inland vessels, this hypothesis can be validated using the proposed methodology. The findings could potentially lead to recommendations, such as suggesting increased separation between these vessel types to mitigate associated risks.

9.2. Future research

Future research may explore the impact of individual features, address data challenges, investigate the inclusion of ship-ship interactions, and account for external conditions. The upcoming sections will delve deeper into these aspects.

In the current research, many features are used to define vessel behaviour. However, the individual impact of these features remains unexamined. Future research could focus on isolating the effect of each feature on clustering results, identifying which play an important role, and determining whether similar results can be achieved using only a part of the features.

Furthermore, additional research could analyse the clustering results. It could be interesting to further evaluate a cluster, is it possible to determine the specific locations within a cluster where patterns deviate from the norm. An observation from this research is that the accidents found at the Schellingwouderbrug lie on the outside of their respective clusters in Figure 6.43. This raises questions about the extent to which outliers within these clusters can be identified by the methodology.

Another challenge is the amount of input data, as accidents are scarce on waterways increasing the chances of finding one with larger data sets. Therefore, this study aimed to use parallel programming with Dask, to handle large volumes of data. While several issues with mismatching python packages have been solved, a new challenge arises in managing numerous partitions of varying sizes. Overcoming the wide range in partition size could potentially be achieved by repartitioning the data set in a more intelligent way. Currently, the data set is repartitioned based on unique vessel names, resulting in large partitions for frequently navigating vessels and small partitions for those passing through once. This introduces skewness over the partitions and can lead to problems. A combination of vessel name and date would probably result in more equal partitions which in general perform better in Dask.

This research provided a brief description of behaviour during ship-ship interactions but does not further apply this aspect in the model. Ship-ship interactions introduce more challenges, as two moving objects both in time and space are involved. Initially, attempts were made to filter these interactions based on temporal and spatial criteria and identify vessel trajectory interactions within bounding boxes. However, this approach did not yield satisfactory results and was not further employed in defining the model's features. Where including the ship-ship accidents is of great importance in the overall safety assessment.

Shipping behaviour is significantly influenced by external factors such as weather, currents, and visibility, as indicated by Shu et al. (2017). When a coupling is made with hydrodynamic or weather models, this can give a more complete insight. For instance, during heavy wind conditions, vessels may adopt different angles concerning the waterway. If such weather conditions are known for specific days, they need not be categorised as anomalous behaviour. Similarly, strong currents can affect vessel speed. In the context of ship-infrastructure, water levels can have a substantial impact. Low water levels can

narrow down waterways, forcing vessels to sail closer to each other, while high water levels can pose challenges for navigating bridges due to reduced headroom. It's important to note that this research primarily focuses on horizontal distances, but vertical interactions could also be significant in these scenarios.

9.3. Recommendations to the field

Throughout this research, several discoveries result in recommendations to the field, specifically concerning accident registration.

In many cases, the absence of accurate and comprehensive accident information presents a challenge. Leveraging AIS data immediately after a observed or reported incident could offer substantial assistance. Storing AIS data from a vessel directly after an incident enables reconstruction of the situation before, during and, after the event. This reconstructed accident data can be valuable in validating or correcting filed reports, enhancing the overall quality of accident data.

Moreover, the AIS data from known accidents can serve as valuable resources for subsequent studies, either as validation or as training data for models. Despite limitations related to the logging of AIS data due to privacy concerns, exploring exceptions or possibilities to use this data solely for the purpose of enhancing safety could be beneficial.

References

Al Shaaili, M., Al Alawi, M., Ekyalimpa, R., Al Mawli, B., Al-Mamun, A., & Al Shahri, M. (2023). Near-miss accidents data analysis and knowledge dissemination in water construction projects in Oman. *Heliyon*, *9*(11), e21607. https://doi.org/10.1016/j.heliyon.2023.e21607

Ashush, N., Greenberg, S., Manor, E., & Ben-Shimol, Y. (2023). Unsupervised Drones Swarm Characterization Using RF Signals Analysis and Machine Learning Methods. *Sensors*, *23*(3). https://doi.org/10.3390/s23031589

Baak, J. (2023). *The ship domain in port areas* [Doctoral dissertation, Delft University of Technology]. http://resolver.tudelft.nl/uuid:02b5f319-d825-48cb-b902-84fd71676a71

Bahmani, B., Moseley, B., Vattani, A., Kumar, R., & Vassilvitskii, S. (2012). Scalable K-Means++. *Proceedings of the VLDB Endowment (PVLDB)*, *5*(7), 622–633. https://doi.org/https://doi.org/10.48550/arXiv.1203.6402

Beenhakker, C., & Schelling, I. (2020, February). *Risicoanalyse aanvaringen van bruggen, sluizen, stuwen en keringen* (tech. rep.). Rijkswaterstaat Water Verkeer en Leefomgeving.

Berglund, R., & Huttunen, M. "AIS data analysis for identification of close encounter situations of vessels". http://www.vtt.fi/inf/julkaisut/muut/2009/SafetySecurityReview09.pdf *VTT Technical Research Centre of Finland*.

Binnenvaartpolitiereglement. "." https://wetten.overheid.nl/BWBR0003628/2017-01-01/0#DeelI_Hoofdstuk6 *Last accessed: aug. 31, 2023*.

Bureau Telematica Binnenvaart. (2009). Wat is AIS? Een introductie van AIS voor de binnenvaart.

Caliński, T., & Harabasz, J. (1974). A Dendrite Method For Cluster Analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

CBS. "Hoeveel vaarwegen zijn er in Nederland?" https://www.cbs.nl/nl-nl/visualisaties/verkeer-en-vervoer/vervoermiddelen-en-infrastructuur/vaarwegen *Last accessed: dec. 28, 2022*.

CBS. "Hoeveel binnenvaartschepen zijn er in Nederland?" https://www.cbs.nl/nl-nl/visualisaties/verkeer-en-vervoer/vervoermiddelen-en-infrastructuur/binnenvaartschepen *Last accessed: jan. 31, 2023*.

CBS. "Snelle indicatoren Goederenvervoer". https://www.cbs.nl/nl-nl/visualisaties/snelle-indicatoren-goederenvervoer

CBS. "Goederenvervoer; vervoerwijzen, vervoerstromen van en naar Nederland". https://opendata.cbs.nl/#/CBS/nl/dataset/83101NED/line?ts=1686304421531

Chen, X., Ling, J., Yang, Y., Zheng, H., Xiong, P., Postolache, O., & Xiong, Y. (2020). Ship Trajectory Reconstruction from AIS Sensory Data via Data Quality Control and Prediction. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/7191296

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2019). *tsfresh Documentation Release 0.12.0* (tech. rep.). Blue Yonder GmbH.

Cui, M. (2020). Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. *Geoscience and Remote Sensing*, *3*, 9–16. https://doi.org/10.23977/accaf.2020.010102

Daranda, A., & Dzemyda, G. (2020). Navigation decision support: Discover of vessel traffic anomaly according to the historic marine data. *International Journal of Computers, Communications and Control*, *15*(3). https://doi.org/10.15837/IJCCC.2020.3.3864

Dask Development Team. "Dask: Library for dynamic task scheduling". https://dask.org

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

De Gelderlander. "Schepen moeten om via Brabant of België door aanvaring". https://www.gelderlander.nl/grave/schepen-moeten-om-via-brabant-of-belgie-door-aanvaring~a632f087/ *Last accessed: sep. 22, 2023*.

De Havenloods. "Bijna werd de Spijkenisserbrug geramd door een zeeschip, het ging maar nét goed! " https://www.dehavenloods.nl/nieuws/algemeen/49947/bijna-werd-de-botlekbrug-geramd-door-een-zeeschip-het-ging-ma *Last accessed: nov. 7, 2023*.

de Ruijter, A., & Guldenmund, F. (2015). The bowtie method: A review. *Safety Science*, *88*, 211–218. https://doi.org/10.1016/j.ssci.2016.03.001

Di Ciccio, C., van der Aa, H., Cabanillas, C., Mendling, J., & Prescher, J. (2016). Detecting flight trajectory anomalies and predicting diversions in freight transportation. *Decision Support Systems*, *88*, 1–17. https://doi.org/10.1016/j.dss.2016.05.004

Dilkhaz, S. (2019). A comparative study of safety leading and lagging indicators measuring project safety performance. *Advances in Science, Technology and Engineering Systems*, *4*(6), 306–312. https://doi.org/10.25046/aj040639

Ehlers, U. C., Ryeng, E. O., McCormack, E., Khan, F., & Ehlers, S. (2017). Assessing the safety effects of cooperative intelligent transport systems: A bowtie analysis approach. *Accident Analysis and Prevention*, *99*, 125–141. https://doi.org/10.1016/j.aap.2016.11.014

Emmens, T., Amrit, C., Abdi, A., & Ghosh, M. (2021). The promises and perils of Automatic Identification System data. *Expert Systems with Applications*, *178*. https://doi.org/10.1016/j.eswa.2021.114975

European Commission. "Promotion of inland waterway transport". https://transport.ec.europa.eu/transport-modes/inland-waterways/promotion-inland-waterway-transport_en *Last accessed: aug. 2, 2023.*

Felski, A., & Jaskolski, K. (2013). The Integrity of Information Received by Means of AIS During Anticollision Manoeuvring. *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, *7*(2), 95–100. https://doi.org/10.12716/1001.07.01.12

Ferroli, P., Caldiroli, D., Acerbi, F., Scholtze, M., Piro, A., Schiariti, M., Orena, E. F., Castiglione, M., Broggi, M., Perin, A., & Dimeco, F. (2012). Application of an aviation model of incident reporting and investigation to the neurosurgical scenario: Method and preliminary data. *Neurosurgical Focus*, *33*(5). https://doi.org/10.3171/2012.9.FOCUS12252

Fournier, M., Casey Hilliard, R., Rezaee, S., & Pelot, R. "Past, present, and future of the satellite-based automatic identification system: areas of applications (2004–2016)". *WMU Journal of Maritime Affairs*. ISSN: 16541642. https://doi.org/10.1007/s13437-018-0151-6

Fujii, Y., & Tanaka, K. (1971). Traffic capacity. *The Journal of navigation*, *24*(4), 543–552.

GeoJSON.io. "." https://geojson.io/#map=2/0/20 *Last accessed: nov. 28, 2023.*

Goerlandt, F., Montewka, J., Lammi, H., & Kujala, P. "Analysis of near collisions in the Gulf of Finland". *Advances in Safety, Reliability and Risk Management*. https://doi.org/10.1201/b11433-409

Graser, A. (2019). MovingPandas: Efficient structures for movement data in Python. *GI_Forum Journal for Geographic Information Science*, *7*(1), 54–68. https://doi.org/10.1553/GISCIENCE2019{_}01{_}S54

GVB. "Lines". https://reisinfo.gvb.nl/en/lijnen?boat&show *Last accessed: sep. 28, 2023.*

Halkidi, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, *17*(3), 107–145.

