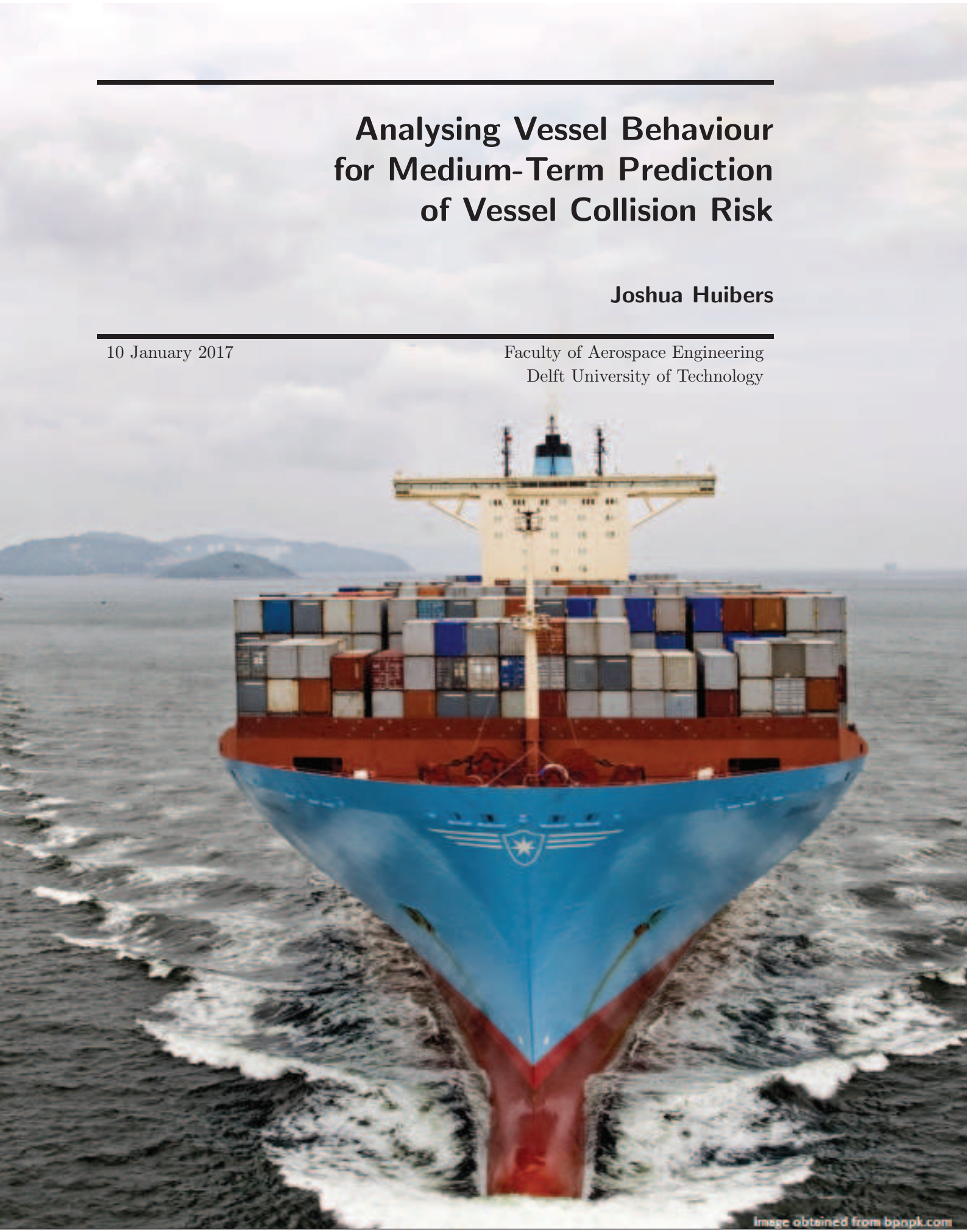
# Analysing Vessel Behaviour for Medium-Term Prediction of Vessel Collision Risk

**Joshua Huibers**

10 January 2017

Faculty of Aerospace Engineering
Delft University of Technology




TU Delft Delft University of Technology

Challenge the future

# Analysing Vessel Behaviour for Medium-Term Prediction of Vessel Collision Risk

Master of Science Thesis

For obtaining the degree of Master of Science in Aerospace Engineering at Delft University of Technology

Joshua Huibers

10 January 2017

Faculty of Aerospace Engineering · Delft University of Technology

Delft University Of Technology
Aerospace Engineering
Control and Operations
Air Transport and Operations

The undersigned hereby certify that they have read and recommend to the Faculty of Aerospace Engineering for acceptance a thesis entitled **"Analysing Vessel Behaviour for Medium-Term Prediction of Vessel Collision Risk"** by **Joshua Huibers** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: 10 January 2017

Safety Chair: _____
prof. dr. ir. H.A.P. Blom

Academic Supervisor: _____
dr. O.A. Sharpanskykh

External Committee member: _____
dr. R.M. Groves

Supervisor Saab Technologies B.V.: _____
ir. R.A. Hogendoorn

2nd Supervisor Saab Technologies B.V.: _____
ir. W.H.L. Neven

# Abstract

Maritime traffic has to deal with the risk of collision on a daily basis. Currently, Vessel Traffic Services Operators are provided with short-term prediction methods, used to resolve potential collisions. Research is done to predict collision risk at a larger time horizon, which is expected to provide a Vessel Traffic Services Operator (VTSO) with more time and information to anticipate upon and prevent situations with high risk of collision from developing. This thesis focusses on forming the basis for the prediction component of this goal. The objective is to provide a basic understanding of the process of medium-term behaviour of vessels. This is done by performing a data analysis of a case study of the vessel traffic off the coast of Rotterdam, investigating which variables can be used to predict the intent of a vessel.

Two aspects of the intent of a vessel are considered: where the vessel intends to end (within the scope of the scene), and which intermediate waypoints it plans to follow. Entry points, exit points and waypoints are clustered using the Density Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) clustering technique. Waypoints are derived by detecting change-points in the course of vessels using binary segmentation (Scott and Knott, 1974). Variables from the dataset are selected and it is investigated which variables can distinguish between different intents, given the entry point of the vessel.

The results are that the variables 'course' and 'destination' can distinguish between routes sufficiently enough to investigate them further. This further investigation for the course variable is due to its dependence on vessel position, among other variables. Other variables may add value also, but then in combination with these two variables. The waypoint determination has not yet been successfully implemented, but it is regarded as a promising means to describe and predict vessel intent in more detail, partially due to the conclusions drawn regarding the course variable.

# Preface

First and foremost, all the honour and credit for any achievement in my life goes to God Most High. Everything I have, and all that I am, I have received from Him. Attaining an MSc Aerospace Engineering degree is therefore something I wish to give back, using it to serve Him and his kingdom, now and in the future. I thank God for this great privilege to serve Him and his creation; without Him I can do nothing.

Leoné, my loving (and lovely) wife, has stood by my side non-stop on all fronts that I have encountered in life in the past four years. I will not forget the many intellectual discussions, the emotional support, the encouragement in faith, the practical help, and the uncontested tender loving care. She is a wonderful person and a great support, enviable by many; not a day goes by that she doesn't make me smile.

My family (including in-laws) have been there for us, encouraging us especially in tough times. I consider it a great blessing to have a family that truly cares, is involved with your life, and is there to help each other. I would specifically like to thank my mother, for her editorial support in the English language for all my major reports.

Koos van der Linden, Daniël Booms, Adnan & Linda John, Laurens Toering, Mark Huizinga and Rianne Huizinga are people I have considered close friends throughout these times. I want to thank them for being good and faithful companions, spending many hours with me in meaningful discussions about faith, life and relationships. However, just as important are the many hours of distraction, fun, sports and games that we spent together. I thank Koos and Rudi for contributing to my thesis by letting their critical eyes glide over the sentences I wrote, and by not withholding their valuable comments.

Thanks is also due to my supervisors at Saab Technologies -René Hogendoorn and Bill Neven- for their input, ideas, insights and patience throughout my thesis. It has been an interesting and educational time to work with them and their company. They, and also the other colleagues from Saab Technologies, as well as the Rotterdam Vessel Traffic services personnel are worth thanking for their expert opinions and their enthusiastic

explanations of the vessel traffic management domain.
I thank Alexei Sharpanskykh, my daily supervisor at the TU Delft, for his patience, explanations, tools and input throughout the entire duration of my thesis. I appreciate his holistic type of thinking, his professional opinion, and his ability to keep me focused and critical. More towards the background of my masters degree, but definitely worth mentioning, professor Henk Blom has played an interesting role. A few subjects of his, but more importantly discussions with him, have enhanced my thinking on the fundamentals of system behaviour, and provided me with tools to express, quantify and even model this.
The list of the persons that have contributed to me leaving the TU Delft as an MSc Aerospace Engineer is far too long to set out here. Many of these people I do not even know by name, because they work in the background. On this list are the many (assistant) professors, phd's, student assistants, fellow students and professional colleagues from which I have gained much knowledge and experience. The secretaries, the janitors, the cleaners, the cafeteria personnel, the administrative personnel, the IT-support departments, the human resource departments, the education/management boards and all other involved people (from the companies I have been at and from the TU Delft) have been essential to this achievement.

Delft, The Netherlands Joshua Huibers
10 January 2017

# Contents

# List of Figures

# List of Tables

# Nomenclature

## Abbreviations

| | |
|---|---|
| **A-KDE** | Adaptive Kernel Density Estimator |
| **AIS** | Automatic Identification System |
| **ATC** | Air Traffic Control |
| **CPA** | Closest Point of Approach |
| **CSV** | Comma Separated File |
| **DBSCAN** | Density Based Spatial Clustering of Applications with Noise |
| **DCPA** | Distance to Closest Point of Approach |
| **DGPS** | Differential GPS |
| **ETA** | Estimated Time of Arrival |
| **GHMM** | Growing Hidden Markov Model |
| **GPS** | Global Positioning System |
| **IMMJPDA*** | Interacting Multiple Model Joint Probabilistic Data Association {*Avoiding Track Coalescence} |
| **IMMJPDA** | Interacting Multiple Model Joint Probabilistic Data Association |
| **IMM** | Interacting Multiple-Model |
| **ITM** | Instantaneous Topological Map |
| **JPDA** | Joint Probabilistic Data Association |
| **KDE** | Kernel Density Estimation |
| **kNN** | K Nearest Neighbour |

| | |
|---|---|
| **MDTC** | Medium Distance To Collision |
| **MMSI** | Maritime Mobile Service Identity |
| **MTCD** | Medium Term Conflict Detection |
| **NAIL** | Neural Associative Incremental Learning |
| **PDA** | Probabilistic Data Association |
| **PDF** | Probability Density Function |
| **PF** | Particle Filter |
| **SA** | Situation Awareness |
| **TCPA** | Time to Closest Point of Approach |
| **TREAD** | Traffic Route Extraction and Anomaly Detection |
| **TSS** | Traffic Separation Scheme |
| **VCRO** | Vessel Conflict Ranking Operator |
| **VTSO** | Vessel Traffic Service Operators |
| **VTS** | Vessel Traffic Services |

# Chapter 1

# Introduction

The transport sector is a major component of the economy and so also of society, and within the transport sector, the largest group by far is sea traffic. Day and night, many forms of cargo are being transported from one country to another. It is a quiet phenomenon that runs far away in the background of our everyday lives, but an essential component nevertheless. In the traffic sector, the main goal is efficiency, but a more crucial factor is safety - more specifically safety in terms of minimizing the risk of collisions. Traffic inherently has a degree of collision risk, and vessel traffic at sea is not an exception. This thesis takes its place within this area, contributing to the decrease of collision risk at sea.

## 1.1 Objective

Vessel traffic involves many types of risk. One risk is that a vessel can collide with another vessel or object. This risk is higher in areas where more ships come together, such as near ports and in choke-points. This risk may be decreased by anticipating what will happen, in order to take preventive actions. For this reason, Vessel Traffic Service Operators (VTSO) are appointed to maintain an overview of traffic scenes, which they use to inform and advise the involved parties. Each VTSO has responsibility for a part of the map. Several VTSO 's operate together from one control centre to manage the traffic. Support is provided by the V3000 system, which gives a visual representation of the map and the locations of all vessels.
One yet unexplored area of risk assessment considered promising is that of medium term prediction of collision risk. It is seen as promising because it may provide a VTSO more time to anticipate and prevent situations with high risk of collision from developing. The medium-term time horizon (15 minutes or more) is a primary area to which this research contributes. Currently, collision (risk) prediction is restricted to approximately 3-5 minutes in advance. Much research has been done with respect to this short time frame for the sake of conflict resolution and collision avoidance. However, when predictions are extended over a longer amount of time, the uncertainties increase, and the regular models used for short term prediction are inadequate.
This brings up the larger research, which this thesis is part of. This larger research has the following objective: To enable VTSO 's to prevent vessel collision risk from developing by designing a method to predict future collision at a medium-term time horizon. This thesis forms the basis for the prediction component of this objective. It focuses on understanding the medium-term behaviour of vessels, for the purpose of prediction. Therefore, the objective of this thesis is formulated as:
*To provide a basic understanding of the process of medium-term behaviour of vessels.*

## 1.2 Research Question

For a larger prediction time horizon, the dynamic model of a vessel becomes less important, while the intent of the vessel (overall aim or planned path) becomes more important (Lancia et al., 2014). Every vessel captain has a plan in mind: where he/she wants to go, when he/she plans to arrive and how to get there. Unfortunately, this information is not (directly) available. There is no 'sail plan' mandate similar to the 'flight plan' regulations in air traffic. This research studies the traffic off the coast of Rotterdam harbour, controlled by the Hook of Holland Vessel Traffic Services Centre. Here, only the vessels that plan to enter the harbour within 24 hours are obliged to provide their destination (e.g. the name of a dock within the harbour) and Estimated Time of Arrival (ETA). Other vessels are free to provide such information, which is not commonly done (reliably). This is an incentive to derive the intent of a vessel based on the data that *is* available. Therefore, the main research question of this thesis is:
*Which variables can be used to predict the intent of a vessel?*

## 1.3 Report Structure

In section 1.4 the case study used in this thesis is described, after which section 1.5 gives an explanation of important definitions and concepts, to facilitate the reader. Then chapter 2 starts off with an extensive literature survey, which is especially useful for the larger research within which this thesis takes place. It is intended to give an understanding of the current state of related research, to state possible contributions, and in general to provide background to the problem and the domain. Chapter 3 describes the dataset used in this research. Chapter 4 then zooms in on the aim of the thesis, and discusses the research question, its sub-questions and the overall approach taken. This is followed by chapter 5, which explains how the dataset was prepared in order to be used for the research. The next step is the exploration of the data, which is discussed in chapter 6. Chapter 7 builds upon that, giving extensive details to the approach of the data analysis performed. It can be seen as the core of this thesis. This is then followed by results and their discussion in chapter 8. A small validation of a component of the results is performed in chapter 9. The thesis concludes in chapter 10, and sends the reader into the future again with recommendations in chapter 11.