Harbers, M. "Ongevalscijfers Scheepvaart 2022 [Letter of government]". https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjV9MD9gfAAxVQ2gIHHeq_CQcQFnoECBUQAQ&url=https%3A%2F%2Fwww.rijksoverheid.nl%2Fbinaries%2Frijksoverheid%2Fdocumenten%2Fkamerstukken%2F2023%2F07%2F13%2Fongevalscijfers-scheepvaart-2022%2Fongevalscijfers-scheepvaart-2022.pdf&usg=AOvVaw00Grft1Cv078Yx1oqcmuVN&opi=89978449 *Ministerie van Infrastructuur en Waterstaat.*

Hofmeijer, A. F. (2019, December). *Monitor Nautische Veiligheid Binnenwateren 2009-2018 Datarapport* (tech. rep.). Rijkswaterstaat Water Verkeer en Leefomgeving.

Hofmeijer, A. F. (2020, February). *Monitor Nautische Veiligheid Binnenwateren 2009-2018 Risicoanalyserapport* (tech. rep.). Rijkswaterstaat Water Verkeer en Leefomgeving.

Hozumi, Y., Wang, R., Yin, C., & Wei, G. W. (2021). UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. *Computers in Biology and Medicine*, *131*. https://doi.org/10.1016/j.compbiomed.2021.104264

Huang, X., Wen, Y., Zhang, F., Han, H., Huang, Y., & Sui, Z. "A review on risk assessment methods for maritime transport". *Ocean Engineering*. ISSN: 00298018. https://doi.org/10.1016/j.oceaneng.2023.114577

International Maritime Organization. (2015, December). *Resolution A.1106(29) REVISED GUIDELINES FOR THE ONBOARD OPERATIONAL USE OF SHIPBORNE AUTOMATIC IDENTIFICATION SYSTEMS (AIS)* (tech. rep.). https://edocs.imo.org/Final

Iphar, C., Ray, C., & Napoli, A. (2020). Data integrity assessment for maritime anomaly detection. *Expert Systems with Applications*, *147*. https://doi.org/10.1016/j.eswa.2020.113219

Jin, X., & Han, J. (2017). K-Means Clustering. *Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning and Data Mining*, 563–564. https://doi.org/https://doi.org/10.1007/978-1-4899-7687-1{_}431

Jollife, I. T., & Cadima, J. "Principal component analysis: A review and recent developments". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. ISSN: 1364503X. https://doi.org/10.1098/rsta.2015.0202

Leland McInnes. (2023, September). *umap Documentation Release 0.5* (tech. rep.).

Leveson, N. (2015). A systems approach to risk management through leading safety indicators. *Reliability Engineering and System Safety*, *136*, 17–34. https://doi.org/10.1016/j.ress.2014.10.008

Marx, D. A., & Slonim, A. D. (2003). Assessing patient safety risk before the injury occurs: An introduction to sociotechnical probabilistic risk modelling in health care. *Quality and Safety in Health Care*, *12*(SUPPL. 2). https://doi.org/10.1136/qhc.12.suppl{_}2.ii33

McInnes, L., Healy, J., Saul, N., & Grossberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *The Journal of Open Source Software*, *3*(29), 861.

Mestl, T., Tallakstad, K. T., & Castberg, R. (2016). Identifying and Analyzing Safety Critical Maneuvers from High Resolution AIS Data. *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, *10*(1), 69–77. https://doi.org/10.12716/1001.10.01.07

Montewka, J., Hinz, T., Kujala, P., & Matusiak, J. (2010). Probability modelling of vessel collisions. *Reliability Engineering and System Safety*, *95*(5), 573–589. https://doi.org/10.1016/j.ress.2010.01.009

MOSWOZ. "Onderzoek naar scheepvaartveiligheid". https://www.noordzeeloket.nl/beleid/interdepartementaal-directeuren-overleg-noordzee/idon-nieuwsbrief/nr-38/moswoz-onderzoek-scheepvaartveiligheid/ *INTEGRAAL BEHEER NOORDZEE*.

Movares. (2013a, December). *Monitoring Nautische Veiligheid 2013 Binnenwateren Deel 1: beleidsrelevante rapportage* (tech. rep.). Rijkswaterstaat Water, Verkeer en Leefomgeving, AfdelingVeiligheidsmanagement en Verkeersveiligheid. Delft.

Movares. (2013b, December). *Monitoring Nautische Veiligheid 2013 Binnenwateren Deel 2: ondersteunend cijfermateriaal* (tech. rep.). Rijkswaterstaat Water, Verkeer en Leefomgeving, AfdelingVeiligheidsmanagement en Verkeersveiligheid. Delft.

Nielsen, D. S. (1971). The cause/consequence diagram method as a basis for quantitative accident analysis. *Risø National Laboratory*, *Risø-M*(1374). https://api.semanticscholar.org/CorpusID:210911257

OpenStreetMap. "." https://www.openstreetmap.org/ *Last accessed: nov. 28, 2023*.

Patel, H. "What is Feature Engineering — Importance, Tools and Techniques for Machine Learning". https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10 *Towards Data Science*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pérez, F. L., & Clemente, J. A. (2007). The influence of some ship parameters on manoeuvrability studied at the design stage. *Ocean Engineering*, *34*(3-4), 518–525. https://doi.org/10.1016/j.oceaneng.2006.02.004

Rao, M. (1969). Cluster Analysis and Mathematical Programming. *Mathematical Programming*, *79*, 30. https://doi.org/10.2307/2283542

Ravankar, A., Ravankar, A. A., Kobayashi, Y., Hoshino, Y., & Peng, C. C. "Path smoothing techniques in robot navigation: State-of-the-art, current and future challenges". *Sensors (Switzerland)*. ISSN: 14248220. https://doi.org/10.3390/s18093170

Rijkswaterstaat. "Hollandsch Diep". https://www.rijkswaterstaat.nl/water/vaarwegenoverzicht/hollandsch-diep *Last accessed: sep. 26, 2023*.

Rijkswaterstaat. "IJ". https://www.rijkswaterstaat.nl/water/vaarwegenoverzicht/ij *Last accessed: sep. 26, 2023*.

Rijkswaterstaat. "Volkeraksluizen". https://www.rijkswaterstaat.nl/water/waterbeheer/bescherming-tegen-het-water/waterkeringen/deltawerken/volkeraksluizen *Last accessed: sep. 26, 2023*.

Rijkswaterstaat. (2018, May). *SOS-Formulier: versie 4.1.1 Handleiding en Toelichting* (tech. rep.). Ministerie van Infrastructuur en Milieu.

Rijkswaterstaat. "AIS: verkeersinformatie scheepvaart". https://www.rijkswaterstaat.nl/zakelijk/verkeersmanagement/scheepvaart/scheepvaartverkeersbegeleiding/river-information-services/automatic-identification-system *Last accessed: jan. 3, 2023.*

Rijkswaterstaat. "Pleziervaart". https://www.rijkswaterstaat.nl/water/scheepvaart/pleziervaart *Last accessed: feb. 01, 2023.*

Rijkswaterstaat. "Scheepsongevallen (Openbaar)". https://maps.rijkswaterstaat.nl/gwproj55/index.html?viewer=Scheepsongevallen_Openbaar.Webviewer *Last accessed: aug. 24, 2023.*

Rijkswaterstaat. "Scheepvaartverkeerswet". https://www.rijkswaterstaat.nl/water/wetten-regels-en-vergunningen/scheepvaart/scheepvaartverkeerswet *Last accessed: aug. 2, 2023.*

Rijkswaterstaat WVL. (2021, May). *Achtergrondrapportage Vaarwegen Integrale Mobiliteitsanalyse 2021* (tech. rep.). Rijkswaterstaat Water Verkeer en Leefomgeving. https://open.overheid.nl/repository/ronl-0e501f1c-e81c-463a-8695-f619461a6049/1/pdf/bijlage-4-achtergrondrapport-3-vaarwegen.pdf

Rijkwaterstaat. (2020, December). *Waterway Guidelines 2020* (tech. rep.). Rijkswaterstaat Water, Verkeer en Leefomgeving, Rijswijk, Rijkswaterstaat, Ministerie van Infrastructuur en Waterstaat.

Rousseeuw, P. J. (1987). *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis* (tech. rep.).

Sainburg, T., McInnes, L., & Gentner, T. Q. (2021). Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, *33*(11), 2881–2907. https://doi.org/10.1162/neco{_}a{_}01434

Satopää, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. *31st International Conference on Distributed Computing Systems Workshops*, 166–171. https://doi.org/https://doi.org/10.1109/ICDCSW.2011.20

Scheepvaartverkeerswet. "." *https://wetten.overheid.nl/BWBR0004364/2021-07-01.*

Shu, Y., Daamen, W., Ligteringen, H., & Hoogendoorn, S. P. (2017). Influence of external conditions and vessel encounters on vessel behavior in ports and waterways using Automatic Identification System data. *Ocean Engineering*, *131*, 1–14. https://doi.org/10.1016/j.oceaneng.2016.12.027

Sklet, S. (2004). Comparison of some selected methods for accident investigation. *Journal of Hazardous Materials*, *111*(1-3), 29–37. https://doi.org/10.1016/j.jhazmat.2004.02.005

Stevens, S. S. (1946). *On the Theory of Scales of Measurement* (tech. rep. No. 2684).

Szlapczynski, R., & Szlapczynska, J. (2017). Review of ship safety domains: Models and applications. *Ocean Engineering*, *145*, 277–289. https://doi.org/10.1016/j.oceaneng.2017.09.020

The European Commission. (2019). Legislation 273. *Official Journal of the European Union*, *62*, 8–9.

Van Der Maaten, L., & Hinton, G. (2008). *Visualizing Data using t-SNE* (tech. rep.).

van der Werff, S. E. "trajectories.ipynb". https://github.com/TUDelft-CITG/AIS-utilities/blob/master/notebooks/attic/trajectories.ipynb

Van Koningsveld, M., Verheij, H. J., Taneja, P., & De Vriend, H. J. (2021, August). *Ports and Water ways Navigating the changing world.*

Varen doe je Samen! "Rijkswaterstaat pakt kruising Hollandsch Diep met Dordtsche Kil aan". https://varendoejesamen.nl/kenniscentrum/artikel/rijkswaterstaat-pakt-kruising-hollandsch-diep-met-dordtsche-kil-aan *Last accessed: sep. 26, 2023.*

Varen doe je Samen. "Meld onveilige situaties met de Vaar Melder-app". https://varendoejesamen.nl/kenniscentrum/artikel/vaar-melder-app#:~:text=Bel%20bij%20spoed%20altijd%20112%20of%20meld%20het%20via%20de%20marifoon.&text=Naast%20het%20monitoren%20van%20de,de%20veiligheid%20op%20het%20water. *Last accessed: aug. 28, 2023.*

Varen doe je Samen! "Doorstroming Moerdijkbruggen voor kleinere binnenvaart verbeterd". https://varendoejesamen.nl/kenniscentrum/artikel/doorstroming-moerdijkbruggen-voor-kleinere-binnenvaart-verbeterd *Last accessed: sep. 26, 2023.*

Vereniging vrienden van veerponten. "veerponten.nl". https://veerponten.nl *Last accessed: sep. 26, 2023.*

Vessel Tracking and Tracing Expert Group (RIS VTT). (2014). *Guidelines on the Installation of the Inland Automatic Identification System* (tech. rep.). www.ccr-zkr.org

Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). *Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization* (tech. rep.). http://jmlr.org/papers/v22/20-1061.html.