## 1.4 Case Study

This thesis works with data on the vessel traffic at sea, off the coast of Rotterdam. This area is controlled by the Hook of Holland Vessel Traffic Services Centre. Figure 1.1 shows a map of this area, with the structure of the Traffic Separation Scheme (TSS) included. A TSS is a network of lanes in the sea, to regulate the flow of traffic in a similar fashion as road traffic lanes. The black lines indicate the structure of the TSS , the large black arrows indicate the regulated direction of travel, and the enclosed anchor symbols indicate the anchorage areas. Between the main East-going and West-going lanes there is a deep draught channel through which vessels with deep draught must travel, with buffer zones around it to separate the deep channel from general traffic. Vessels that are not constrained by draught may not pass through the deep channel, but they may cross it. In the figure, also key places are indicated to understand the map. The most important is Hook of Holland, which is the origin or destination of most traffic coming through this TSS , since the TSS is built around the traffic going in and out of Rotterdam harbour However, other traffic does also cross this Rotterdam-traffic at the junctions. Also in this figure, the border is given of the area considered, based on the coverage area of the Automatic Identification System (AIS) - a transponder system to send positional, identification and other data.

## 1.5 Definitions

In order to facilitate the reader in the rest of this thesis, this section gives an overview of the definitions used in this thesis, with the help of an illustration. In fig. 1.2, an

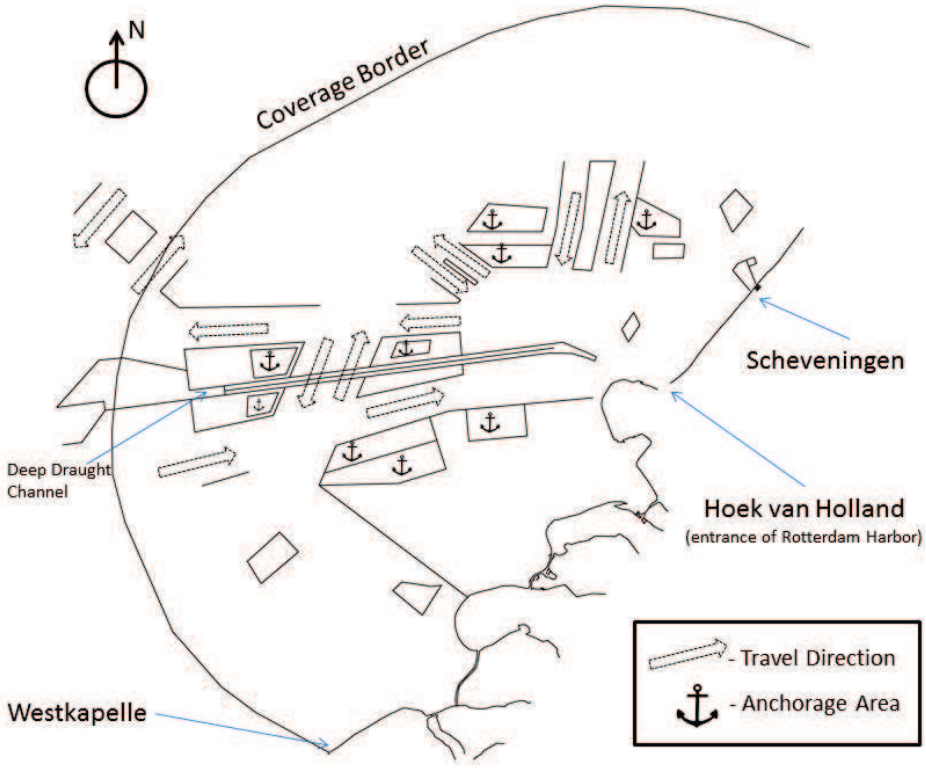



**Figure 1.1:** A map of the area considered as case study in this thesis. The lines show the structure of the TSS .

illustration is a zoomed in, simplified view of the TSS . The black lines indicate the structure, and the large black arrows indicate the regulated direction of travel. The deep draught channel and its buffer zones are also depicted for context.
To explain the definitions used, the focus is first on the blue ship ( ). Generally speaking, vessels are on their way to a specific 'destination'. This could for example be a dock in the Port of Rotterdam. To get there, a path is planned by the captain of the ship, mostly based on efficiency. This 'planned path' is indicated in the figure with a dashed line ( ). The last point of this planned path within the scope of the map and the time frame is called the 'goal' of a ship: . This can also be an area in the middle of the map, such as the anchoring area (the goal of the red ship). Vessels generally go from waypoint to waypoint on their way to their destination. The waypoints within the scope of the map and time frame are referred to as 'sub-goals' ( ). All the components (destination, goal, sub-goals, planned path) together are referred to as the 'intent' of the vessel, which is the plan that is principally in the mind of the vessel captain.
When another vessel (or object), such as the red ship, affects the behaviour of a vessel, this is called an 'interaction'. Interactions, as well as environmental conditions, cause

a vessel to deviate from the planned path, resulting in an 'adapted plan' (⇢). If the red ship communicates its plan to the blue ship, the blue ship can adapt its plan in a better way. Finally, the execution of the adapted plan results in an 'actual path' (→■), which depends on the skills of the crew, the vessels, and the environment.
All the above definitions have been explained in terms of positions. Please note that intent and paths also contain other components such as time, velocities and rudder setting. An overview of these basic definitions is given in table 1.1; further definitions will be stated where necessary.

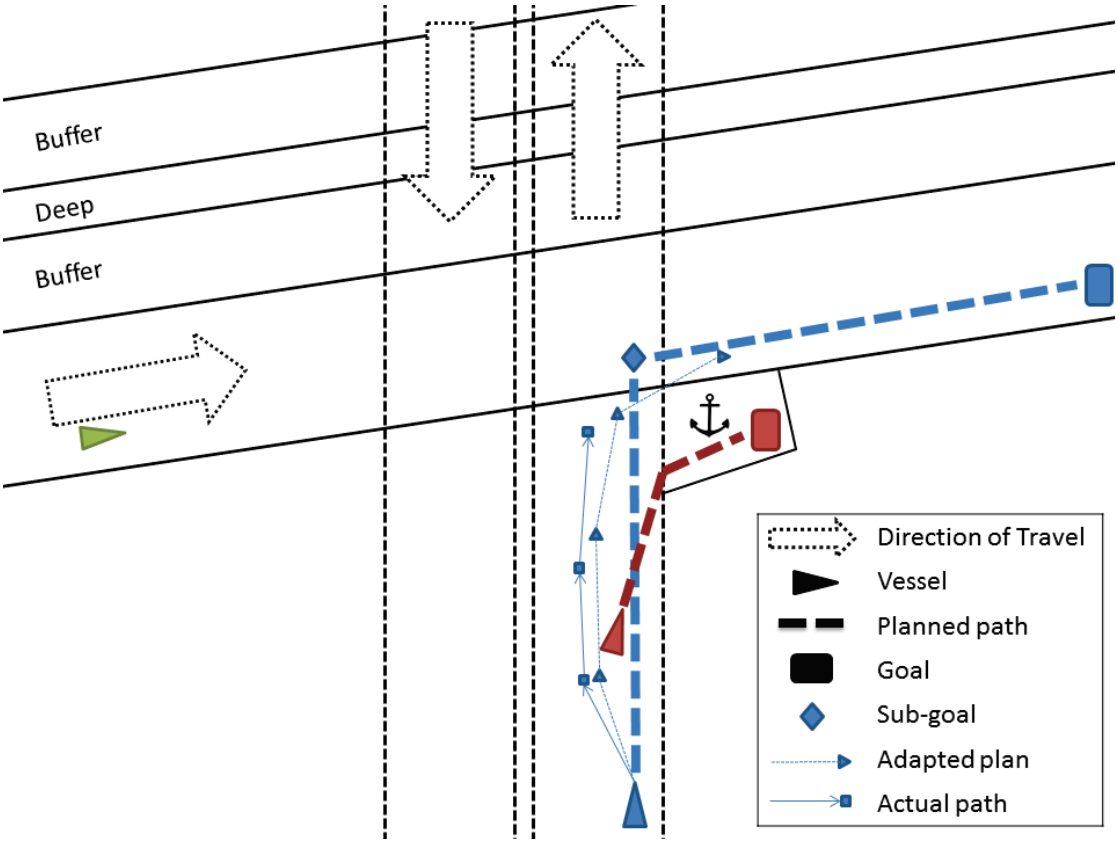



**Figure 1.2:** An illustration of several concepts and definitions as used in this document (see section 1.5 for explanation)

| **Term** | **Definition** | **Mark** (fig. 1.2) |
|---|---|---|
| Destination | Final aim of vessel | - |
| Planned path | Path planned to destination by captain of ship |  |
| Goal | Last point of planned path within map scope |  |
| Sub-goal | Waypoint on the planned path |  |
| Intent | Plan in mind of captain (includes above terms) | - |
| Interaction | When a vessel plan is affected by another vessel or object | - |
| Adapted plan | Intended deviation from planned path |  |
| Actual path | Result of executing the adapted plan |  |

**Table 1.1:** Definitions used in this thesis.

# Chapter 2

# Literature Survey

In this chapter, various approaches and concepts from literature are detailed and their contributions to the research are assessed. Section 2.1 describes the literature that is mostly related to risk, followed by section 2.2 which describes literature that is mostly related to prediction. Finally, section 2.3 brings all the literature together in a discussion. The aspects of prediction and risk are interlinked in many ways. For example, the prediction uncertainty can be an indicator of risk, while the other way around the level of risk can influence behaviours or interactions which influence future states. In this chapter, risk and prediction are principally treated as separate concepts, but the relevant linkages between them will also be highlighted.

## 2.1 Risk

The occurrence of a collision, and so also the risk of a collision occurring, has multiple components. This section starts with literature that is strictly related to spatio-temporal factors, and ends with comprehensive measures of collision risk.

### 2.1.1 Spatio-temporal overlap

One aspect of collision risk is (near-) spatio-temporal overlap of two or more objects, or the probability thereof. In the field of collision detection, this is very closely related to spatio-temporal prediction, since the risk is measured based on the amount of overlap (expected). Classic measures in this field are the Closest Point of Approach (CPA) and the ship domain. The CPA is the point where the distance between two moving objects will be the closest. Distance to Closest Point of Approach (DCPA) is the distance to that point, while Time to Closest Point of Approach (TCPA) is the time before the objects will be at CPA . The formulae for these parameters are detailed in Lenart (1999) A ship domain (originally introduced by Fujii et al. (1974)) is a geographical region around a vessel; when entered by another vessel, it is considered a (near-) collision.

#### Montewka et al

Montewka et al. (2010) model the probability of vessel collisions in a certain area using a simulator and a novel measure of risk. This measure of risk is called the Medium Distance To Collision (MDTC), which is defined as the distance below which two vessels cannot perform any manoeuvre to avoid collision. The value of this distance was determined experimentally using a hydrodynamic model of the vessels and a Monte Carlo simulation, simulating many different encounters between ships by varying the velocity, position, and course. For every meeting an evasion manoeuvre was performed. If it resulted in a close shave, it was stored as the MDTC for that initialized setting. The application chosen by the authors (detecting collisions in a simulator) is not directly relevant for this thesis, but the risk measure is: it can be used as a component of how much risk an encounter has. The smaller the MDTC , the safer an encounter between two ships. This is only dependent on the length of the ships, their type, and their relative speeds, positions and courses.

#### Li and Pang

Li and Pang (2013) assess the risk of collision using a Dempster-Shafer theory approach. They define three membership functions -dependent risk indicators- which depend on DCPA , TCPA and relative distance. The uncertainty of these input parameters is then used to ultimately derive the joint basic probability assignment as a measure of collision risk. The relevance of this approach is that the uncertainty of measurements and

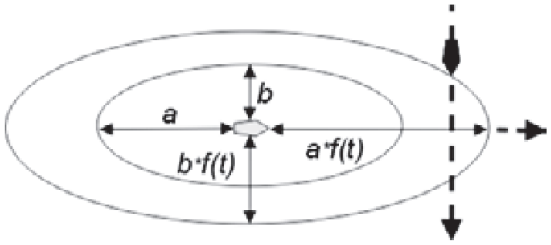



**Figure 2.1:** Approach factor $f$ in an elliptical ship domain with dimensions $a$ and $b$ Szlapczynski (2006).

parameters are explicitly included in the degree of risk. Unfortunately, the parameters used to assess the risk ( DCPA and TCPA ) are strongly related to a close encounter, and not directly related to a longer time horizon risk assessment. Nevertheless, the principles applied can possibly be adapted to account for measurement and prediction uncertainty in the assessment of a future traffic state.

### Szlapczynski

The concept of a ship domain, which is the area around a vessel that is not to be intruded by another object, is commonly applied in the maritime collision detection field. It is commonly used to define an encounter or a collision. Szlapczynski (2006) proposes an alternative to the DCPA risk measure, by adapting the ship domain into a measure of risk. He defines an approach factor, which indicates how far a ship domain must be scaled up from a baseline size in order to touch another ship. The larger the approach factor, the further away the other vessel is. In the case of a circle-shaped ship domain, the approach factor is simply directly proportional to the range (relative to the domain radius). However in the case of any other shape, the approach factor depends on the constrained parameter. In fig. 2.1 an elliptical ship domain with parameters $a$ and $b$ is scaled up to touch another ship. In this case, parameter $a$ is constrained. There are many shapes of ship domains (Wang et al., 2009) but this approach factor can account for any of them.
The promising part of this risk measure is that it is not restricted to close encounters, and the concept is quite simple to apply. Unfortunately this risk measure only takes into account relative distance, not relative speed, bearing or orientation. However, it may be possible to expand this measure by including the time change of the measure (for speed or change in bearing). That would leave relative orientation still unaccounted for.

### Kuwata et al

A common risk measure method in collision avoidance for robots is that of velocity obstacles. However, this approach is not very common in the maritime field. Kuwata

et al. (2014) have applied this concept in order to plan collision avoiding paths for autonomous vessels. The basic idea of a velocity obstacle is that it contains all the possible velocity vectors that will (likely) result in a collision with a target object. This concept is a potential risk indicator, that can also take into account velocity obstacles from multiple vessels. Consider the physically feasible velocity space of a single vessel. If a large part of this space is occupied by velocity obstacles, the options of a vessel are limited. If its speed lies within a velocity obstacle, there is a risk of collision. This concept can also be used to predict vessel behaviour, assuming that the ship-master has knowledge of his options.

#### 2.1.1.1 Expert-based Approaches

Several approaches use expert opinion to define a measure for collision risk. This is implicitly related to conflict complexity, since the measure of risk is defined by the judgement of experts. This thus includes their estimation of the level of difficulty of conflict resolution. The general disadvantage of these methods is that they are especially suitable for close encounters. If a future traffic state contains much uncertainty, the values for these risk indicators may be difficult to determine.

**Zhang et al**

Zhang et al. (2015) designed a Vessel Conflict Ranking Operator (VCRO), which includes the factors distance, rate of distance change, and relative orientation. It is defined by a custom function (which depends on the previously mentioned factors), of which the parameters are estimated based on expert assessment of a set of typical two-vessel encounter scenarios. The value of the VCRO is a conflict severity *rank* indicator, so it only indicates if one encounter is more severe than another. It does not indicate a real value of risk.

**Bukhari et al**

Bukhari et al. (2013) have devised a method that does have a real value of risk as an output. Using DCPA , TCPA and rate of bearing change as inputs, they built a fuzzy inference system which links the risk of these parameters together. Each of the parameters was given a set of linguistic values (e.g. from 'Positive big' to 'Negative big'), from which membership functions were built with expert opinion. Subsequently the combinations of the variables were used as scenarios, each of which was given a linguistic value for risk. For this degree of risk also a membership function was built. This method can be very promising if expert opinion can be used properly. It can be expanded further to include factors such as weather, environment and crew skill, because experts can also evaluate the influence of these factors on scenarios using linguistic terms.

#### Goerlandt et al

Goerlandt et al. (2015) take it one step further than Bukhari et al. (2013). They start by investigating which parameters are commonly used to assess an encounter situation, and how relevant each of these measures are. They also introduce their own parameters. In determining the relevance, they distinguish between different COLREG classifications of an encounter: overtaking, crossing, and head-on, and a new classification: unexpected turn. By expert elicitation the relevance of each measure is determined. The resulting parameters included are seven proximity indicators (e.g. range and orientation), projected positions (e.g. DCPA ), booleans (e.g. if a ship is turning), static parameters (e.g. ship length), and environment (visibility and time of day). Interestingly the wave conditions and ship type were discarded due to the low relative importance (as assessed by the consulted experts). The greater advantage of this method is that it also includes environmental factors, and again can be expanded to include human factors. The paper also includes an extensive analysis on which factors are truly relevant.

### 2.1.2 Traffic Complexity

Traffic complexity is an aspect of risk that is correlated to the workload of a VTSO and his/her capability to understand and resolve a situation. Therefore this covers a relevant aspect of risk. Consider the following cases. An evolving situation that has a high risk of collision, but is not complex, can be easily recognized and resolved by the VTSO , given that the VTSO can influence the situation. However an evolving situation with yet a low risk of collision, but is complex, has potential to develop risk, because the VTSO (and other parties) have more difficulty to evaluate and resolve the situation. In the maritime field this appears to be a novel concept, but in the air traffic field this concept is currently being studied extensively.

#### Wen et al

Wen et al. (2015) aim to evaluate a maritime traffic situation using a traffic complexity method. They are the first to introduce the concept of traffic complexity to the maritime domain. The authors define traffic complexity as an indicator of the degree of crowding and the risk of collision in a specific area. It includes two types of complexity: density complexity and conflict complexity. Density complexity is a function of relative distance between ships, the geographical environment, and the ship type. Conflict complexity consists of two components: angle complexity (function of relative angle) and the convergence complexity (function of relative motion). The angle complexity is determined based on research done by Montewka et al. (2010), which relates the MDTC risk measure to encounter angles (among others).
Though the authors include a ship type component in their density measure, in this paper all vessels are considered to be of the same type. Distinguishing between vessel types is considered future work by the authors. The added value of this paper in the

traffic complexity area, is that it already has incorporated much domain knowledge from the maritime traffic domain, which the following papers lack (as they originate from the air traffic domain).

### Delahaye and Puechmorel

Delahaye and Puechmorel (2000) propose two measures for traffic complexity: a geometrical indicator and 'topological entropy'. The geometrical indicator comprises density, degree of convergence, and sensitivity. These three aspects are depicted as dimensions in a 'complexity' coordinate system. The definitions set in this paper for density and convergence have been adapted by Wen et al. (2015). The sensitivity component is a measure of how sensitive the relative distance between two aircraft is to small changes in speed or heading. Topological entropy is a traffic complexity measure that addresses overall flow complexity. This measure is aimed at traffic flows and is thus especially suitable for evaluation of the complexity of a traffic scheme, not for medium-term prediction. However, the geometrical indicator *is*, and hence may prove useful. The sensitivity complexity can potentially be used to account for uncertainties in prediction.

### Rahman et al

A different kind of traffic complexity measure by Rahman et al. (2012) is the solution-space based measure. When an aircraft encounters (one or more) other aircraft, there is a limited number of safe manoeuvre possibilities to avoid collision. Complexity is defined as the number of safe manoeuvre possibilities divided by the total number of manoeuvre possibilities. This is depicted in fig. 2.2 by a ring of speed and heading combinations, where the unsafe combinations are shaded grey. The complexity is then the non-shaded area divided by the total area. The contribution of this complexity measure is that it closely captures the conflict resolution aspect of complexity, and accounts for multiple aircraft simultaneously. Note that this approach has similarities with Kuwata et al. (2014) (discussed in section 2.1.1).

### Wang et al

Wang et al. (2015) developed a graph-based framework for traffic complexity. In a traffic scene, each aircraft is modelled as a node, and two nodes are connected by an edge, if their relative distance is lower than a set threshold value. From this framework, based on how the graph evolves, different measures of complexity are derived. 'Degree' signifies the amount of edges connected to a node, 'connection rate' is the amount of edges that appear or disappear over time, 'clustering coefficient' sums the degrees of the nodes connected to a node, the 'importance' of a node is its degree relative to the average degree, and finally the 'network structure entropy' indicates the homogeneity of the degrees of the nodes. This last measure is based on the fact that if a set of aircraft

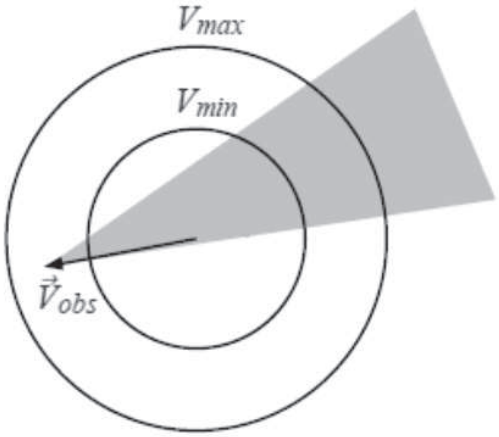



**Figure 2.2:** Ring of possible speed and heading combinations (between $V_min$ and $V_max$). $V_obs$ is the observed speed vector of another aircraft, which induces the grey area indicates unsafe possibilities. Adapted from Rahman et al. (2012).

is clearly more important, it is easier for a controller to understand which aircraft needs the most attention.
The value of this approach is that it can present a generic traffic situation in a simple manner, can assess the complexity at a local scale and a macro-scale. Unfortunately the method only considers proximity of aircraft, implicitly including speed by the evolving connection rate. However, there is potential to build on this framework to include different measures such as velocity, and possibly weighting factors.

### 2.1.3 Socio-Technical Systems

Another viewpoint of risk is to approach the system as a whole, i.e. a socio-technical system. Here, the different players (humans, environment and objects) are commonly called 'agents' if they can influence the environment or other agents. The risk emerges from the interactions between these agents. A large portion of these interactions involves the observing or sharing of information from one agent to another.

#### Vanek et al

Vaněk et al. (2013) deploy an agent-based model to capture the complex dynamics of the maritime transportation system in the Indian Ocean. Their focus is on modelling the effects and risks of piracy in that region. In order to do this, they devise individual-ship models for three types of vessels: merchants, pirates, and military. In each of

these models, a vessel switches between different activities, depending on the parameters, interactions with the environment and other agents. For example, a merchant ship switches between 'docking', 'route-planning', and 'cruising' depending on the time, and subsequently may switch to 'hijacked' following a sequence of events (such as not observing a pirate vessel, or not requesting help from a military vessel, or a military vessel being out of range or too late). Though the piracy aspects of this approach may not be relevant for this thesis, the modelling approach may prove useful, since it may capture the interactions between vessels and other agents, and the risk that emerges from that. Also it could be used to explicitly model the element of risk that environmental factors bring (such as non-communication).

### Lefèvre et al

Lefèvre et al. (2012) link into part of a social-technical system, by comparing the expected behaviour of a driver to the inferred intent of that driver. This is done at a road intersection. Based on observations of the states of a car in discrete time, the manoeuvre intention (adapted plan), and expected behaviour (to stop or not) of its driver are estimated using a Dynamic Bayesian Network. This means that the probabilities of arcs change over time and are influenced by other drivers. This model captures effects such as a driver intending a manoeuvre based on the states of other drivers, which may or may not be an objectively expected manoeuvre, and which may or may not be executed correctly. This type of approach blends the line between risk and prediction by finding a probability that an object will not perform the expected behaviour The context, however, lacks overlap with the maritime domain, since the manoeuvre intention is discretised into four options (the four possible directions from an intersection). A maritime TSS is much closer to a continuous space.

### Blom and Sharpanskykh

Blom and Sharpanskykh (2015) have developed a mathematical framework to model Situation Awareness (SA) in socio-technical systems. Every agent is represented by a state vector. This state vector comprises three main state types. These are the actual state of the own agent, self-awareness (the state of the own agent according to the own agent), and situation awareness (the state of another agent according to the own agent). This can be done recursively: an agent A can be aware of the situation awareness of agent B about the state of agent C. The situation awareness states are updated by observation, messaging, and interpretation events.
The relevance of this model is that it can be used for the human and environmental causes to collision risk. Since human factors such as communication have been identified as a common cause for developing collision risk, modelling this may prove useful in both prediction as well as risk assessment. However, direct information about these states will likely not be available, they will need to be inferred from the data, perhaps in a manner like Lefèvre et al. (2012).

**Bouarfa et al**

Bouarfa et al. (2013) deploy an agent-based model to simulate a specific socio-technical system, in order to identify emergent behaviour affecting the safety of the system. The case studied is an area on an airport where a taxiing route crosses a runway. The human actors -agents- involved are pilots in the involved aircraft and a runway traffic controller. The technical agents considered are the two aircraft and the Air Traffic Control (ATC) system. The remaining 'agent' is operational conditions, which includes the runway configuration and visibility conditions. After defining the simulation model, Monte Carlo simulations have been performed to obtain the emergent system behaviour, from which a safety assessment is made.
The all-encompassing approach of this piece is of interest in this research. The technical aspects (e.g. aircraft models) as well as the human cognition aspects and the environmental conditions are all taken into account. This is done not only per component, but rather also includes the interactions between the components and so the behaviour of the system as a whole is modelled This is its major advantage. An important requirement for such an approach, however, is to have detailed sub-models and knowledge of the individual behaviours of the agents considered. This requirement may be fulfilled further down the line in maritime research.

## 2.2 Prediction

In traffic literature, prediction is generally done based on a model for vehicle dynamics or a planned path, or based on historical data. Depending on the amount and the kind of information available, different methods are used. This chapter focuses on the approaches found that contribute to this thesis. Goal, intent (sub-goals), trajectories and future states (position/velocity) are spatio-temporal information types that generally are predicted.

### 2.2.1 Model-based Prediction

Model-based prediction methods are -unlike data-driven methods- not directly dependent on (historical) data. They are built based on knowledge and assumptions about reality, not depending on historical data.

#### 2.2.1.1 Medium Term Conflict Detection

Medium Term Conflict Detection (MTCD) -not to be confused with Minimum Distance To Collision (MDTC) - is a large field in the air traffic management domain that has some similarity with the problem at hand. It is aimed at detecting spatio-temporal overlap by predicting the future states of aircraft. This links one risk aspect (spatio-temporal overlap) and prediction into one approach. However the general disadvantage

is that these methods are based on knowledge of a flight plan, which is information that is not available in the maritime domain. A vessel generally does have a *planned path* (e.g. sailing plan), but this information is not shared.

#### Prandini and Hu

Prandini and Hu (2016) define air traffic complexity as the limitation of manoeuvrability of an aircraft. This complexity is said to increase as aircraft come within a certain range of each other, thus their definition is closely related to the density of the traffic. They predict this complexity on a mid-term time horizon. Their approach is quite similar to many MTCD methods applied in the air traffic domain. Each aircraft has a flight plan, and will deviate from it due to pilot error, wind and preventive avoidance manoeuvres by the pilot. This paper contributes by linking mid-term prediction to complexity evaluation, even though the definition of complexity may not be very useful.

#### Althoff et al

A concept commonly applied in this area is 'reachable sets'. This is the total set of possible states that an object can reach within a given amount of time. Althoff et al. (2009) evaluate the safety of a planned path of an autonomous car by estimating the stochastically reachable sets of the own car and other traffic participants in a discrete state space. This discrete state space is the basis for a Markov chain model, where transition probabilities between cells are determined based on the current acceleration, speed, and position of an object. The future position is then simulated multiple times to obtain a picture of the stochastically reachable sets. It is taken into account that it is unknown what turn another traffic participant will make at an intersection, though road geometry is used as a constraint. The relevance of this method is in its simple approach in modelling the vehicles: only the acceleration, speed and position are modelled Also, it does not assume a planned path of another vehicle to be known.