Wang, Y., Yang, L., Song, X., & Li, X. (2021). Ship classification based on random forest using static information from AIS data. *Journal of Physics: Conference Series*, *2113*(1). https://doi.org/10.1088/1742-6596/2113/1/012072

Waterkaart Live. "Schellingwouderbrug, in Schellingwoude: openingstijden en contact". https://waterkaart.net/gids/brug.php?naam=Schellingwouderbrug *Last accessed: sep. 28, 2023.*

Waterkaart Live. "Waterkaart". https://waterkaart.net *Last accessed: sep. 28, 2023.*

Waterrecreatie Nederland. (2016, November). *Basisvisie Recreatietoervaart Nederland 2015-2020* (tech. rep.).

Watersportverbond. "Regioteam geeft inspraak bij gebiedsontwikkeling rond Oranjesluizen Amsterdam". https://www.watersportverbond.nl/nieuws/regioteam-geeft-inspraak-bij-gebiedsontwikkeling-rond-oranjesluizen-amsterdam/ *Last accessed: sep. 26, 2023.*

Widyantara, I. M. O., Hartawan, I. P. N., Karyawati, A. A. I. N. E., Er, N. I., & Artana, K. B. (2023). Automatic identification system-based trajectory clustering framework to identify vessel movement pattern. *IAES International Journal of Artificial Intelligence*, *12*(1), 1–11. https://doi.org/10.11591/ijai.v12.i1.pp1-11

Wright, L., & Van Der Schaaf, T. (2004). Accident versus near miss causation: A critical review of the literature, an empirical test in the UK railway domain, and their implications for other sectors. *Journal of Hazardous Materials*, *111*(1-3), 105–110. https://doi.org/10.1016/j.jhazmat.2004.02.049

Yan, Z., Song, X., Zhong, H., Yang, L., & Wang, Y. (2022). Ship Classification and Anomaly Detection Based on Spaceborne AIS Data Considering Behavior Characteristics. *Sensors (Basel, Switzerland)*, *22*(20). https://doi.org/10.3390/s22207713

Zhang, W., Goerlandt, F., Montewka, J., & Kujala, P. (2015). A method for detecting possible near miss ship collisions from AIS data. *Ocean Engineering*, *107*, 60–69. https://doi.org/10.1016/j.oceaneng.2015.07.046

Zhou, Y., Daamen, W., Vellinga, T., & Hoogendoorn, S. P. (2019). Ship classification based on ship behavior clustering from AIS data. *Ocean Engineering*, *175*, 176–187. https://doi.org/10.1016/j.oceaneng.2019.02.005

A Appendix A: VesseltypeERI

Table A.1: ERI code and ship description (Vessel Tracking and Tracing Expert Group (RIS VTT), 2014)

Code	Ship type	Code	Ship type
8000	Vessel, type unknown	8310	Pushtow, one tank/gas barge
8010	Motor freighter	8320	Pushtow, two barges at least one tanker or gas barge
8020	Motor tanker	8330	Pushtow, three barges at least one tanker or gas barge
8021	Motor tanker, liquid cargo, type N	8340	Pushtow, four barges at least one tanker or gas barge
8022	Motor tanker, liquid cargo, type C	8350	Pushtow, five barges at least one tanker or gas barge
8023	Motor tanker, dry cargo as if liquid (e.g. cement)	8360	Pushtow, six barges at least one tanker or gas barge
8030	Container vessel	8370	Pushtow, seven barges at least one tanker or gas barge
8040	Gas tanker	8380	Pushtow, eight barges at least one tanker or gas barge
8050	Motor freighter, tug	8390	Pushtow, nine or more barges at least one tanker or gas barge
8060	Motor tanker, tug	8400	Tug, single
8070	Motor freighter with one or more ships alongside	8410	Tug, one or more tows
8080	Motor freighter with tanker	8420	Tug, assisting a vessel or linked combination
8090	Motor freighter pushing one or more freighters	8430	Pushboat, single
8100	Motor freighter pushing at least one tankship	8440	Passenger ship, ferry, cruise ship, red cross ship
8110	Tug, freighter	8441	Ferry
8120	Tug, tanker	8442	Red cross ship
8130	Tug freighter, coupled	8443	Cruise ship
8140	Tug, freighter/tanker, coupled	8444	Passenger ship without accomodation
8150	Freightbarge	8450	Service vessel, police patrol, port service
8160	Tankbarge	8460	Vessel, work maintainance craft, floating derrick, cable-ship, buoy-ship, dredge
8161	Tankbarge, liquid cargo, type N	8470	Object, towed, not otherwise specified
8162	Tankbarge, liquid cargo, type C	8480	Fishing boat
8163	Tankbarge, dry cargo as if liquid (e.g. cement)	8490	Bunkership
8170	Freightbarge with containers	8500	Barge, tanker, chemical
8180	Tankbarge, gas	8510	Object, not otherwise specified
8210	Pushtow, one cargo barge	1500	General cargo vessel maritime
8220	Pushtow, two cargo barges	1510	Unit carrier maritime
8230	Pushtow, three cargo barges	1520	Bulk carrier maritime
8240	Pushtow, four cargo barges	1530	Tanker
8250	Pushtow, five cargo barges	1540	Liquified gas tanker
8260	Pushtow, six cargo barges	1850	Pleasure craft, longer than 20 metres
8270	Pushtow, seven cargo barges	1900	Fast ship
8280	Pushtow, eigth cargo barges	1910	Hydrofoil
8290	Pushtow, nine or more barges	1920	Catamaran fast

Table A.2: ERI codes classified as inland vessels

Code													
8010	8020	8021	8022	8023	8030	8040	8050	8060	8070	8080	8090	8100	8110
8120	8130	8140	8150	8160	8161	8162	8163	8170	8180	8210	8220	8230	8240
8250	8260	8270	8280	8290	8310	8320	8330	8340	8350	8360	8370	8380	8390
8490	8500												

B

Appendix B: Additional information on current method of nautical safety determination

Table B.1: Effectscore table used in the Monitor Nautische Veiligheid (Hofmeijer, 2020)

Effect class	Safety, Health, Society	Environmental damage	Economic damage	Effect score SOS
5 - Very serious	Multiple deaths or injuries	Extensive damage to flora and fauna in a large area,recovery taking years	Disruption of the fairway for more than 7 days and/or material damage exceeding 100 million	100.000
4 - Serious	One dead or missing	Serious disruption for more than 1 year in a medium-sized area. Partial recovery of environmental values possible within a few years	Disruption of the fairway for multiple days, and/or material damage between 15 million and 100 million	10.000
3 - limited	Multiple severely injured	Local disruption, medium-sized, and mostly temporary damage to flora and fauna, unwanted environmental impact lasts a maximum of 1 year. Full environmental recovery is possible	Disruption of the fairway for 1 day, and/or material damage between 1 million and 15 million	1.000
2 - light	One severely injured	Short-term exceedance of threshold values in a small area without lasting damage to flora and fauna. Full recovery is assured.	Disruption of the fairway on the order of 2 hours, and/or material damage up to 1 million	100
1 - Very light	Light injury	Threshold values for pollution are not exceeded	Less than 1 hour of disruption, and/or material damage on the order of tens of thousands of euros	10
0 - Nil	No casualties	no environmental damage	No economic damage	0

C

Appendix C: Accidents registered in public SOS database

AIS data is available for the months of January, April, July, and October 2019, for the same months the SOS-database is used to check for accidents recorded during these months. The recorded accidents in categories ship-ship, ship-object and ship-infrastructure can be found in tables C.1 and Table C.2.

Table C.1: Accidents recorded in public SOS-Database, Hollands Diep

Classification	Registration number	Month	Year	Accident	Vessels involved
non-significant	201952942	April	2019	ship-infrastructure	Only inland vessels
significant	201950336	July	2019	ship-ship	Only inland vessels
non-significant	201952610	July	2019	ship-ship	Only inland vessels
non-significant	201950605	July	2019	ship-infrastructure	Only recreational
significant	201953881	July	2019	ship-infrastructure	Only recreational
non-significant	201953434	July	2019	ship-ship	Inland-Seagoing
non-significant	201953414	October	2019	ship-object	Only other

Table C.2: Accidents recorded in public SOS-Database, the IJ

Classification	Registration number	Month	Year	Accident	Vessels
non-significant	201946407	January	2019	ship-infrastructure	
non-significant	201946723	January	2019	ship-infrastructure	
non-significant	201950244	April	2019	ship-ship	Only inland
non-significant	201950393	July	2019	ship-ship	Only inland

D

Appendix D: Additional figures location of interest

D.1. Hollands Diep

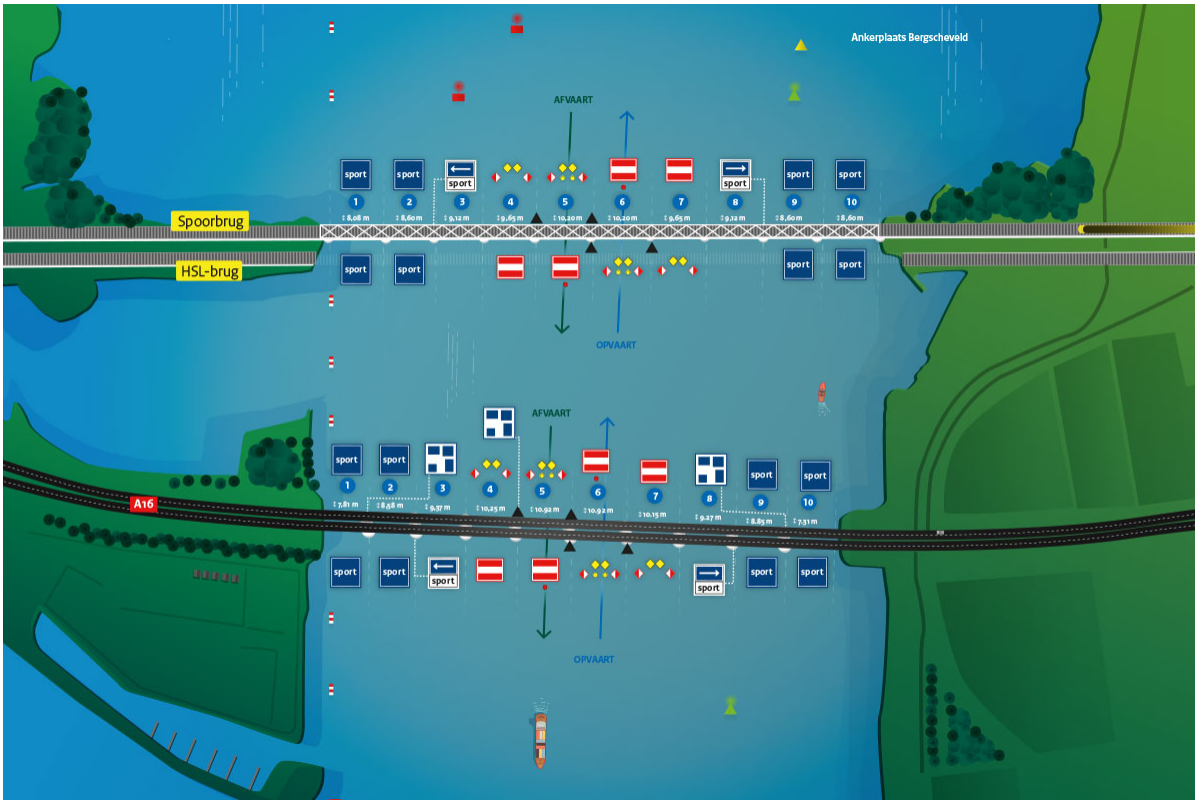


Figure D.1: Rules for passge Moerijkdbrug (Varen doe je Samen!, 2018)

D.2. The IJ

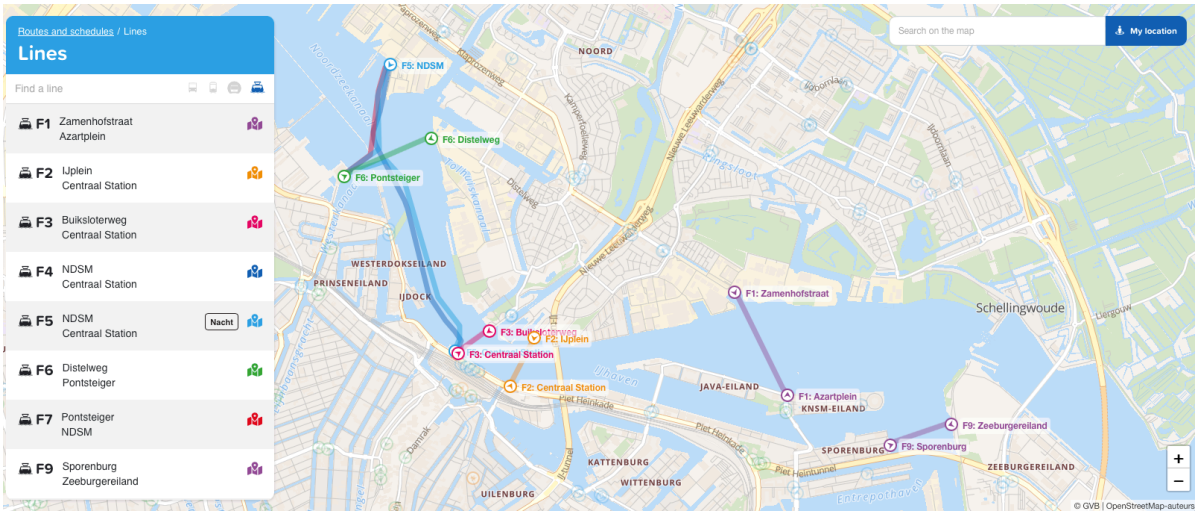


Figure D.2: Ferry lines crossing the IJ (GVB, 2023)

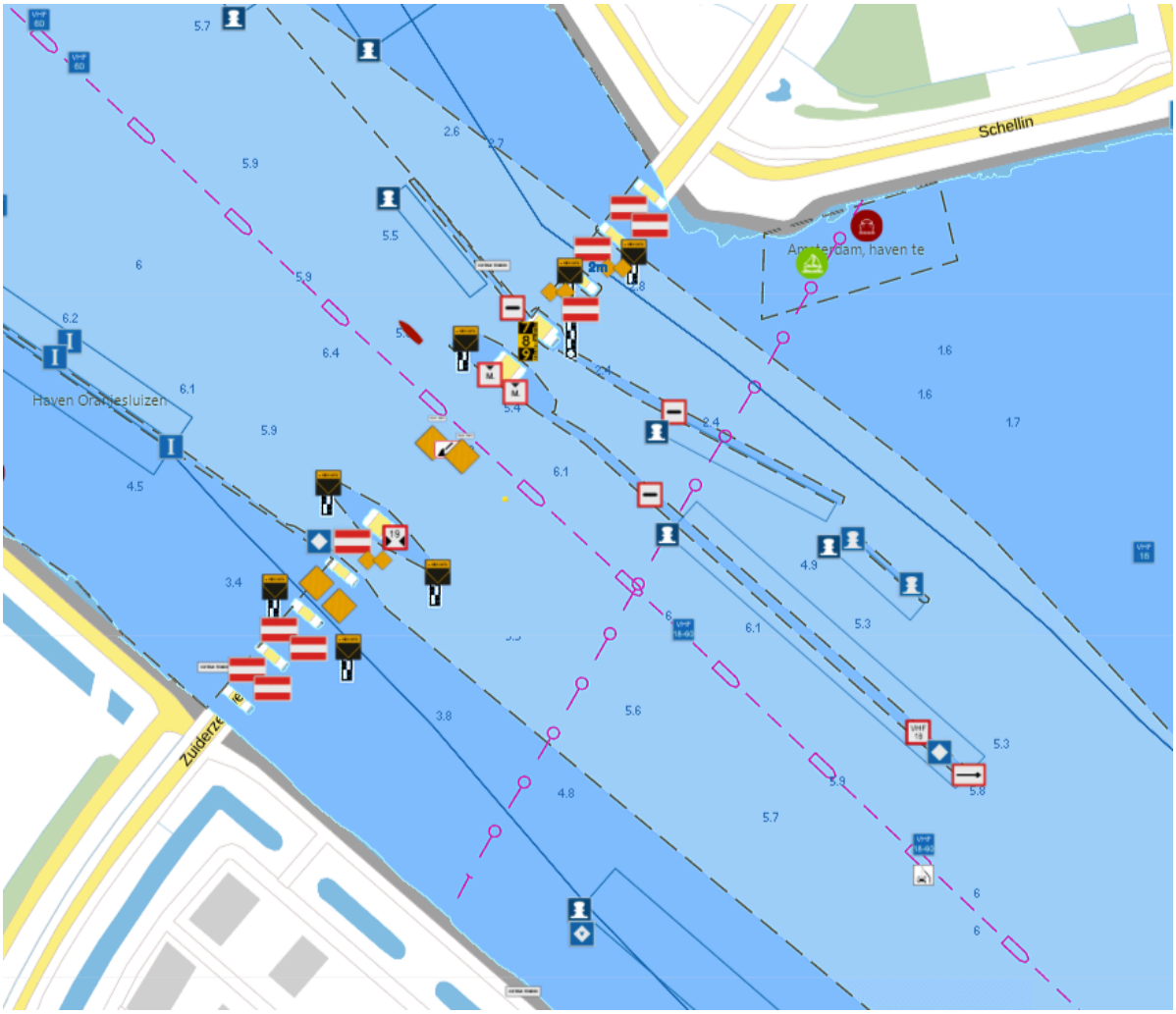


Figure D.3: Rules for passage Schellingwouderbrug (Waterkaart Live, 2023b)

E

Appendix E: Feature tables

Table E.1: Overview of general features

General		
Vessel type		
	vessel_type_rec	vessel_type_inland
Speed-related		
	speed_maximum speed_median	speed_minimum speed_standard_deviation
Acceleration-related		
	acc_maximum acc_median	acc_minimum acc_standard_deviation
Direction-related		
	COG_maximum COG_median	COG_minimum COG_standard_deviation
Manoeuvring-related		
	ROT_maximum ROT_median l_b_ratio	ROT_minimum ROT_standard_deviation

Table E.2: Overview of ship-infrastructure interaction features

Ship-infrastructure		
Distance		
	minimal _distance	
200 meter around bridge		
	speed_near_bridge_max_200	speed_near_bridge_min_200
	speed_near_bridge_std_200	speed_near_bridge_median_200
	acc_near_bridge_max_200	acc_near_bridge_min_200
	acc_near_bridge_std_200	acc_near_bridge_median_200
	COG_diff_near_bridge_max_200	COG_diff_near_bridge_min_200
	COG_diff_near_bridge_std_200	COG_diff_near_bridge_median_200
	ROT_near_bridge_max_200	ROT_near_bridge_min_200
	ROT_near_bridge_std_200	ROT_diff_near_bridge_median_200
400 meter around bridge		
	speed_near_bridge_max_400	speed_near_bridge_min_400
	speed_near_bridge_std_400	speed_near_bridge_median_400
	acc_near_bridge_max_400	acc_near_bridge_min_400
	acc_near_bridge_std_400	acc_near_bridge_median_400
	COG_diff_near_bridge_max_400	COG_diff_near_bridge_min_400
	COG_diff_near_bridge_std_400	COG_diff_near_bridge_median_400
	ROT_near_bridge_max_400	ROT_near_bridge_min_400
	ROT_near_bridge_std_400	ROT_diff_near_bridge_median_400

Table E.3: Overview of features generated with tsfresh

Tsfresh		
Speed-related		
	speed_quantile_q_0.1	speed_quantile_q_0.9
	speed_number_peaks_n_1	speed_number_peaks_n_3
	speed_number_peaks_n_5	speed_number_peaks_n_10
	speed_number_peaks_n_50	
Acceleration-related		
	acc_quantile_q_0.1	acc_quantile_q_0.9
	acc_number_peaks_n_1	acc_number_peaks_n_3
	acc_number_peaks_n_5	acc_number_peaks_n_10
	acc_number_peaks_n_50	
Direction-related		
	COG_quantile_q_0.1	COG_quantile_q_0.9
	COG_number_peaks_n_1	COG_number_peaks_n_3
	COG_number_peaks_n_5	COG_number_peaks_n_10
	COG_number_peaks_n_50	
Manoeuvring-related		
	ROT_quantile_q_0.1	ROT_quantile_q_0.9
	ROT_number_peaks_n_1	ROT_number_peaks_n_3
	ROT_number_peaks_n_5	ROT_number_peaks_n_10
	ROT_number_peaks_n_50	

F
Appendix F: Results

Additional figures depicting the results of ship infrastructure interaction for Hollands Diep and the IJ are presented in this appendix.