### 2.2.1.2 Maneuver Recognition

#### Houenou et al

Houenou et al. (2013) predict the trajectories of cars in an area, based on movement measurements. This method is three-fold: first the type of manoeuvre that the car is making is determined, then the possible future trajectories are determined, from which the most likely trajectory is selected. The latter parts of this approach are not relevant for this thesis due to their short-term nature, and road traffic related dynamics. However, the first part is interesting, because detecting what type of manoeuvre a vessel is making can be of value for the prediction of its trajectory. For example, if one can detect that the blue ship in fig. 1.2 is initiating an overtaking manoeuvre, or even that the next

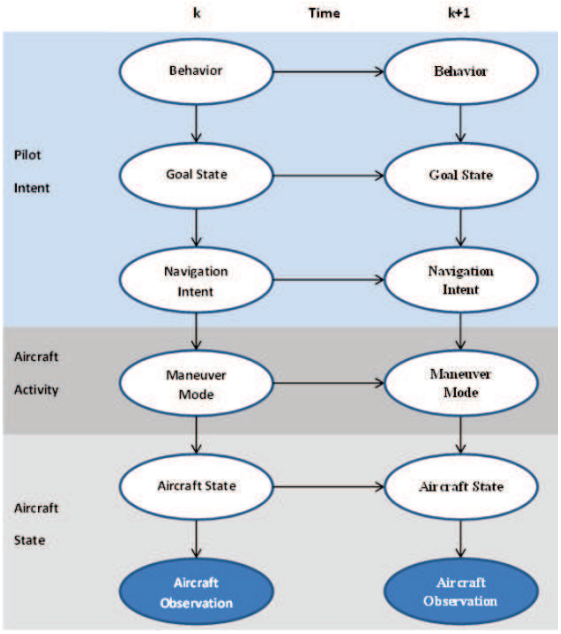



**Figure 2.3:** Hierarchical intent model by Lowe and How (2015)

manoeuvre is a right turn, it would be helpful in inferring the adapted plan or planned path. The authors here derive the manoeuvre mode (maintain lane, change lane, or turn) by detecting the deviation from the centreline

### Lowe and How

Lowe and How (2015) focus on prediction of aircraft in the uncontrolled airspace, like Lancia et al. (2014). However, these authors focus on short-term prediction, and especially concentrate on inferring the intent of a pilot. They have designed a hierarchical intent model shown in fig. 2.3. Here the observation consists of the measurements made, the state is the actual state, the manoeuvre mode is a single mode (e.g. constant velocity or coordinated turn). Then, navigation intent is the sequence of navigation states the pilot intends to take to reach its goal state - the final state the pilot intends to reach. Finally the behaviour is classified as complying or non-complying (i.e. manoeuvring as expected towards a goal or not).
For all layers of the hierarchical intent model, estimations are made by predicting the next time step using the current estimation of layers above and below, and subsequently evaluating the previous prediction, in order to adjust the estimation. The value of this hierarchical intent model is that it addresses all the layers of the problem: from the most unknown behaviour classification, to the observation state. It explicitly models the relationships between these layers.

### 2.2.2 Data-driven Prediction

There is a lack of information when it comes to a planned path. Vessels do not have the obligation of declaring a sailing plan, like aircraft pilots have to file a flight plan. This difficulty in modelling a ship's intended path can possibly be solved with the use of historical data. An important thing to note is that these methods implicitly include environmental and behavioural effects through the historical patterns, though these elements may not be explicitly taken into account.

#### 2.2.2.1 Trajectory Prediction

Trajectory prediction predicts the planned path of an object, with the use of historical data, focusing on the trajectory that the object will take, as opposed to predicting where the object is headed in the end.

**Lancia et al**

A research area that has received much less attention is that of general aviation. Lancia et al. (2014) venture to predict the future trajectories of light aircraft in the uncontrolled airspace. However, different from the common field of MTCD (section 2.2.1.1), in general aviation the availability of a flight plan is generally low. This aspect makes the problem closer to the prediction problem in the maritime domain. The authors use historical data to predict the trajectory of an aircraft. These historical trajectories are first broken up into line segments and 'turning points'. These turning points are then clustered, to obtain nominal trajectories. Based on the entire registered historical path of an aircraft, its future trajectory is estimated using a Particle Filter (PF). This kind of approach is typical for motion pattern prediction. This approach also takes into account the possibility that a pilot may change plans during flight.

**Pallotta et al**

Pallotta et al. (2013) propose an unsupervised and incremental learning methodology for vessel motion pattern prediction, which they have named Traffic Route Extraction and Anomaly Detection (TREAD). They apply a 'vectorial' representation of traffic: trajectories are a collection of waypoints with straight paths in between for the sake of computational cost. They use the Density Based Spatial Clustering of Applications with Noise (DBSCAN) methodology (Ester et al., 1996) to cluster waypoints, because it does not require the number of clusters a priori, can find arbitrarily shaped clusters, and can filter outliers. Special clusters are stationary waypoints (typical anchorage/harbours), and track initiation or termination, called entry/exit points respectively (e.g. map edge, harbour). Routes are defined as a pair of an entry point and an exit point.
Every new ship entering the scene is compared with the existing set of routes, and

if matching, is added to a list of vessels for a route, and to lists of vessels for each waypoint it passes. If it does not match, a new route is created. If enough ships transit a route, the route is activated (usable for prediction). Kernel Density Estimation (KDE) (non-parametric) is used to estimate the Probability Density Function (PDF) of routes. These PDF 's are used to classify routes. Route prediction is done based on the transited trajectory of the ship, in a similar manner to Hidden Markov Models (HMMs), distinguishing between actual states and observations. The prediction is expressed as probabilities that the vessel will follow an activated route. An illustration of what this looks like is given in fig. 2.4.
The advantages of this approach are: the context is the maritime environment, the environment is represented in a semi-continuous manner (since waypoints and routes are added as necessary), and the planned path and goal/destination are explicitly derived. The method does not include time or seasonality, however, and neither can it adapt to changing patterns by, for example, removing routes.

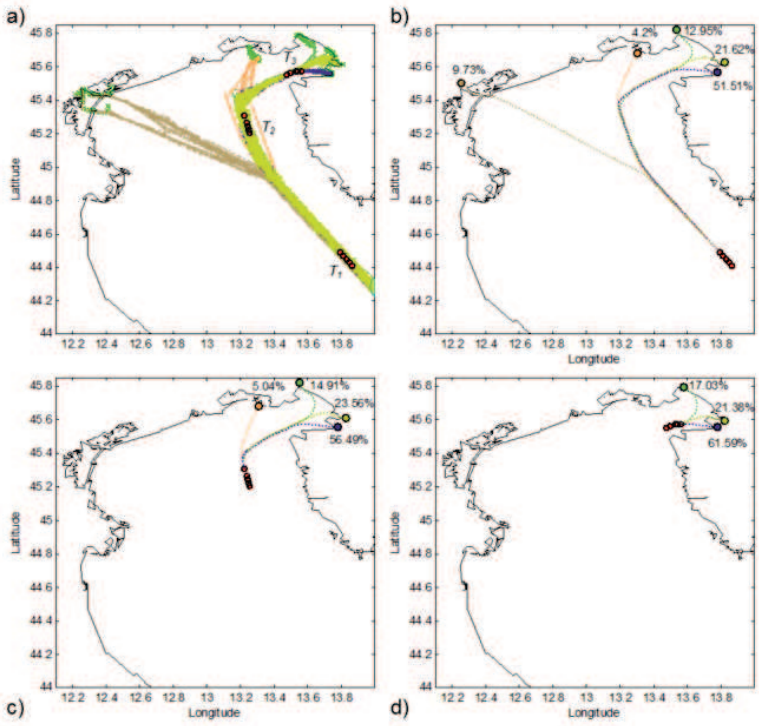



**Figure 2.4:** Illustration of trajectory prediction by Pallotta et al. (2013). Prediction is expressed as the probability that the vessel will follow a certain route. Plot a summarizes the following three plots. The plots b,c,d follow chronologically, indicating time. The colours indicate different routes, and the red dots indicate a series of measurements of a vessel.

#### Ristic et al

Ristic et al. (2008) take a statistical approach: they model vessel motion patterns as (Gaussian) PDF 's per location, not by path. These PDF 's are estimated with an Adaptive Kernel Density Estimator (A-KDE) based on a historical dataset. Subsequently prediction is performed by using a Particle Filter (PF) on the possible proceeding sequences of positions from the current position. Interestingly, this method only uses recorded position to predict future positions. Perhaps it is expandable to other states. If the assumption of Gaussian distributions is valid, then this becomes a very simple but effective way of estimating the future position of a vessel. In case the distributions are not Gaussian, another distribution or even another estimator may be possible.

#### Zandipour et al

The series of papers (Zandipour et al., 2008; Rhodes et al., 2007; Bomberger et al., 2006) aim to predict the future position of a vessel by learning historical data incrementally using the Neural Associative Incremental Learning (NAIL) algorithm (Rhodes, 2007), which is built to learn taxonomic relationships between concepts or objects. The correlation learned is between the current vessel state (position and velocity) and its future position at a single time horizon.
Promising aspects of this approach are that it learns the frequent patterns quickly, and has the capacity to unlearn less frequent patterns through weight decay. This means that it will adapt to changes in the environment or the paths, and different learning sets can be adapted for data sets of different seasons.
One lacking point is that the prediction is done for a single time horizon. The method would have to be expanded to predict a trajectory. Also, the method only relates the current state vector (position and velocity) with the future position. Previous states are not taken into account. Also future speed is not predicted, and other factors such as environment, time or vessel type are not taken into account. Maybe the method can be expanded to include these factors as well.

### 2.2.2.2 Goal and Route Prediction

Another type of approach to predicting the planned path is by first estimating the goal of a vessel, to subsequently use this estimate to derive the planned path from it.

#### Simmons et al & Krumm et al

Simmons et al. (2006) predict the goal of a driver using a Hidden Markov Model (HMM) on the historical trajectories of that driver. The map is represented by a graph, where each vertex is an intersection, and each (direction dependent) edge is a road. The hidden states are the current position and additional factors such as time or day of the week.

Every observation is used to estimate the hidden transitions between states. Travelling time is also measured but not applied further.
Krumm and Horvitz (2006) developed the so-called 'predestination' approach. It uses a grid representation of a city, each cell possibly the goal of a driver. Based on the dominant ground type of a cell (e.g. pasture, industry, or water) and a user survey, an initial probability -of being a goal- is assigned to the cell. Then, as a driver travels from cell to cell, the probability of all the cells are updated based on how likely it is that the driver will end in a cell. This is based on the assumption that a driver will take an efficient route (distance-wise), and that the driver will complete his travels in a 'normal' amount of time (based on a distribution).
In (Krumm et al., 2013) the research is continued by deriving the planned path from the goal. However, in this research the same graph representation as in Simmons et al. (2006) is used, so a driver goal becomes the last road intersection that it will pass. Subsequently, the planned path is quite simply derived by assigning each road segment the probability of being traversed: the sum of the probabilities of all the nodes downstream from the road segment. An illustration of this is given in fig. 2.5.
A promising aspect of these approaches is that they can be extended with many states. The disadvantage is that they are built on the simplistic, discrete representation of a road network (while the marine environment is much more continuous). The 'predestination' method is more detailed than Simmons et al.' method, and presents all probabilities of candidate goals for each time step, while Simmons et al.' method may be more simple to expand.

### Best and Fitch

Best and Fitch (2015) estimate the goal and planned path for pedestrians in a continuous environment. The authors basically take the same approach as Krumm et al. (2013), except that initially all goals are assumed to have equal probability (a priori knowledge not needed). However, due to the continuous nature of the environment, simple summation of downstream goal probabilities is not possible, which is why the authors perform a number of Monte Carlo simulations of goals. For each sampled goal a trajectory is drawn with random perturbations for each time step ahead (i.e. a 'random walk' biased towards the sampled goal). The benefit of this method lies in that it does not require any a priori knowledge of the area, and that it is applicable in a continuous environment. So though the method performs the same task, it is actually not a data driven approach, since it only depends on the historical trajectory of an object only, not on historical trajectory database.

### Vasquez et al

Vasquez et al. (2009) learn the motion patterns of pedestrians and vehicles using a Growing Hidden Markov Model (GHMM) approach in combination with an Instantaneous Topological Map (ITM). The ITM approach is promising because it can represent a

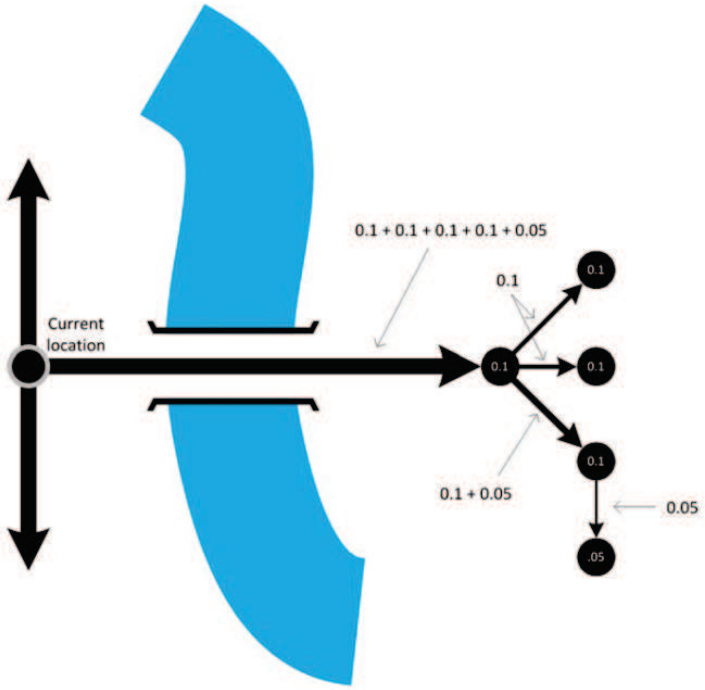



**Figure 2.5:** The probability of each road segment (edges) is the sum of the probabilities of the downstream destinations (nodes). (Krumm et al., 2013)

continuous state space of a model based on motion patterns in a semi-continuous manner since nodes of the graph are added and removed based on incrementally added trajectories. Areas with heavy traffic will have more nodes than areas with scarce traffic. This is similar to the maritime environment. The GHMM approach links the prediction to the ITM. It models the current state, and the intended state (goal) of an object. The result is a directed graph (in ITM structure) with all possible ways to get from the current state to the intended state.
This approach contributes to this thesis because the intended path can be derived from the goal, current state and the past trajectory. Another useful property of this approach is that the state space can be expanded (e.g. adding velocity, time, or even abstract states such as 'manoeuvre mode').