F.1. Hollands Diep

F.1.1. Characteristics input data

Table F.1: Trip characteristics Hollands Diep

Characteristic	
Total number of trips	2322
Average trip length	61 data points
Longest trip	1791 data points
Shortest trip	11 data points
Percentage recreational	0.2 %
Percentage inland	83.3 %
Percentage other type	16.5 %

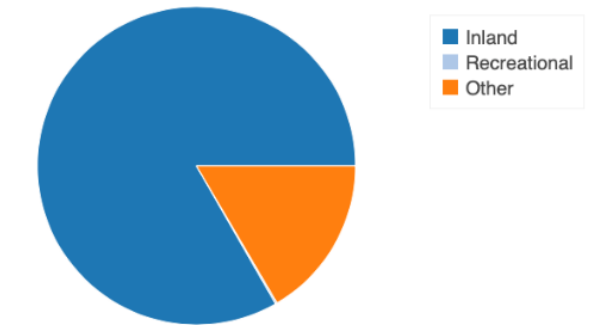


Figure F.1: Distribution of vessel categories over the data, Hollands Diep

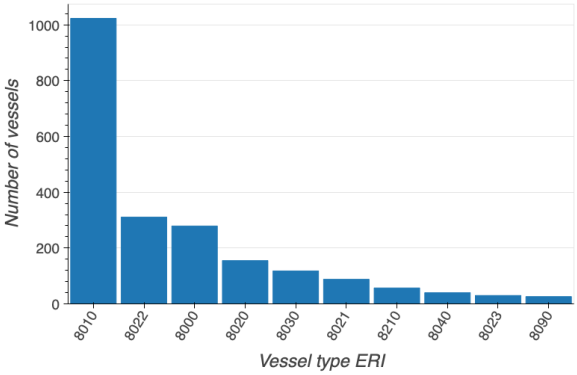


Figure F.2: Top 10 vessel types, Hollands Diep

F.1.2. Evaluation of the clustering

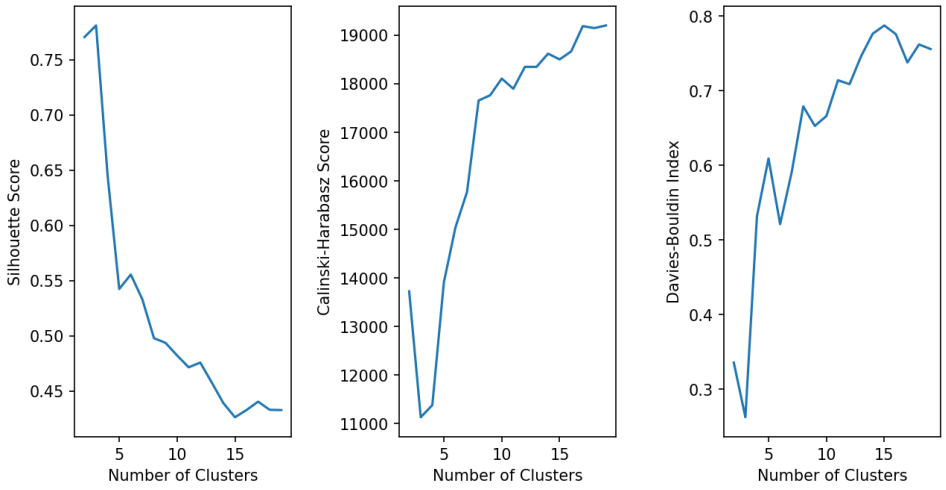


Figure F.3: Scores over the number of clusters for the clustering near the Moerdijkbrug, Hollands Diep

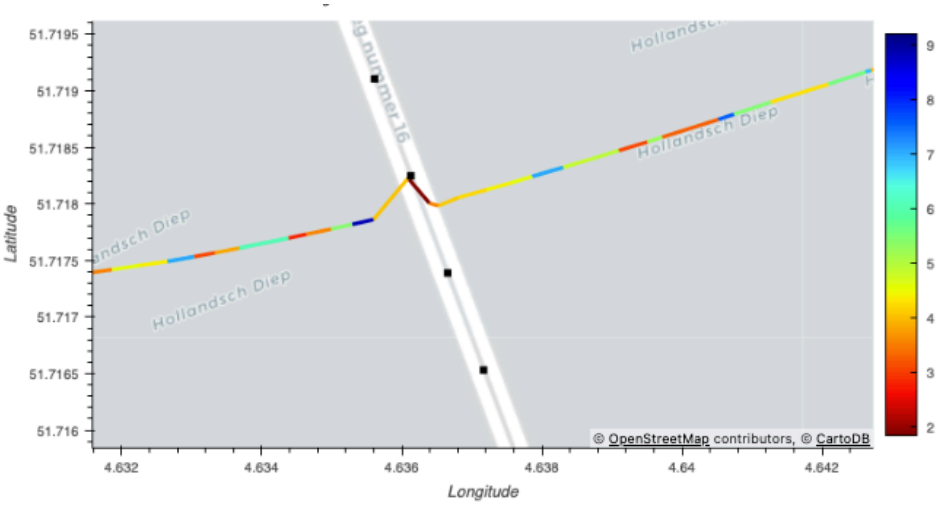


Figure F.4: Zoom to trajectory 2236 close interaction with the bridge, colour bar indicating vessel speed [m/s], Hollands Diep

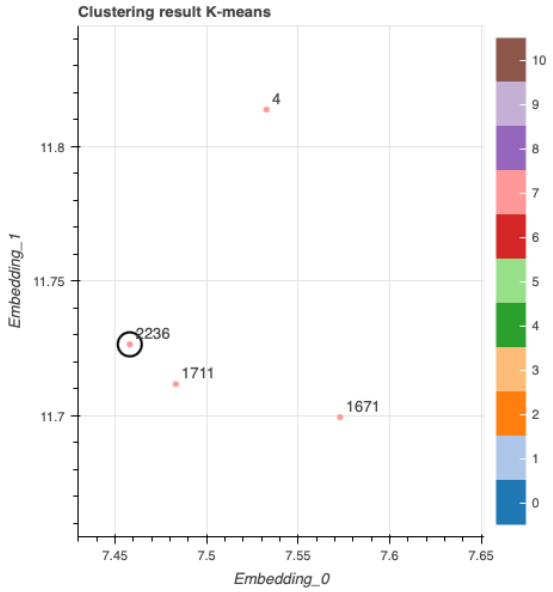
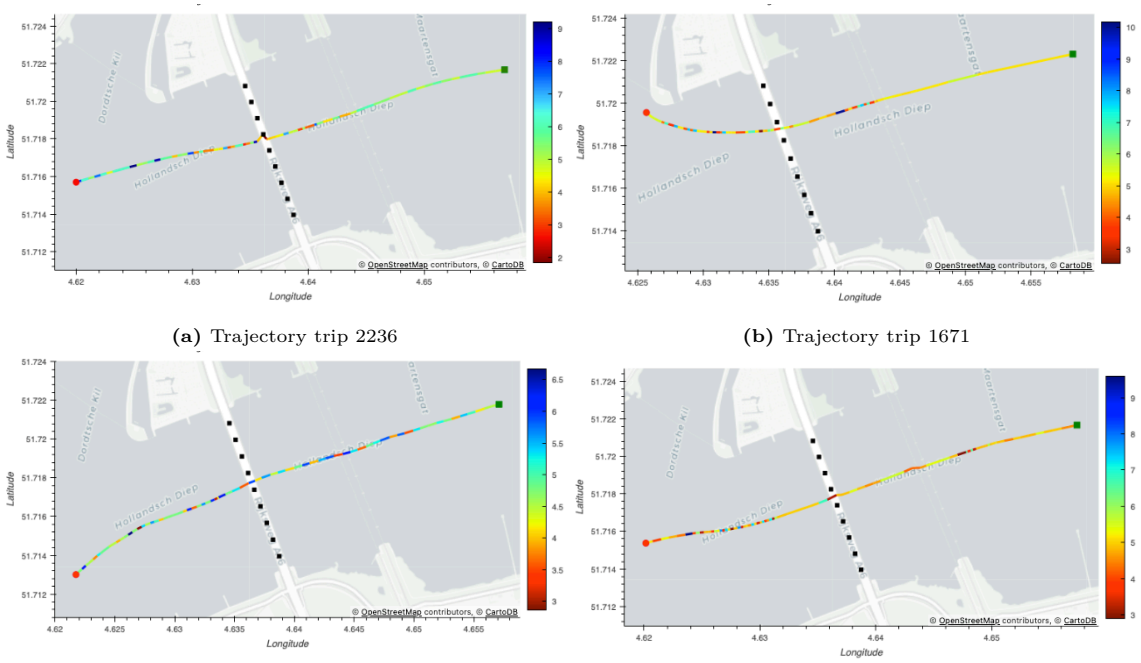


Figure F.5: Zoom to trip 2236 in clustering to find points close by, Hollands Diep



(a) Trajectory trip 2236

(b) Trajectory trip 1671

(c) Trajectory trip 1711

(d) Trajectory trip 4

Figure F.6: Trajectory plots of trips close to trips 2236 in clustering, colour bar indicating vessel speed [m/s], Hollands Diep

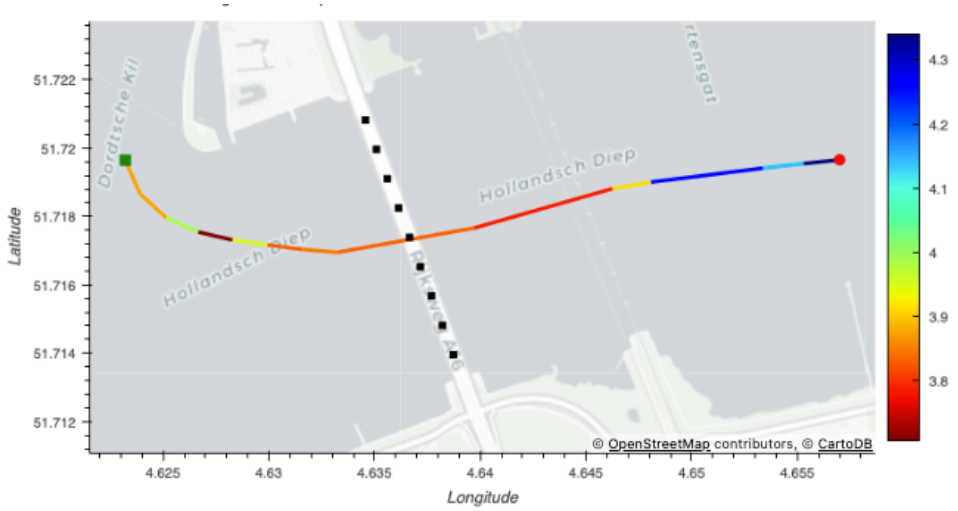


Figure F.7: Trajectory plot trip 1429, 6.5 meter encounter to bridge pillar, colour bar indicating vessel speed [m/s], Hollands Diep

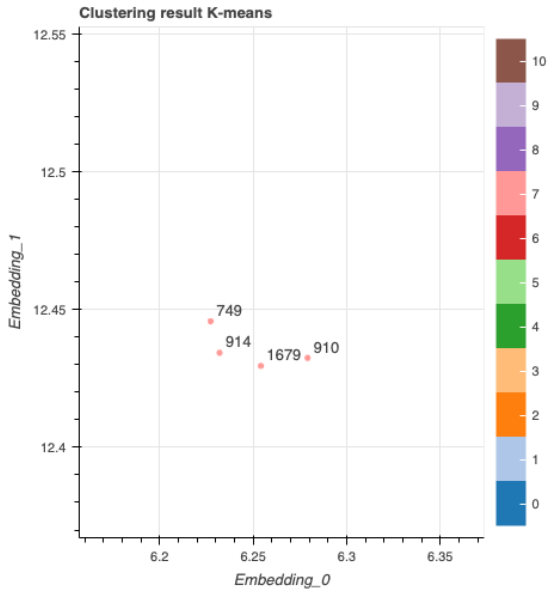
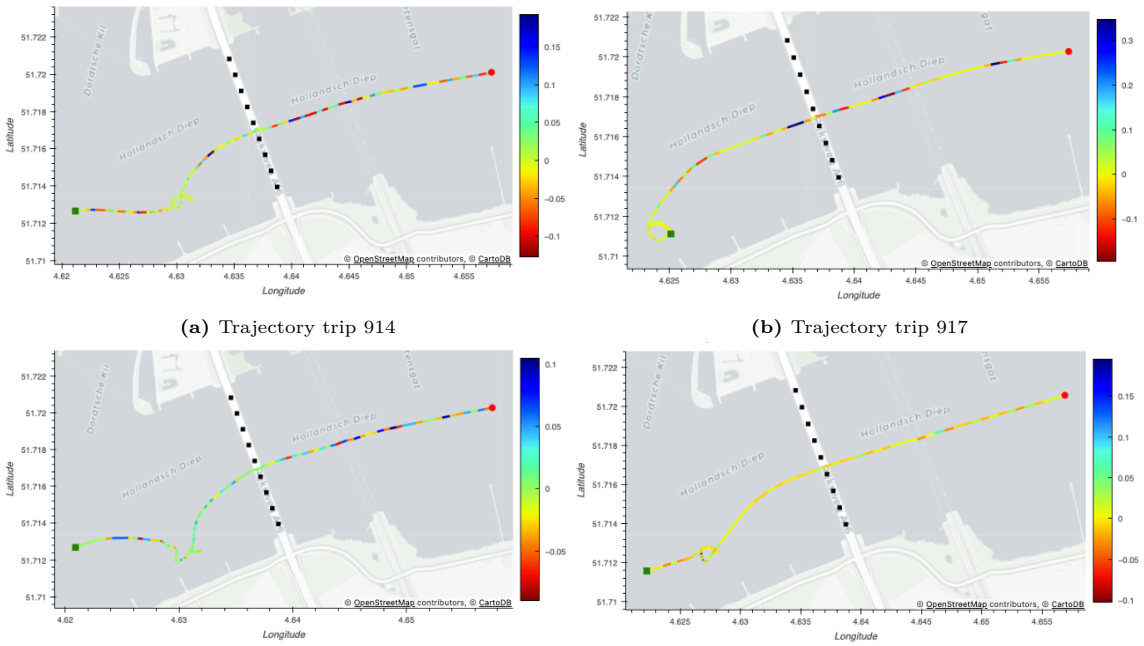


Figure F.8: Zoom to trips in clustering with high number of peaks in acceleration, Hollands Diep



(a) Trajectory trip 914

(b) Trajectory trip 917

(c) Trajectory trip 1679

(d) Trajectory trip 749

Figure F.9: Trajectory plots of trips with high number of peaks in acceleration, colour bar indicating vessel acceleration [m/s^2], Hollands Diep

F.2. The IJ lock complex included

F.2.1. Characteristics input data

Table F.2: Trip characteristics the IJ lock complex included

Characteristic	
Total number of trips	2248
Average trip length	333 data points
Longest trip	4161 data points
Shortest trip	11 data points
Percentage recreational	1.4 %
Percentage inland	62 %
Percentage other type	36.6 %

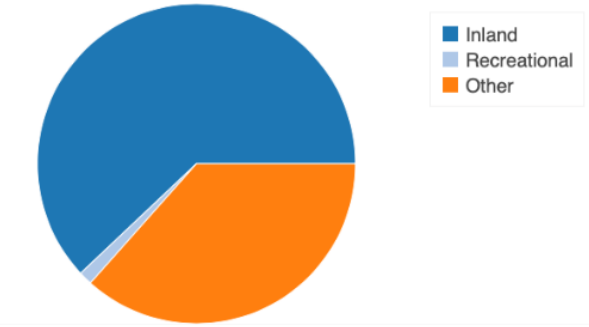


Figure F.10: Distribution of vessel categories over the data, the IJ lock complex included

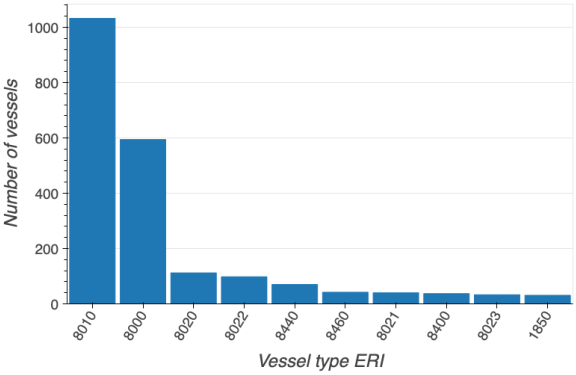


Figure F.11: Top 10 vessel types, the IJ lock complex included

F.2.2. Evaluation of the clustering

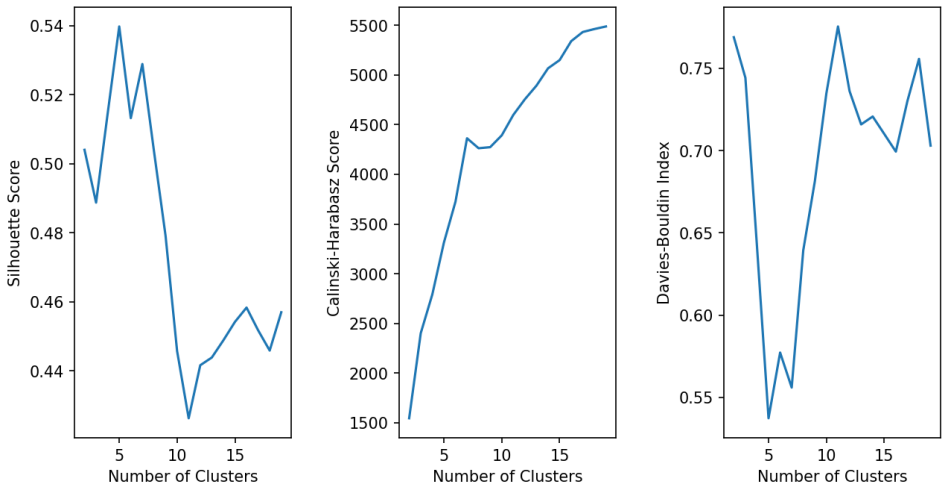


Figure F.12: Scores over the number of clusters, the IJ lock complex included

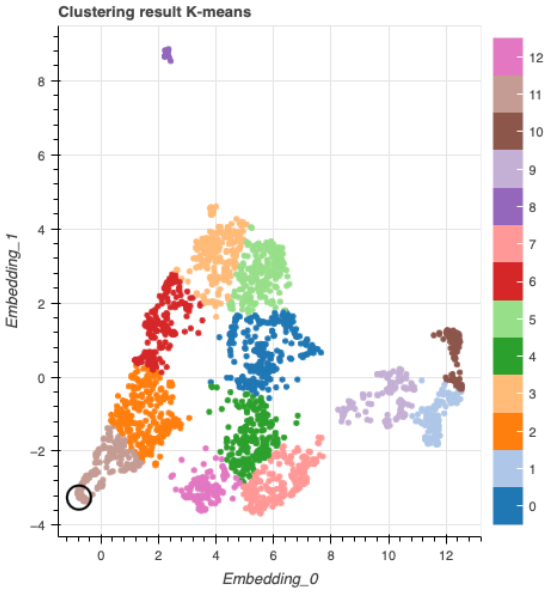


Figure F.13: Scatter plot K-means clustering at Schellingwouderbrug, recreational vessel outside cluster highlighted, the IJ lock complex included

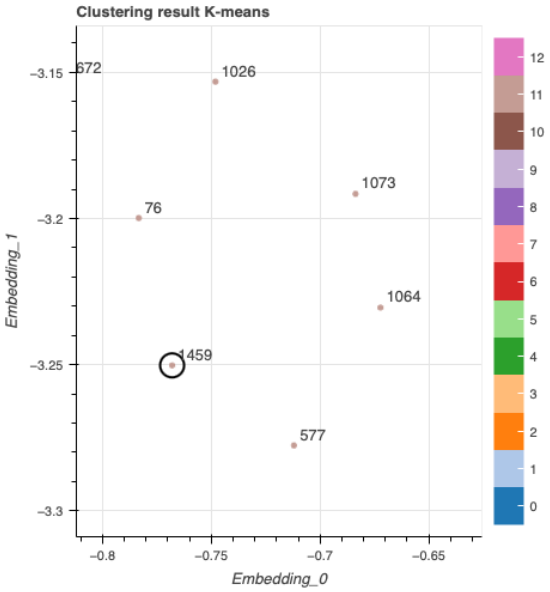
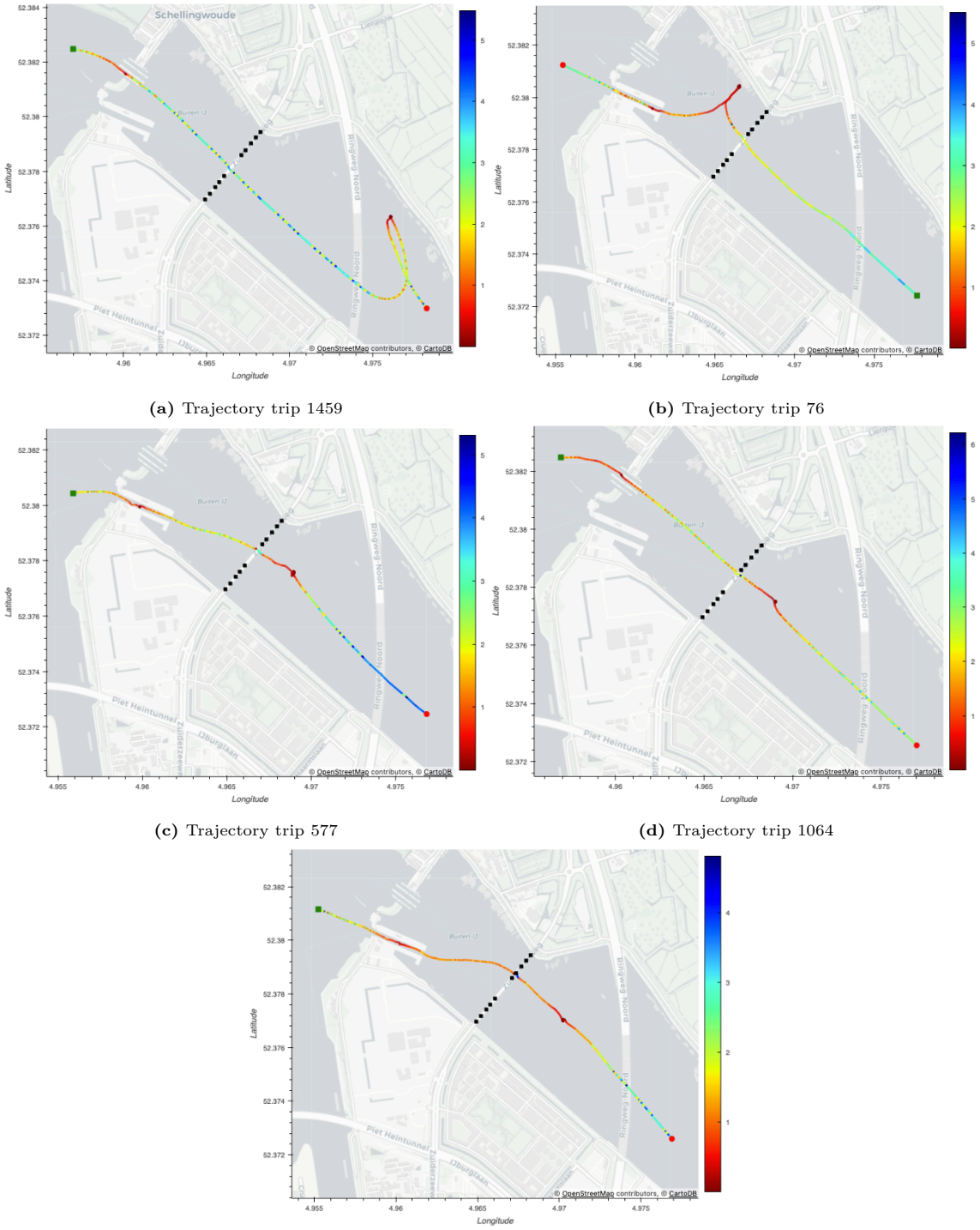