### Ikeda et al

Ikeda et al. (2013) are a special case of goal prediction. Their method predicts long-term pedestrian behaviour based on the 'sub-goal' concept. A sub-goal is a point towards which a pedestrian generally heads when it comes into view. See fig. 2.7 for an illustration. It has much similarity with the waypoint concept used in path-finding. In the light of this thesis, it is related to the sub-goal (◆) in fig. 1.2. The sub-goals are derived

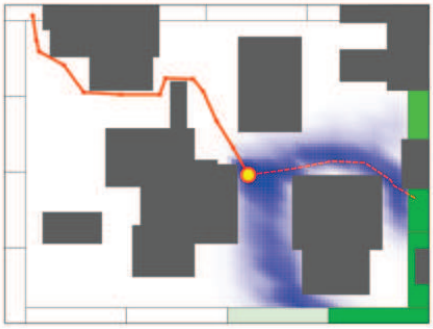


**Figure 2.6:** The pedestrian (orange) has walked a red trajectory. His goal (green) and planned path (blue) are predicted. The darker the color, the higher the likelihood of the goal or planned path. (Best and Fitch, 2015)

by assessing the direction in which pedestrians travel (from each grid cell), and checking where these pedestrian flows overlap each other. Then the behaviour of a pedestrian at each point in time can be modelled as a sequence of movements towards one of the sub-goals In that way a sub-goal sequence is derived to describe the past trajectory of a pedestrian. Transition probabilities are derived between these sub-goals. These transition probabilities, together with the current position, velocity, and preferred velocity are used to predict the future trajectory.
The promising part of this approach is that it has much similarity with the behaviour of vessels. Vessels commonly aim for waypoints (sub-goals) and seem to have a preferred velocity (the most efficient velocity). Also, the behaviour is modelled in continuous space; though the historical data is first set up in a grid layout, the sub-goals and trajectories are defined in continuous space.

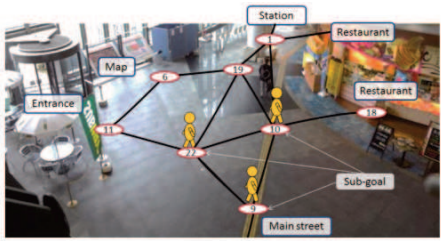



**Figure 2.7:** Pedestrians move from sub-goal to sub-goal, aiming for the next sub-goal as soon as it becomes visible. (Ikeda et al., 2013)

### 2.2.3 Interactions

When predictions are done based on historical trajectories (section 2.2.2), more detailed information is lost due to averaging or clustering mechanisms. Such predictions can

at best predict the 'planned path'. In this section, methods that model interactions between a vessel and another vessel or a vessel and the environment are discussed. This could help in deriving the 'adapted plan' from the planned path.

### Xu et al

Xu et al. (2015) simulate vessel traffic flow in an inland waterway. A useful sub-model of their simulator is the 'vessel behaviour model'. This model treats a vessel as an agent with decision-making and communication capabilities, aimed to capture the collision avoiding interactions between vessels when one vessel sails behind another with a higher velocity. This results in an overtaking or following behaviour, depending on the risk estimated by the agent, and the communication with the other agent. Unfortunately this covers only one type of interaction (approach from behind), disregarding other approach angles. However this type of model could possibly be used to model the interactions between the vessels in a simple manner.

### Chauvin and Lardjane

Chauvin and Lardjane (2008) investigate the interaction behaviour of ferries that cross a main traffic lane in the Dover Strait. The purpose of the model is to determine which vessel will take action first, and on which side of the other vessel it will pass. Based on interaction statistical data, a logistic regression model is found to be the best fitting model to the data. This model depends on the types of two interacting ships, the behaviour of the other ship, as well as their relative encounter parameters such as speeds, angles and CPA measures. The value of this approach is that it uses encounter parameters to predict what will happen, rather than estimate the risk (section 2.1.1). This captures the human aspect in how the vessels respond to an encounter.

### Helbing and Molnár

Helbing and Molnár (1995) propose the social force model in order to model the behaviour of pedestrians. This model has since been widely applied in the prediction of pedestrian behaviour The movement of a pedestrian is firstly determined by his goal position. To get there he/she has a planned path of straight edges, similar to the sub-goal concept of Ikeda et al. (2013). This results in a desired direction. The pedestrian wants to do this in a comfortable manner, resulting in a corresponding desired velocity. If not travelling at this speed, this results in an acceleration force in the desired direction.
On the way to the goal the pedestrian encounters several repulsive and attractive forces. Other pedestrians have a repulsive effect on the pedestrian, though some have an attractive force (e.g. friends). Walls or borders act repulsively. These forces are stronger within the line of sight of the pedestrian, and weaker behind him/her. The combination of all these (dynamic) forces results in a pedestrian motion. This method could even be

expanded to include weather influences by adding or adjusting forces. This approach can possibly capture the interactions between vessels.

#### Elfring et al

Elfring et al. (2014) combine the GHMM prediction approach of Vasquez et al. (2009) with the social force model (Helbing and Molnár, 1995) to predict pedestrian motion more accurately. They first apply the former approach, and subsequently refine the prediction using the latter model. Their motivation is the lacking social forces model predictions, and the semi-discrete map of Vasquez et al. (2009) where a continuous representation is desirable. This is the most promising prediction method yet found, since it accounts for longer term prediction as well as for interactions with other objects. However, the interaction forces in the maritime environment may not be similar, nor the behaviour of pedestrians.

## 2.3 Discussion

The aim of this chapter is to summarize and interlink all of the presented literature. There are three main points of discussion in predicting collision risk on a medium-term time horizon within the maritime context. The first point is the causes for collision risk and how these are taken into account. The second is that the time horizon at which predictions need to be done. A larger time horizon causes a build-up of uncertainty. The third point is that in the context of this research, there is a lack of information about the intent of vessels and their communication. This section starts with the causes for collision (section 2.3.1), and subsequently summarizes and discusses the literature per group (sections 2.3.2 to 2.3.6) in the light of the above mentioned discussion points. Finally, the contributions of the different pieces of literature are put in an overview in section 2.3.7.

### 2.3.1 Collision Causes

The causes of a vessel collision can be roughly divided into four main groups. The first group is spatio-temporal: if two vessels are too close and heading towards each other, a collision may be inevitable. The second group is human factors: error, skill, communication and behaviour The third group is environmental causes: visibility, tides, sea state, wind and currents. The fourth group is linked to the vessel characteristics: its dimensions, manoeuvrability, cargo and equipment. Bear in mind that these groups are not independent but linked closely in various manners. Considering the fourth group, these factors are taken into account by nearly all methods, by way of classification.

### 2.3.2 Spatio-temporal risk factors

Methods that focus on the spatio-temporal problem treat vessels on the observed spatio-temporal level, which is very tangible. Abstract levels such as cognition and decision are not included. These methods are generally relatively easy to grasp and verify/validate.

#### 2.3.2.1 Two-vessel encounter

In the case of an encounter of two vessels, several measurements for risk have been proposed. Each of these methods use spatio-temporal measures related to proximity, speed and orientation as an input. MDTC (Montewka et al., 2010) indicates the closest distance that two vessels can come to each other, based on a manoeuvrability model of the vessels. The basis of the measure is very meaningful, because it aims at the actual resolvability of the situation. Using a manoeuvrability model makes the resulting risk measure very accurate, but difficult to extend to all vessels, due to the large variety of vessel characteristics, and the lacking information thereof.
Other approaches (Li and Pang, 2013; Bukhari et al., 2013) use DCPA and TCPA , which themselves are commonly used as risk indicators, as components for their risk measures. These methods are simple in their application, and account for uncertainty (in)directly. However, one disadvantage of the above risk measures is that they are all especially valid to measure the risk of close encounters, not for larger relative distances. The approach factor proposed by Szlapczynski (2006) does not have this limitation, and is also very simple to implement. With some additions to accommodate speed and bearing change, this measure may become very promising.
Also the velocity obstacles method implemented by Kuwata et al. (2014) is applicable to multi-vessel encounters, and is theoretically applicable to all distances. However, the practicality decreases with distance, due to increase of uncertainty.
The expert-based, fuzzy method by Goerlandt et al. (2015) is also focused on close encounters only, but includes visibility and time of day as input, which the above methods do not. The advantage of the expert-based approaches is that they can easily include human and environmental factors in their approach.

#### 2.3.2.2 Traffic Complexity

Collisions, however, do not always involve two vessels only; multiple vessels can be involved (and often are). To assess the risk of multi-ship encounters, the risk measures discussed in section 2.3.2.1 are generally summed for each pair of ships. This unfortunately does not capture the interdependence of the interactions of ships.
Traffic complexity measures generally do account for multi-vessel relationships. Besides that, they also are valuable because they cover other aspects of risk, namely the workload of the VTSO and the resolvability of encounters. Wen et al. (2015) seem to be the first to introduce this concept into the maritime domain. They capture two-vessel interactions by their conflict complexity (building on the concept of MDTC ), and the multi-vessel

interactions by density complexity. Delahaye and Puechmorel (2000), besides density complexity, contribute a measure of uncertainty by calculating how sensitive a situation is to speed and course changes. Rahman et al. (2012) base complexity on the possible manoeuvre that each aircraft is safely capable of. Wang et al. (2015) provide a graph-based framework to assess complexity based on the evolution of the graph.

### 2.3.3 Socio-technical system

Approaches aimed at the socio-technical system as a whole add another dimension to the multi-vessel problem. The behaviours and interactions of human and environmental agents that are involved are also modelled Agent-based models (Vaněk et al., 2013; Bouarfa et al., 2013) design a model for every agent (person, vessel/aircraft, environment) separately, as well as the relationships between them. From the network of these sub-models a system risk emerges.
Useful for the human and environmental factors is the mathematical SA framework by Blom and Sharpanskykh (2015). It can be used in a multi-agent system, to capture the awareness that each agent has of other agents and of itself. Therefore the intent, communication and interactions of the agents can be explicitly modelled
This same thinking is applied in part in Bouarfa et al. (2013), namely in the modelling of human behaviour This approach, however, also includes models for the technical systems, and their interactions with the human agents, and the environment. In contrast with Vaněk et al. (2013), the technical systems are detailed more. Lefèvre et al. (2012) focus on the human interaction and error by offsetting expected behaviour to actual behaviour of a human driver. The behaviour of a driver is predicted, and based on the expected behaviour relative to the actual behaviour, the risk is determined.

### 2.3.4 Model-based prediction

Prediction can be done based on a model for vessel dynamics. However this is limited to the short term. For the medium-term horizon, additional information on intent is necessary to maintain acceptable accuracy. This additional information can be on intended path (e.g. flight plan) or on intended manoeuvre.

#### 2.3.4.1 Medium-Term Conflict Detection

The air traffic research field of Medium-Term Conflict Detection (MTCD) predicts the future position of aircraft based on their flight plans. Subsequently the risk is assessed based on spatio-temporal overlap measures which have similarities with those discussed in section 2.3.2.1. For this thesis however, information on the intent of vessels is lacking, so generally the models applied in MTCD cannot be used. However if the intent of a vessel can be inferred (e.g. in section 2.3.5)to sufficient extent, then the methods from this research field may prove valuable.

Prandini and Hu (2016) aim to predict traffic complexity (density) on the medium-term time horizon. This is different from most MTCD approaches, which are focused on the spatio-temporal overlap of two aircraft. This is for example the case in an earlier paper (Lygeros and Prandini, 2002), but this paper also includes an explicit model for wind. This approach is typical to many MTCD approaches, which model the aircraft behaviour and model the behaviour of the wind. These aspects may still prove useful if the condition for the intent is met - as mentioned earlier.
The other concept commonly applied in MTCD , but also in road traffic conflict detection is that of (stochastic) reachable sets (Althoff et al., 2009), which ties into the idea of explicitly including the prediction uncertainty that comes along with predicting spatio-temporal overlap.

#### 2.3.4.2 Manoeuvre Recognition

Prediction can also be done based on the internal model of the human operator of a vehicle. The methods of Houenou et al. (2013) and Lowe and How (2015) derive the type of manoeuvre that a human operator is engaged in, based on spatial measurements. This information is at a higher abstraction level (human cognition), which is then used to better predict what the vehicle will do. Houenou et al. (2013) have a relatively simple 'manoeuvre mode' model, but Lowe and How (2015) apply multiple layers of abstraction, predicting on each level. These methods contribute by involving the human aspect (discussed in section 2.3.3) in spatio-temporal prediction.

### 2.3.5 Data-driven Prediction

Model based prediction can be done on a short-term, by straightforward propagation of the current state forward into time. However on the medium-term time horizon this approach becomes erroneous, and sufficient accuracy can only be achieved if the intent of the vessel is known. The problem of lacking of information on the intent can potentially be solved by using historical data. This section discusses the contributions of data-driven prediction methods.
Trajectory-based prediction (Lancia et al., 2014; Pallotta et al., 2013) clusters historical trajectories, based on which planned paths are predicted. Point-based prediction links the current position (or state) to a single future position or state (Ristic et al., 2008; Zandipour et al., 2008). Goal-based prediction (Simmons et al., 2006; Krumm et al., 2013; Best and Fitch, 2015; Vasquez et al., 2009; Ikeda et al., 2013) uses the current and/or past states of an object to predict what its goal state will be.
The maritime environment under consideration is structured in a manner similar to road traffic, except that there is much more freedom in manoeuvring space, speed and even direction of travel. Therefore a continuous representation of the map is preferred. Maritime traffic patterns depend much on seasonality and environmental conditions (visibility and weather). Therefore it is an advantage if an approach can take these aspects into account. Also, traffic patterns change over time which requires an approach

| Type | Authors | Map | Seasonality & Environment | Adapt to changes | Independent of Historical Trajectory |
|---|---|---|---|---|---|
| Trajectory | Lancia et al. (2014) | Graph/Continuous | No | No | No |
| | Pallotta et al. (2013) | Graph/Continuous | Season | No | No |
| Point | Ristic et al. (2008) | Continuous | No | No | Yes |
| | Zandipour et al. (2008) | Grid | Season | Yes | Yes |
| Goal | Simmons et al. (2006) | Graph | Season* | No | Yes |
| | Krumm et al. (2013) | Grid/Graph | No | No | Yes |
| | Best and Fitch (2015) | Continuous | No** | Yes | No |
| | Vasquez et al. (2009) | Graph (changing) | No | Yes | No |
| Sub-Goal | Ikeda et al. (2013) | Graph/Continuous | No | No | No |

*Not included, but explicitly made possible
**Independent of both factors, can learn any new environment

**Table 2.1:** An overview of the data driven prediction methods and their contributions

that can adapt to these changes. Finally, dependence on the historical trajectory of an object for prediction can be seen as a disadvantage. In table 2.1 an overview is given in which areas each approach contributes.