Figure F.14: Zoom towards recreational vessel outside recreational cluster to indicate neighbouring points, the IJ lock complex included



(a) Trajectory trip 1459

(b) Trajectory trip 76

(c) Trajectory trip 577

(d) Trajectory trip 1064

(e) Trajectory trip 1073

Figure F.15: Trajectory plots of trips close to trips 1469 in clustering, colour bar indicating vessel speed [m/s], the IJ lock complex included

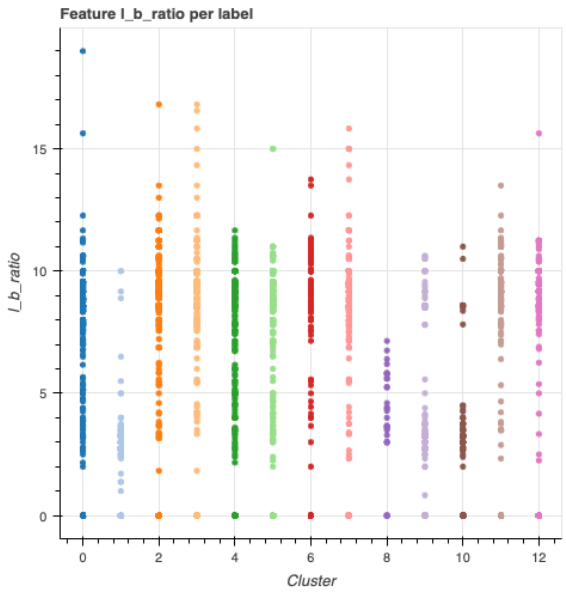
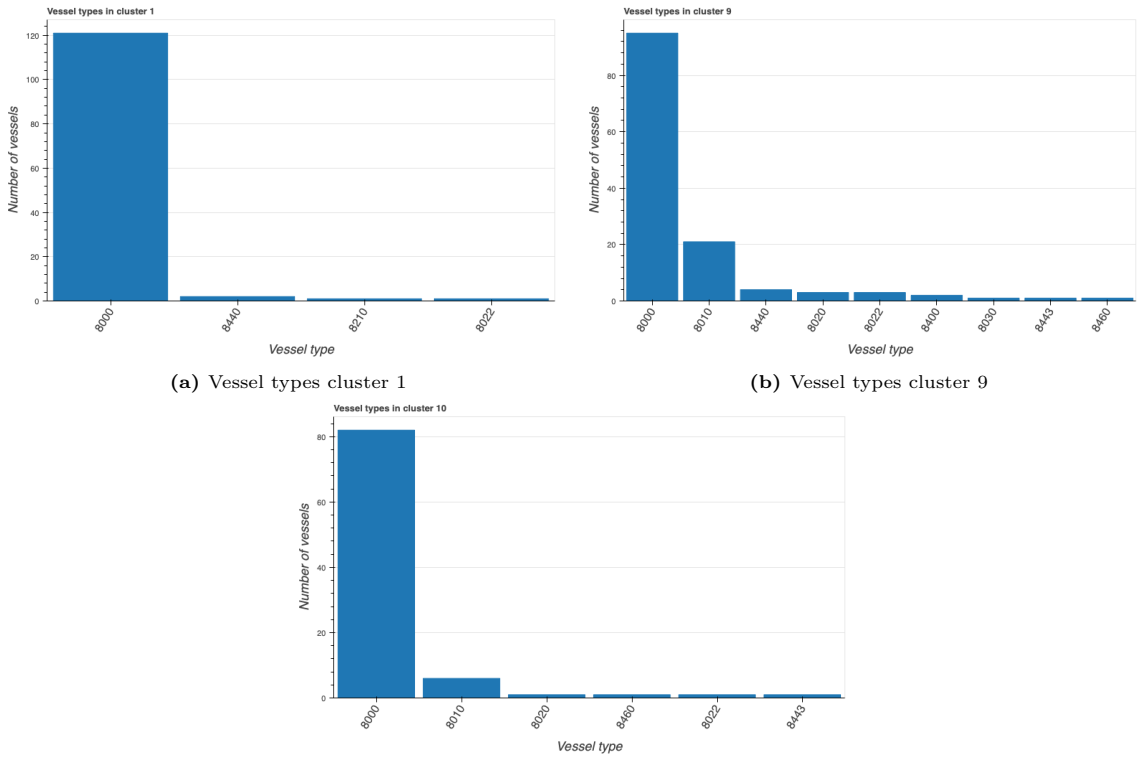


Figure F.16: Feature L/B ratio over the clusters, the IJ lock complex included



(a) Vessel types cluster 1

(b) Vessel types cluster 9

(c) Vessel types cluster 10

Figure F.17: Distribution of vessel type across cluster 1, 9 and 10, the IJ lock complex included

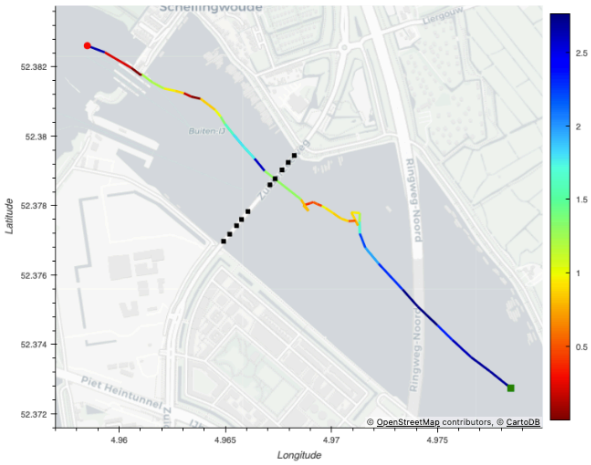


Figure F.18: Trajectory plot of trip 845, 0.9 [m] from Schellingwouderbrug, colour bar indicating vessel speed [m/s], the IJ lock complex included

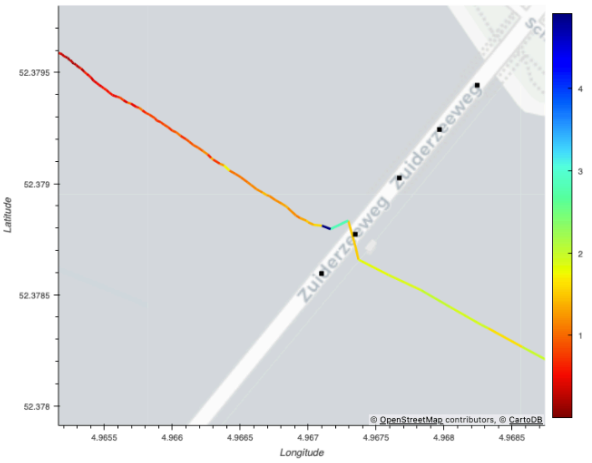


Figure F.19: Zoom trajectory plot of trip 1344, 1.5 [m] from Schellingwouderbrug, colour bar indicating vessel speed [m/s], the IJ lock complex included

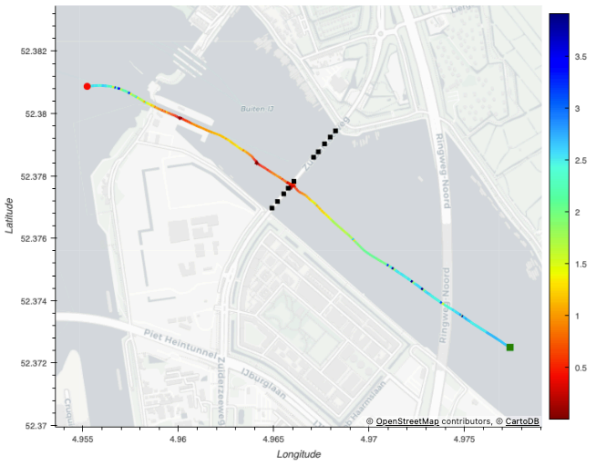


Figure F.20: Trajectory plot of trip 2218, 0.5 [m] from Schellingwouderbrug, colour bar indicating vessel speed [m/s], the IJ lock complex included

Figure F.21: Zoom trajectory plot of trip 2218, 0.5 [m] from Schellingwouderbrug, colour bar indicating vessel speed [m/s], the IJ lock complex included

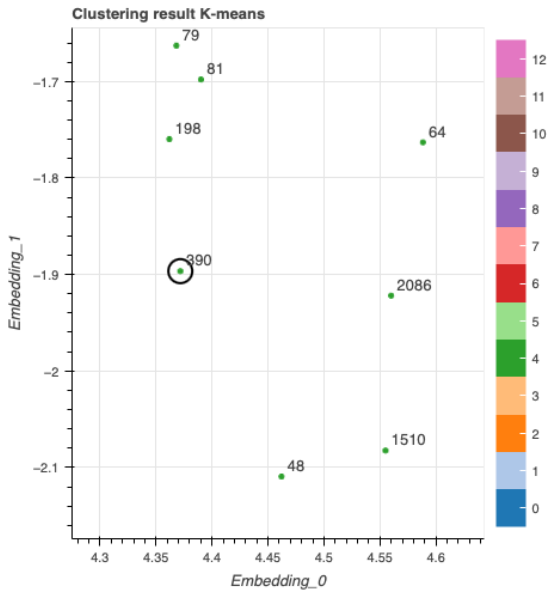
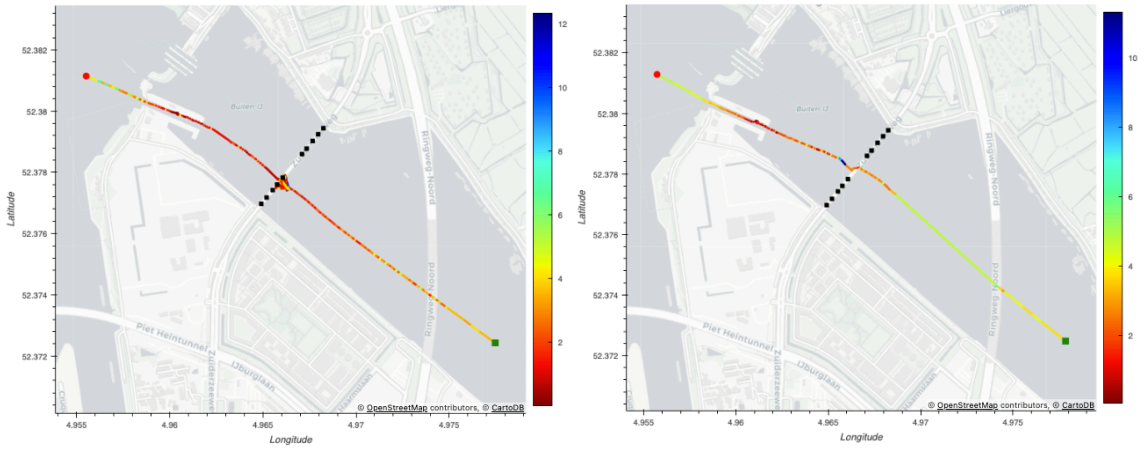
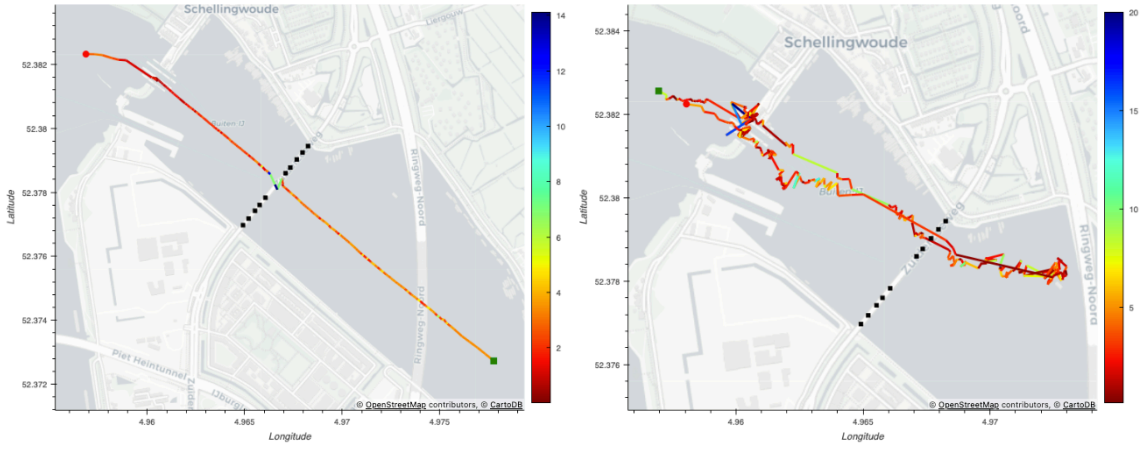


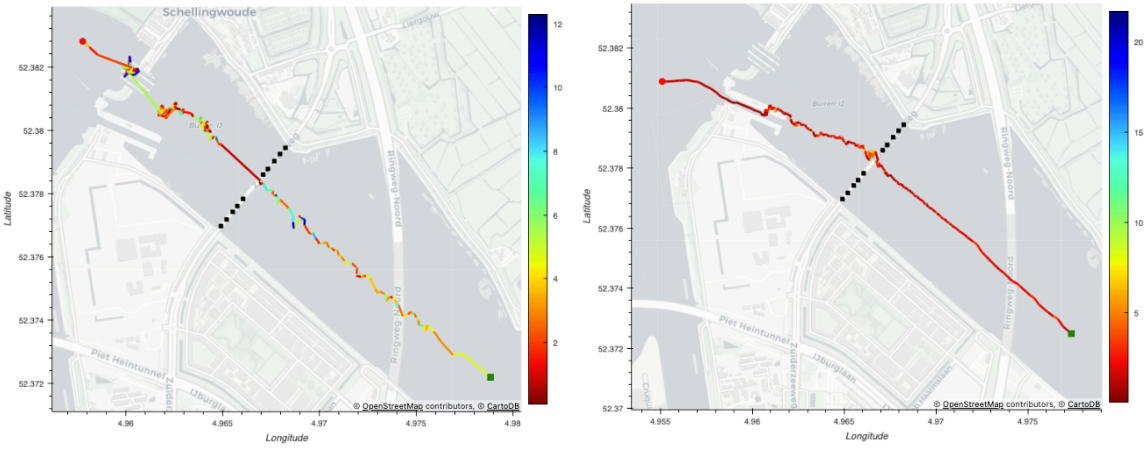
Figure F.22: Zoom towards trip 390 in embedding to indicate neighbouring points, the IJ lock complex included



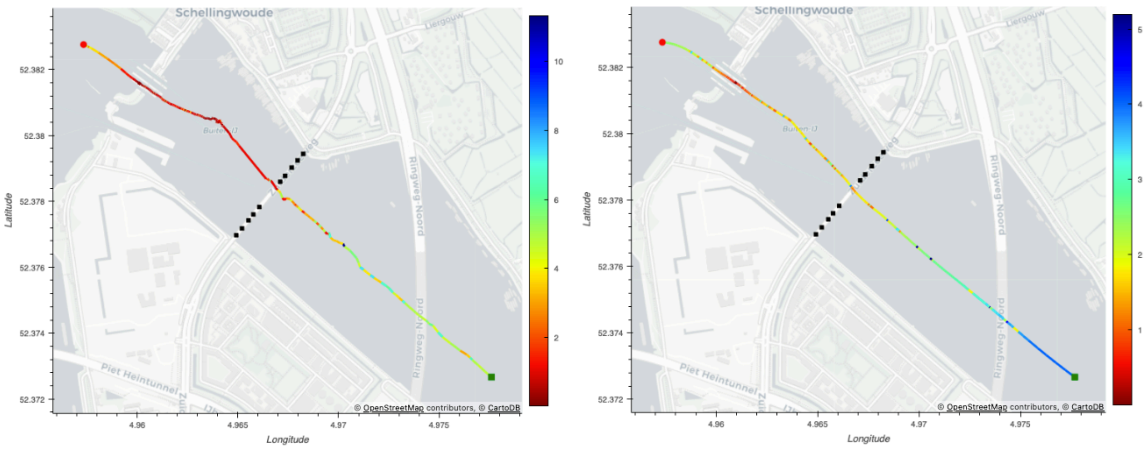
(a) Trajectory trip 390 **(b)** Trajectory trip 49



(c) Trajectory trip 64 **(d)** Trajectory trip 79



(e) Trajectory trip 81 **(f)** Trajectory trip 198



(g) Trajectory trip 1510 **(h)** Trajectory trip 2086

Figure F.23: Trajectory plots of trips close to trips 390 in clustering, colour bar indicating vessel speed [m/s], the IJ lock complex included

F.3. The IJ lock complex excluded

F.3.1. Characteristics input data

Table F.3: Trip characteristics the IJ lock complex excluded

Characteristic	
Total number of trips	2222
Average trip length	156 data points
Longest trip	3897 data points
Shortest trip	11 data points
Percentage recreational	1.4 %
Percentage inland	62.7 %
Percentage other type	35.9 %

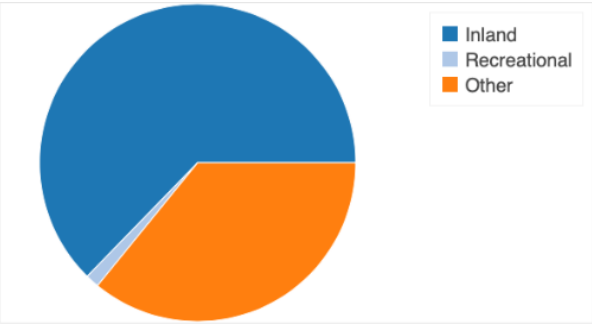


Figure F.24: Distribution of vessel categories over the data, the IJ lock complex excluded

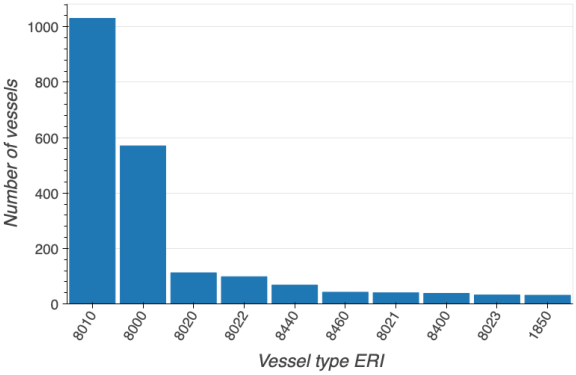


Figure F.25: Top 10 vessel types, the IJ lock complex excluded

F.3.2. Evaluation of the clustering

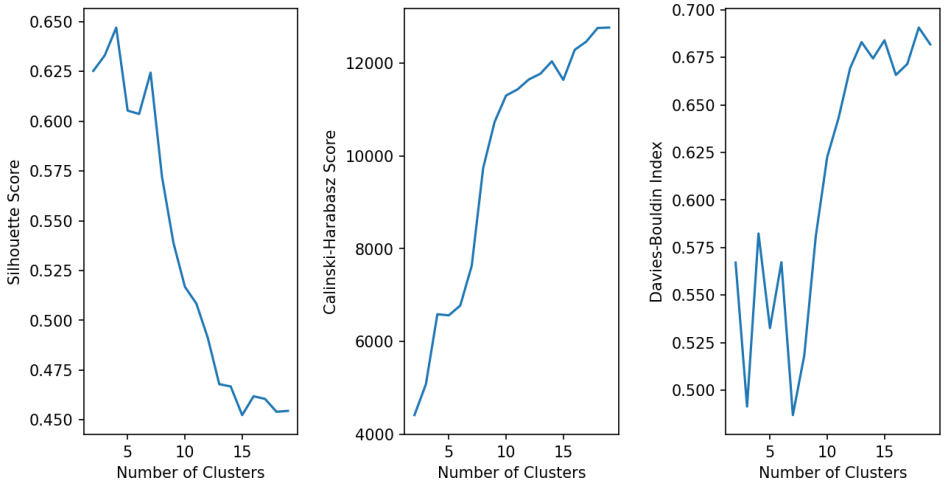


Figure F.26: Scores over the number of clusters, the IJ lock complex excluded

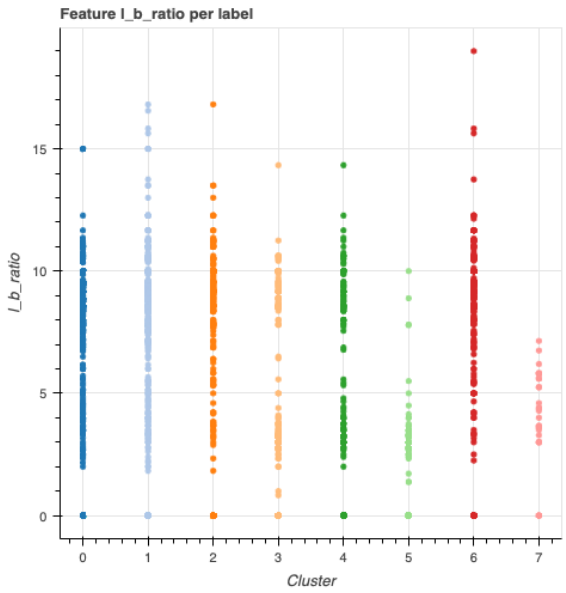


Figure F.27: Feature L/B ratio over the clusters, the IJ lock complex excluded

List of Figures

List of Tables