### 2.3.6 Interactions Prediction

Methods that model interactions between vehicles are promising in that they may deduce the adapted path. Xu et al. (2015) model the interactions between overtaking vessels. This is done by modelling each vessel as a basic decision-making, communicating agent, which decides to adjust his speed based on the encounter and communication. Chauvin and Lardjane (2008) model the interactions between crossing vessels with a logistic regression model. It probabilistically predicts which vessel will take action first, and in which direction that action takes place.
The social forces model (Helbing and Molnár, 1995) applies a series of forces which determine the movement of a pedestrian relative to other pedestrians and the surroundings. Elfring et al. (2014) build on this concept, employing the GHMM approach of Vasquez et al. (2009) to predict long-term pedestrian behaviour, and the social forces model to predict interactions. The advantage of the combined approach is that using a more simple, discrete model is preferable for longer-term prediction, whereas a more detailed, continuous model is preferable for the detailed interactions. This of course includes the adaptive benefit of the GHMM approach.

### 2.3.7 Overview of contributions

In table 2.2, the contributions of each literature piece is given. The literature on grounding risk is left out due to its low correlation with the rest. The first two columns indicate if a piece of literature predicts states, assesses risk, or both. The next three columns represent which factors of collision risk are accounted for (this can be done in risk *and/or* prediction). If marked with a '0', this signifies that the authors have explicitly stated the option of including such factors, or else that the method clearly displays the potential of including it. For example, Vaněk et al. (2013)'s agent-based approach is aimed at risk assessment and includes spatio-temporal and human factors, they include interactions between agents, and their model has clear potential to also include an environmental agent.
The 'uncertainty' column indicates if a method explicitly models uncertainty as a component of risk (or prediction). The word 'interactions' is set true for methods that model interactions between agents/vehicles. This is linked to risk for some approaches, and linked to prediction by others (1st and 2nd column). Methods that aim to infer internal intent layers are marked in the 'internal intent' column. Finally, in the last column, methods that (can) be applied at a medium-term time horizon are marked positive. All the data-driven methods are marked positive in the final column; the differentiation of their contributions can be found in table 2.1.

### 2.3.8 Grounding

Grounding risk is related to collision risk since it is a 'collision' with an underwater object or with the seabed. In literature it is often treated alongside collision, since vessels sometimes ground in an attempt to avoid collision. In the considered case study, the only area under threat of grounding is that of the deep draught channel. The vessels with a deep draught have a chance of grounding if they deviate from their planned path. Only the channel itself is deep enough for these vessels, so straying out of the channel would result in grounding. Grounding has been extensively studied in literature, but since it is only a small component of this problem and because the situation is very simple (one straight channel), it is not included in this literature survey.

| Category | Authors | Area | | Factors | | | Considerations | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Risk | Prediction | Spatio-Temporal | Human | Environmental | Uncertainty | Interactions | Internal Intent | Time Horizon |
| Two-vessel Encounter | Montewka et al. (2010) | + | | + | | | | | | |
| | Li and Pang (2013) | + | | + | | | + | | | |
| | Szlapczynski (2006) | + | | + | | | | | | + |
| | Kuwata et al. (2014) | + | 0 | + | | | | | | 0 |
| | Zhang et al. (2015) | + | | + | 0 | 0 | | | | |
| | Bukhari et al. (2013) | + | | + | 0 | 0 | | | | |
| | Goerlandt et al. (2015) | + | | + | + | + | | | | |
| Traffic Complexity | Wen et al. (2015) | + | | + | | | | | | |
| | Delahaye and Puechmorel (2000) | + | | + | | | 0 | | | |
| | Rahman et al. (2012) | + | 0 | + | | | | | | |
| | Wang et al. (2015) | + | | + | | | | | | + |
| Socio-technical system | Vaněk et al. (2013) | + | | + | + | 0 | | | | |
| | Lefèvre et al. (2012) | + | + | + | + | | | + | + | |
| | Blom and Sharpanskykh (2015) | + | | | + | + | | + | | |
| | Bouarfa et al. (2013) | + | + | + | + | + | + | + | + | |
| MTCD | Prandini and Hu (2016) | + | + | + | | + | | | | + |
| | Althoff et al. (2009) | + | + | + | | | | | | |
| Manoeuvre recognition | Houenou et al. (2013) | | + | + | + | | | | + | |
| | Lowe and How (2015) | | + | + | + | | | | + | |
| Data-driven prediction | Lancia et al. (2014) | | + | + | | | | | | + |
| | Pallotta et al. (2013) | | + | + | | | | | | + |
| | Ristic et al. (2008) | | + | + | | | | | | + |
| | Zandipour et al. (2008) | | + | + | | | | | | + |
| | Simmons et al. (2006) | | + | + | | | | | | + |
| | Krumm et al. (2013) | | + | + | | | | | | + |
| | Best and Fitch (2015) | | + | + | * | * | 0 | | | + |
| | Ikeda et al. (2013) | | + | + | | | | | | + |
| | Vasquez et al. (2009) | | + | + | | | | | | + |
| Interaction prediction | Xu et al. (2015) | | + | + | | | | | | + |
| | Chauvin and Lardjane (2008) | | + | + | | | | | | + |
| | Helbing and Molnár (1995) | | + | + | + | | | + | | |
| | Elfring et al. (2014) | | + | + | + | | | + | | + |

Legend

| | |
|---|---|
| + | yes |
| 0 | potentially |
| "blank" | no or not applicable |
| * | implicit or indirectly |

**Table 2.2:** An overview of the literature and their contributions

# Chapter 3

# Dataset Description

This purpose of this chapter is to provide a thorough description of the dataset used in this research. It originates from the Vessel Traffic Services (VTS) system developed by Saab Technologies B.V., called V3000, for the case study as discussed in section 1.4. It provides an overview of the origin of the data, the reliability of the data, its format and further notes that are relevant for this research.

## 3.1 From source to dataset

In figure 3.1, an overview is given of the data, from source to dataset format. The data is obtained from three sources: vessel AIS transponders, radar stations, and the harbour database. The radar data provides time-stamped positional plots provided by the radar data processor, which turns the raw radar data into usable radar plots (by noise filtering, detection thresholds etc.). The harbour database provides data for most vessels going in and out of Rotterdam. This includes data on identity, dimensions, draught, destination, and type. The AIS transponders provide data on state, identity, dimensions, draught, destination, ETA, type and status information. There is a discernment between static, semi-static and state data. State data is defined as the collection of variables that describe the state of the dynamical system of a vessel. Static data is defined as the collection of variables that essentially does not change over time (e.g. identity). Semi-static data is defined as the collection of variables that changes very slowly over time (e.g. vessel draught), where the change is significant enough within the time span that a vessel is considered.
The state data is processed using a method based on the Interacting Multiple Model Joint Probabilistic Data Association {*Avoiding Track Coalescence} (IMMJPDA*) filter by Blom and Bloem (2006). The word 'based' is used because the track coalescence problem is handled in a different manner. Within the SAAB system, the result of this process is called a 'track', which is an estimate of the actual state of a vessel. The (semi-) static data from either source, is called a 'plan' within the SAAB system. From here on, semi-static data and static data are grouped under the term 'static data'.
Tracks are linked to plans in two different ways. Database plans are linked to vessels manually, by a VTSO , who connects the information from the database to vessels he sees on-screen. AIS plans are linked to their corresponding tracks based on the identity information that is sent along.

## 3.2 Reliability and Accuracy

This section discusses the reliability and accuracy of the dataset (based on its sources). First the state data is discussed, then the static data.

### State data

The data obtained from the AIS transponders depends on the crew and sensor equipment on board a vessel. The state data is generally measured using a Global Positioning System (GPS) device, the accuracy of which depends on its type and make. Based on historical values, SAAB sets the accuracy of the AIS location at 95% within 30m radius around the given location. Some vessels carry high accuracy equipment, such as Differential GPS (DGPS); the accuracy is then estimated at a 15m radius. The speed and course of a vessel are direct derivatives of the GPS measurements. The orientation of a

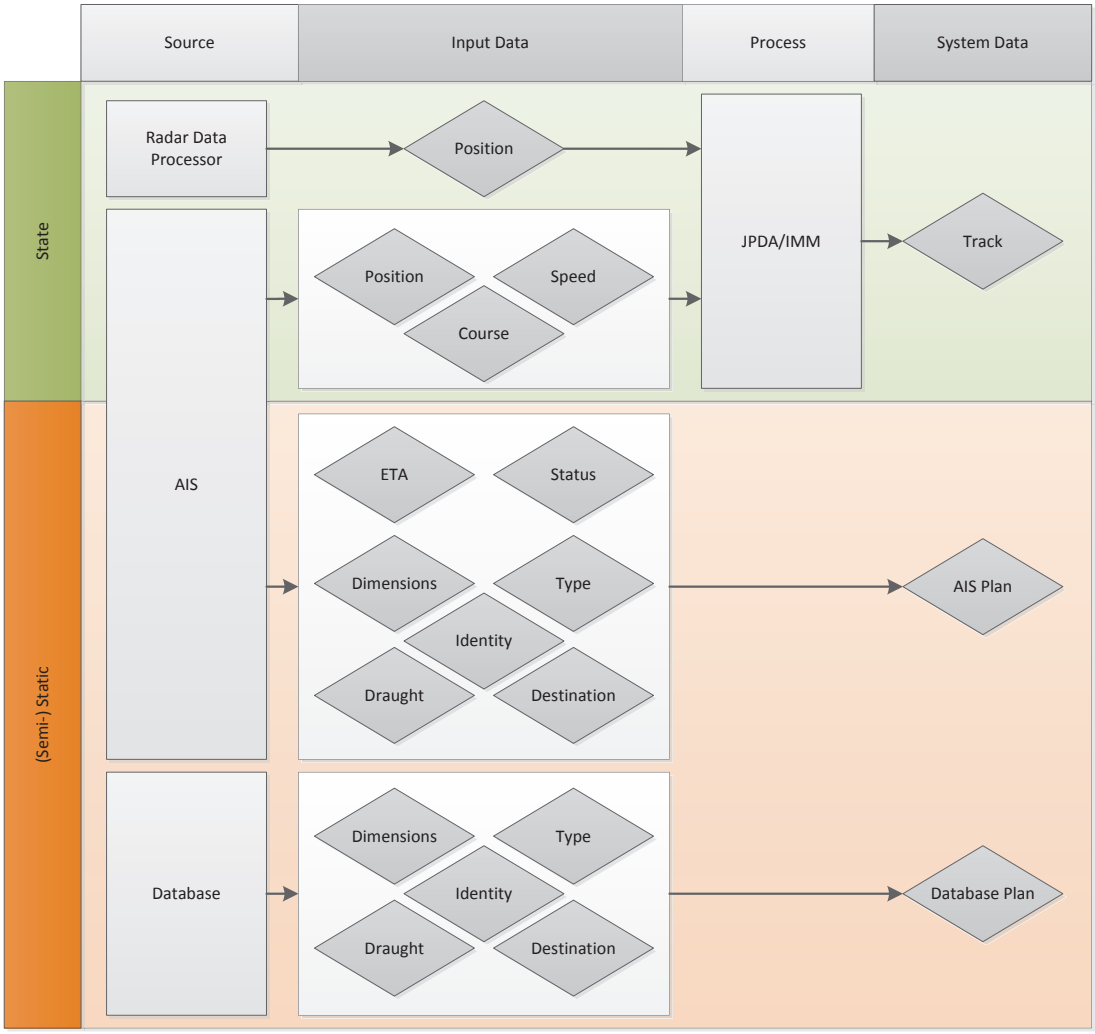



**Figure 3.1:** Overview of data used (form source to dataset).

vessel (if available) originates from a compass sensor. The frequency of AIS messages depends on the state of a vessel. A vessel should, when underway, emit a state message at least every 3-10 seconds (ITU, 2010) but it is often less.
The radar sensors each have an accuracy of 20m in range, and 0.3°azimuth. This means that at 5km range from the sensor, the azimuth accuracy is equivalent to 26m, and at 50km range it is 260m. This assumes a guaranteed detection of a vessel. The detection chance of a vessel mostly depends on the signal strength reflected back by the vessel, which decreases with distance.
In figure 3.2, the coverages for all the sensors are shown (overlaying the TSS). The accuracy of the fused data (tracks), is not directly known, but it has been confirmed that it is in most cases more accurate than the most accurate data source used.

### Static data

The static data given by the database is governed by harbour authorities, and thus follows strict protocol. This includes checks on the information sent, and checks on manual entries. Therefore this data is considered very reliable. However, the static data coming from the AIS transponders, is generally entered manually without protocol, and therefore often contains errors. This even includes mistakes in identity information. The static data from AIS transponders is transmitted approximately every 6 minutes when underway (ITU, 2010).

## 3.3 Data Storage and Format

Tracks are distributes by the V3000 system every three (3) seconds, while plans are updated whenever new information is received from an AIS transponder or from the harbour database. However, all plan data is repeated -'updated'- every 60 seconds. These updates are sent as messages to the user interface, as well as to the logging system. There are three types of messages stored that are relevant for this research (though many more messages are logged by the system, but they are filtered out): track update, plan update, and plan delete. The latter is a message to announce that a plan is terminated. For tracks, this is done in a track update message, using a status flag. In section 3.1 the data was referred to by general category (e.g. identity) only; detailed description on the data fields is given in appendix A.
Every track is given an identity number (trackID), and also every plan is given an identity number (planID). Each plan is a container that can hold one AIS plan and/or one Database plan. If a track has been linked to a plan, the corresponding planID is registered in the 'planId'-field of the track.
A track does not necessarily have to have a plan; a vessel/object may have been detected by radar without knowledge of its identity through AIS or Database. Also, a plan does not necessarily have to be linked to a track. For example, the system may have registered a vessel in the area (e.g. a planned arrival report in the database), while the vessel has

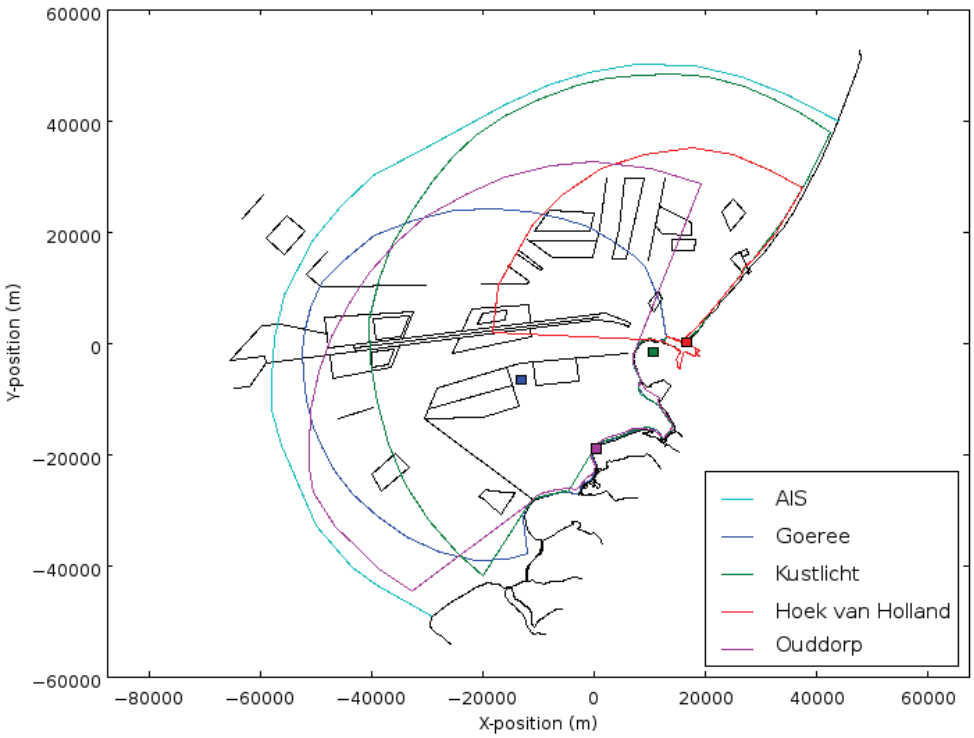



**Figure 3.2:** Illustration of the sensor coverages. The cyan border displays the AIS coverage. Each square indicates the location of a radar, and the border with the corresponding colour is the coverage of that radar.

not yet entered the radar/transponder coverage area.
In the logging system the messages are stored in binary format. Data is kept up to three months after occurrence, after which it is discarded to make room for new logging. The dataset used for this research is one set of approximately three months (84 days), from 14 May 2013 16:03 until 8 August 2013 23:59. For the sake of computing resources (time, memory and storage) and overview, the data from the three relevant messages was first converted into CSV format before use. In appendix A the data fields, types and availability are given. In table A, the fields actually used in this research are listed by category. The name column indicates the name with which the variable is referred to in text.

| Message | Category | Name | System Name | Description | Unit |
|---|---|---|---|---|---|
| Track | Identification | MMSI | MMSI | AIS identification MMSI | |
| | | plan number | plan.planId | Linked plan identification number | |
| | | track number | trackNumber | Track identification number | |
| | Geodetic State | course | course | Geodetic Course w.r.t. North | rad |
| | | speed | speed | Geodetic Speed | m/s |
| | Euclidean State | x position | xPosition | System coordinates x-position | m |
| | | y position | yPosition | System coordinates y-position | m |
| | | x velocity | xVelocity | System coordinates x-velocity | m/s |
| | | y velocity | yVelocity | System coordinates y-velocity | m/s |
| | General State | turn rate | turnRate | Turning Rate | rad/s |
| | | time | stamp | Timestamp (seconds since 1970) | s |
| | Static | (vessel) breadth | breadth | Vessel Breadth | m |
| | | (vessel) length | length | Vessel Length | m |
| | | (AIS) navigation status | navigationStatus | (AIS) Navigation status (e.g. Underway) | |
| | Journey Determination | anchor number | anchor.anchorNumber | Anchor identification | |
| | | buoy number | buoyData.id | Buoy identification | |
| | | track type | classificationType | Internal Track Type (e.g. stationary object) | |
| | | object type | objectTypes | Internal Object Type List | |
| | | track status | status | Tracking Status (e.g. lost or buoy) | |
| Plan | Container | plan number | planId | Plan identification number | |
| | | time (plan) | stamp | Timestamp (seconds since 1970) | s |
| | | (vessel) length | length | Vessel Length | m |
| | | (vessel) breadth | breadth | Vessel Breadth | m |
| | Database | (vessel) breadth | 0.breadth | Vessel Breadth | m |
| | | (vessel) length | 0.length | Vessel Length | m |
| | | (vessel) draught | 0.draught | Vessel Draught | m |
| | | destination (database) | 0.destination | Database Destination | |
| | | vessel type (database) | 0.vesselType.vesselType | Database vesselType | |
| | AIS | MMSI | 2.MMSI | Maritime Mobile Service Identity | |
| | | IMO number | 2.imoCode | IMO identification | |
| | | (vessel) length | 2.length | Vessel Length | m |
| | | (vessel) breadth | 2.breadth | Vessel Breadth | m |
| | | (vessel) draught | 2.draught | Vessel Draught | m |
| | | (vessel) air draught | 2.heightOverKeel | Vessel 'Air Draught' | m |
| | | destination (AIS) | 2.destination | AIS Destination | |
| | | (AIS) navigation status | 2.navigationStatus | AIS Navigation Status | |
| | | UN transport mode | 2.unTransportCode.mode | United Nations Transport Mode | |
| | | UN transport code A | 2.unTransportCode.codeA | United Nations Transport Code A | |
| | | UN transport code B | 2.unTransportCode.codeB | United Nations Transport Code B | |

**Table 3.1:** Data fields used in this research, with names and description. Some fields are interlinked or double. For example, speed is a vector product of x velocity and y velocity (but in a different reference frame).

# Chapter 4

# Research Method

This chapter describes the focus of this research. With a given context and aim described in chapter 1, and the current related advances in research described in chapter 2, and the dataset available (chapter 3), this chapter digs deeper into the research question, explaining its sub-questions. Subsequently, the research methodology is explained in a nutshell.

## 4.1 Research Questions

As stated in chapter 1, the main objective of the thesis is *to provide a basic understanding of the process of medium-term behaviour of vessels*. For medium-term behaviour the dynamic model of a vessel becomes less important, while the intent of the vessel (destination, ETA or planned path) becomes more important (Lancia et al., 2014). Unfortunately, information about this intent is not (directly) available, because there is no 'sail plan' mandate similar to the 'flight plan' regulations in air traffic. Only the vessels that plan to enter the harbour within 24 hours are obliged to provide their destination and in some cases their ETA . Other vessels are free to provide such information, which is not commonly done (reliably). This is an incentive to derive the intent of a vessel based on the data that *is* available. Therefore, the main research question of this thesis is:
*Which variables can be used to predict the intent of a vessel?*

The data is structured for tracking purposes (monitoring the state of a vessel). To investigate intent, the data needs a structure that regards the entire trip that a vessel makes (approximately capturing the intent of a vessel). To be able to generalize conclusions, statistical evidence is needed. However, each trip of each vessel is unique, which is why a form of grouping the trips is necessary. This grouping needs to be done based on some aspect of vessel intent (e.g. planned path), in order for it to be meaningful for the medium-term prediction of vessel behaviour. If these different groups of intent are established, prediction can be done only if it is possible to distinguish between these groups, using the available data. This results in the following sub-questions:

1. What is a useful way to group the data based on intent?
2. Which variables would allow distinguishing between intents?
3. To what extent can historical data provide sufficient support for a medium-term prediction method?

## 4.2 Research Methodology

As a method to answer these research questions, the following steps have been taken in this research (listed here and discussed below).

1. Preparation
   (a) Formatting
   (b) Cleaning

   (c) Filtering
   (d) Selection
2. Exploration
3. Analysis
   (a) Group journeys by route
   (b) Group routes by starting point
   (c) Investigate whether/which variables distinguish routes
   (d) Investigate whether planned paths can be derived
4. Validation

First, the dataset needed to be prepared in terms of format, cleaning, filtering and selection. The format of the data is based on the principle of a journey. A journey is defined as all data belonging to a single vessel from the point that it enters the map or starts moving, up until the point that it leaves the map or stops moving.
Second, in order to gain insight into the available data, the dataset was explored by plotting the selected variables in multiple ways and relative to each other, outliers were investigated and preliminary conclusions were drawn to be used further on in the thesis.
Then, the most important part of the research is the analysis. Here, the journeys in the data were grouped by route. A route is defined as as a pair of spatial positions: the starting point and ending point of a journey. Journeys with similar starting points and similar ending points were grouped together into one route. This was done using the clustering method DBSCAN . This grouping criteria captures the intent of a vessel with minimal assumption. The path between the points is not assumed, neither is time considered. It is only assumed that the origin and the aim of the vessel is similar within a group.
These route clusters, in turn, were grouped by their starting points, for the purpose of comparing routes with the same starting points to each other. For each variable it was investigated within each of these route groups how suitable the variable is to distinguish between the different routes of similar origin. This was done by examining distribution plots of each variable, colour coded by route. A variable is considered distinguishing if for most of the possible values of that variable, the likelihood that the respective vessel will take one specific route is larger than 80%.
Then for each journey, a planned path (series of waypoints) was derived, that is, a procedure was set out and applied, but has not yet worked. A planned path describes the intent of a vessel in more detail than a route. Therefore it was expected to improve the prediction at a medium-term time horizon.
Validation was done by sampling 200 journeys and predicting which route each of these journeys would take based on their starting point and the most distinguishing variable. This prediction was then compared to their actual route to validate how suitable the variable is for distinguishing between routes.

# Chapter 5

# Data Preparation

In this chapter all the steps that have been taken to prepare the data are explained in detail. The steps roughly follow the CRISP-data mining methodology (Chapman et al., 2000): formatting, processing/transformation, cleaning, filtering and selection. Context specific steps have also been made, which are explained where needed. Note that the steps in data preparation and steps in data exploration together form an iterative process, not a purely sequential process. Therefore the steps do follow a logical flow, but not necessarily a chronological/sequential flow. The chapter concludes with which variables were selected and why.

## 5.1 Timespan selection

When the number of tracks and the number of plans were plotted versus time, it was seen that there was a 2.5 hour gap in the data, on 17-06-2013 between 10:43 and 13:15. Since for validation purposes it was planned to split the dataset into two parts, this gap in the data was chosen as a convenient position to split the dataset in two.
Another important cut in the data was to remove all data after UTC 01-08-2013 00:00. This was done because the TSS was changed at this time, causing a significant change in traffic behaviour This finally leaves the first part at a size of approximately 34 days, and the second part at approximately $44\frac{1}{2}$ days. It is important to note that the results shown are from the first part of the dataset.

## 5.2 Format

Since the analysis of the data is centred around moving objects, a suitable data format is a time-based registry of all known vessels with all their corresponding variables. To get the data into this format, an algorithm was written which converts the stream of stored messages into a time-based monitoring of plans and tracks. This algorithm maintained a set of current tracks and a set of current plans. Each time a track update message is received, the system status flag is checked to determine if the track is still active. If it is, its data is registered in the set of current tracks. If the track is not yet in the current tracks, it is added. If the track update status is flagged as 'lost', the data corresponding to the track number is removed from the current tracks. For plan update messages, a similar procedure is followed, except that the removal of plans is determined by a separate plan delete message.
Since this algorithm sequentially monitors the currently active tracks and plans, it can be used to sample the data by saving/storing the data at a regular time interval. The time interval used in this research is 60 seconds, to decrease the processing time, and for easy interpretation (one time sample is one minute). Each time sample contains a number of data samples. In this research a data sample is defined as one time sample of the data of one vessel.

## 5.3 Processing/Transformation

Some of the data fields need to be processed to give useful data. This section describes how the following variables were derived: UN transport code, navigation status, vessel type, acceleration, time of day, length, breadth and draught, and destination code. Time of day is of interest to investigate for example the influence of visibility, and acceleration is of interest because it is a basic motion model descriptor of vessel behaviour The remaining derived variables (length, breadth, draught, UN transport code, navigation status, and vessel type) are of interest to categorize vessels and behaviours

The categorical data which are given as (a set of) integers/bitsets were processed using conversion tables. These tables can be found in appendix B.
Acceleration was considered of possible interest to this research. Therefore it was calculated in the Euclidean reference frame, using the Euclidean velocity fields (x-Velocity, y-Velocity) and time field (stamp). Also, time of day is stored separately, by rounding each time value down to the nearest hour, storing it in the field 'hour'.
For the fields of MMSI , destination, length, breadth, and draught, there sometimes is overlap of information from the two separate sources: Database and AIS . In these cases, the data from the database is used rather than that of AIS , due to its reliability, and the AIS data is discarded entirely. In some cases, the V3000 system has determined its preference, and has stored it in the track. In these cases, the information from the track message is used.
Destinations have been grouped/coded according to a custom coding scheme (detailed in appendix B), because the number of destinations is very large, as well as the number of aliases per destination. If a destination is given by the database, it follows a standard protocol and format, which defines unambiguously what the destination is. Most of these destinations are in, around, or beyond Rotterdam harbour, implying that the vessel is headed for the harbour, which is why they are grouped into the same destination code 'RDAM'. Most destinations received by AIS are in free text form, with no convention. Therefore only destinations which are distinct and without spelling mistakes are coded. This implicitly increases reliability by excluding indistinct entries - which imply unreliable manual entries - but decreases availability. Fortunately some vessels follow the UN convention of location codes UNECE (2013), which makes it easier to identify the destination. Also, according to expert opinion, most vessels that use this convention follow certain protocols, making their data more reliable.

## 5.4 Journey Determination

A journey is defined as the entire time that an identified vessel enters the traffic scene, up until it leaves the traffic scene. A vessel is identified if the 'track' is linked to a 'plan'. The geographical boundaries of the traffic scene are defined by the coverage of AIS (see figure 3.2) and by the land boundary. Therefore a journey will start when it comes within these boundaries, and end when it leaves these boundaries. Another way that a journey is seen as ended is if the vessel stops moving, for example to go for anchor. The opposite is also true; when a vessel starts moving again, a new journey has begun. A threshold of 0.2m/s was chosen for this, There are other ways in which a journey can end, such as if the track or its identification (plan) is lost.
Also, the system has multiple ways in which it monitors the track status, navigation status or type of object (object type, buoy number, anchor number). In this research, vessels that are anchored or moored are not of interest, therefore all data points are filtered out where the navigation status indicates 'moored' or 'anchored' or where the object type is not a vessel. A journey will also end if the track is identified as moored,

anchored or an object type other than vessel.
In the following list, the conditions for a journey are summed up:

- the track must be identified by the system (be linked to an active plan)
- the track must be identified as a vessel (object type)
- the track must be inside the geographical boundaries ( AIS coverage area and land)
- the track must have a speed larger than 0.2 m/s
- the track must not be identified as a buoy, ship at anchor
- the track must not be lost or unconfirmed (based on track status)

In short, a journey is defined as the collection of data samples of a vessel from the first time a message with its identification meets the conditions for a usable track, up until the point that one of these conditions is not met. Every journey is given a number for processing purposes.

## 5.5 Connecting Journeys

A track may be temporarily lost, or it may go in and out of the AIS coverage area, or it may briefly be misidentified. These are reasons for an actual journey to be broken up into multiple journeys by the procedure described in 5.4. Therefore reconnecting these partial journeys is a necessary step in the preparation of journey data as used in this research.
If two partial journeys have the same identity -Maritime Mobile Service Identity (MMSI) - it may very well be that they are of the same actual journey. If the partial journey starts later than another partial journey, and starts with a state similar to the last state of the earlier partial journey, both partial journeys are likely one and the same journey. Therefore, journeys are grouped by MMSI . Within each group, the first data point of each journey is compared with the last of each other journey that ended before the considered journey began. Based on the difference in position and time, the connection speed (x,y) is determined; this is the speed that would be necessary to bridge the spatio-temporal gap between the compared points. This connection speed is compared with both the registered final speed of the first journey, and also to the initial speed of the second journey. If in both cases the connection speed is similar to the compared speed, and if the time difference is less than 10 minutes (maximum allowable time gap), it is considered to be the same journey.

## 5.6 Journey Filtering

Not all journeys are suitable for analysis, because they are too long, too short, non-moving, too fast or very unreliable. Therefore filtering entire journeys is a needed preparation step. The filtering choices made in this research are discussed in this section.
Journeys were considered immobile if the highest speed during the entire journey was below 0.5m/s, because ocean currents can already cause a vessel to drift at this speed. For each journey, the distance travelled was compared to the total duration of the journey, resulting in an 'average speed'. Journeys where the average speed was below 0.5m/s and the maximum speed below 2m/s were filtered out. The combination of the condition with the maximum speed is because some journeys may for example drift around at 0.3m/s for a very long time before starting to accelerate to normal speed.
Journeys with few data-points ($< 10$) or short durations ($< 5$ minutes) were cleaned out because such small journeys lack meaning. The duration limit could be raised even higher, but 5 minutes was chosen to exclude as few journeys as possible. Journeys with large time gaps ($> 10$ minutes) were removed for this thesis (in future work they should be split into multiple journeys).
Vessels that are engaged in fishing are considered to be too unpredictable in their behaviour, which is why all journeys with navigation status indicates 'engaged in fishing' are removed.
Finally, journeys were removed of which it was not clear if it is a full journey, i.e. there is no knowledge when it started or ended. Journeys which have been registered in the first time sample are removed from the dataset, since it it not known when these journeys have commenced. Likewise, journeys that have no known ending are also filtered out.
For each of these filtering conditions, spatial plots were made to confirm the paths of these journeys. The vast majority of the removed journeys were on the edge of the coverage area, or inside anchorage areas. After all connecting and filtering was performed, a total number of 20,515 journeys remained.

## 5.7 Selection

From all the available data fields, a selection was made in consultation with expert staff, based on relevance to the research. The selected fields are shown in table 5.1, with their descriptions. The possible values for navigation status have been described in table 5.2, and the possible values for vessel type can be found in table 5.3. The state of a vessel needs to be described in a complete way, since the vessel state is an inherent part of prediction. Therefore the fields for acceleration, speed and position were chosen, as well as the turn rate and course of a vessel. The position was described in the system reference frame for convenient analysis (simple distance measurements), with a small penalty in accuracy; a more accurate investigation would use the latitude and longitude, requiring more calculation steps. The vessel dimensions and category were regarded as possible causes for vessel behaviour (e.g. choice of path or final destination), as well as destination, navigation status and time of day.

| Group | Name | Code | Description | Possible values | Unit |
|---|---|---|---|---|---|
| State | X position | POS_X | X-position in system reference frame | -70,000 to +70,000 | m |
| | Y position | POS_Y | Y-position in system reference frame | -70,000 to +70,001 | m |
| | Acceleration | ACC | Absolute acceleration | 0 to 2 | $m/s^2$ |
| | Speed | SPEED | Absolute speed | 0 to 40 | m/s |
| | Course | COURSE | Course w.r.t. North, clockwise positive | 0 to $2\pi$ | rad |
| | Turn rate | TURN | Turning rate, clockwise positive | 0 to 0.5 | rad/s |
| Dimensions | Length | L | Vessel length | 0 to 380 | m |
| | Width | W | Vessel width | 0 to 50 | m |
| | Draught | D | Vessel draught | 0 to 22.55 | m |
| Category | Type | TYPE | Vessel type | see table 5.3 | |
| Other | Destination | DST_CODE | Destinations coded by geographical group | e.g. 'RDAM' | |
| | Navigation Status | STAT | Navigation Status according to AIS standard | see table 5.2 | |
| | Hour | HOUR | Time of day, grouped per hour, rounded down | 0-23 | |

**Table 5.1:** Description of the variables used in this research.

| Codename | Description |
|---|---|
| ENG | Under way using engine (default) |
| ANCH | At anchor |
| NUC | Not under command |
| RM | Restricted manoeuvrability |
| CBD | Constrained by draught |
| MOOR | Moored |
| GROUND | Aground |
| FISH | Engaged in fishing |
| SAIL | Under way sailing |
| HSC | Reserved for HSC (High Speed Craft) |
| WIG | Reserved for WIG (Wing In Ground) |
| RES01 | Reserved for future use |
| RES02 | Reserved for future use |
| RES03 | Reserved for future use |
| SART | SART active |
| UNDEF | default or SART under test |

**Table 5.2:** Description of values in the navigation status variable. IMO (2001)

| Letter | Description |
|---|---|
| P | Pilot on board indication |
| I | Ship with IMO cargo indication |
| Sp | Special ship indication |
| A | Anchor ship indication |
| G | Sea ship indication |
| Sv | Service vessel indication |
| C | Channel ship indication |
| D | Dredge ship indication |
| R | River ship indication |
| T | Tug boat indication |
| B | berthed ship indication |

**Table 5.3:** Description of the values for vessel type. Multiple values can be taken, which is by a merge of the letters (e.g. 'GIP').

# Chapter 6

# Data Exploration

In order to gain insight into the available data, several different kinds of plots were generated (section 6.1). The data is viewed from two angles: default view - each data sample is counted once, and journey view - each journey is counted once. In the journey view, numeric data was averaged over the data samples belonging to the journey, and for nominal/ordinal data the most occurring value was taken (usually the only value). From here on in this report, when referring to a plot or a distribution the term 'default' is only used when contrasted with the journey view.
Using these plots, outliers were investigated (section 6.2) and treated, after which preliminary observations were made (section 6.3) which are to be used further on in this research.

## 6.1 Plotting Variables

For each variable, the distribution was plotted for both views. For numeric data, a histogram, a boxplot and a dot plot were created (e.g. figure 6.1). For nominal data, bar charts were created (e.g. figure 6.2).
Each variable was also plotted against each other variable. The type of plot is shown in table 6.1. In order to gain insight on the spread of numeric variables, boxplots were made when plotting numeric variables against nominal variables. For the sake of viewing distribution, trellis -or jittered- plots were generated in the opposite case. In cases where a coloured bar plot was not sufficiently clear, a conditional plot was created. Also, spatial plots were generated, displaying the geographic distribution of numeric variables (grid averages).

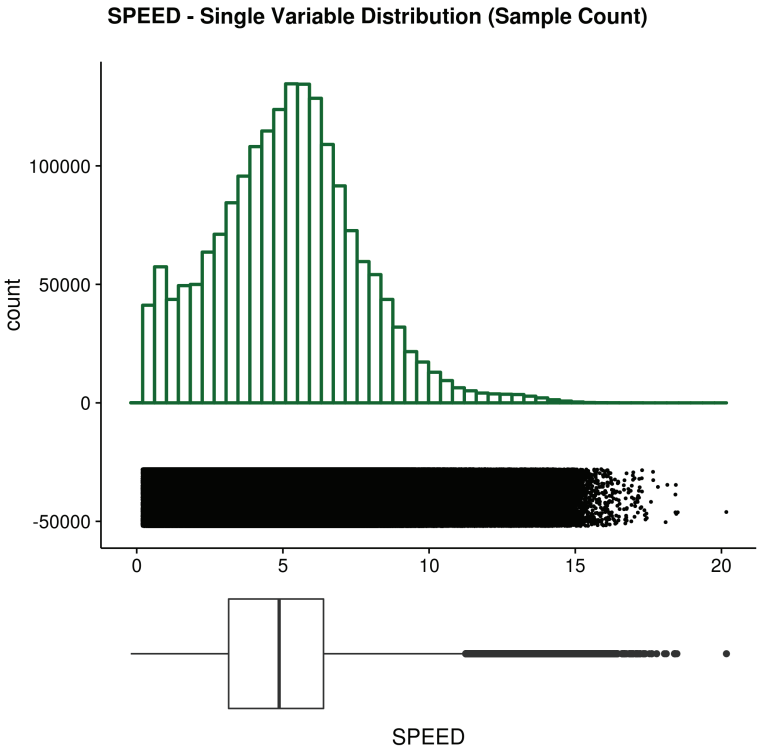



**Figure 6.1:** An example plot of the distribution of one of the numeric variables - in this case speed.

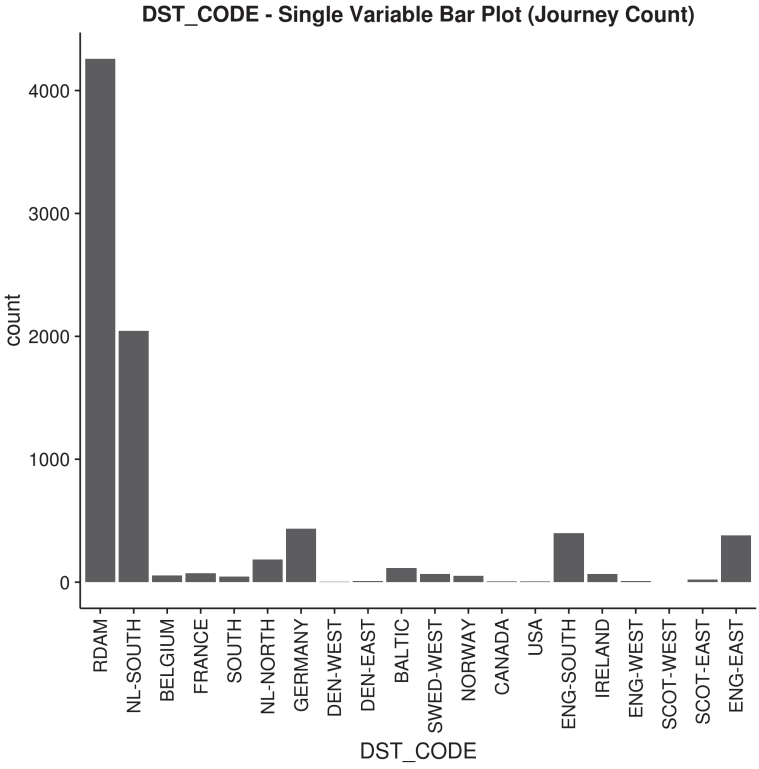



**Figure 6.2:** An example plot of the distribution of one of the nominal variables - in this case destination.

| | | X-axis | |
|---|---|---|---|
| | | Numeric | Nominal |
| Y-Axis | Numeric | Scatter | Box |
| | Nominal | Trellis | Coloured Bar |

**Table 6.1:** Plot type based on variable type per axis.

## 6.2 Outlier Investigation

Outliers were investigated starting with the single variable plots. After having filtered all the journeys as discussed in section 5.6, the only clear remaining outliers were in the dimensions of vessels: length, breadth and draught. These outliers were treated in two steps. First, the values were regarded separately. Then the ratios between the dimensions were checked for outliers.

In august 2013, the longest vessel in the world was 380m long, and the widest vessel in the world 50m wide. Also, the maximum draught possible in the deep water channel -and the entire area- is 22.55m. Any vessel exceeding these values is by definition considered an outlier. These were investigated on a case by case basis due to the low number of journey cases, in order to find an explanation for the values. Possible causes considered were: mixed up dimensions (e.g. used width instead of length), typographical error (e.g. decimal incorrect), unusual vessel type, and unusual behaviour. Solutions chosen were as follows:

**Dimension mix-up** switch values back

**Typographical** undo assumed typographical error

**Unusual Vessel Type** leave data as is, but consider journey as outlier

**Unusual Behavior** leave data as is, but consider journey as outlier

Each solution was verified by the inter-dimension ratios, ship type and speed range. If no solution was found sufficient, all dimensional variables were removed for the investigated data point or for the entire journey in the case of many data points.

The second step was to check for outliers in terms of inter-dimension ratios. Scatter plots of each dimension vs each of the other dimensions (as described in section 6.1 were used to identify these outliers. Here it was also visible that on the lower side there were also outliers: vessels with very low draught values. These outliers (low values and off-ratios) were also treated in the manner discussed above.

## 6.3 Preliminary observations

This section discusses the most relevant/notable observation made during the exploration of the dataset. The importance of it is to gain an understanding of the behaviour of the vessels in general. The first thing observed during data exploration is how often each variable is available in the dataset (as shown in table 6.2). What is especially notable is that the number of samples that contain a turn rate is very low, but that most journeys do have an average turn rate. This is because the system does not register turn rates below a certain detection threshold. Otherwise, the variables are mostly available, though it must be kept in mind that combinations may not be.

### 6.3.1 State variables

When regarding the turn rate distribution (fig. 6.3), most turn rates are very close to zero, having 50% of the datapoints in a span of 0.005 rad/s. This makes sense, since the traffic is at open sea, where manoeuvring much is not necessary, and not preferred either, since it is costly. The same goes for acceleration: a large proportion of accelerations is around the order of $10^{-2}$ m/s$^2$ , both for the journey averages as well as for the default distribution. What is also interesting, is that the journey distribution generally has higher values than the default distribution, which implies that generally longer journeys have lower accelerations.
Both course distributions have multiple peaks (figs. 6.4 and 6.5), the most notable at roughly $\frac{1}{2}\pi$ and $1\frac{1}{2}\pi$, which are associated with directions East and West, implying that most journeys are passing through the main East-West traffic lanes of the TSS . Speed has a very smooth default distribution.

### 6.3.2 Dimension variables

Draught (fig. 6.6) has a mean around 5 meters, with a notable peak at 3 meters. There is also quite a peak at 5 meters, but this is because many draught entries are rounded off, setting them exactly at 5 meters, which is enhanced by the bin size of the histogram. Though draught may have quite a smooth distribution, length definitely doesn't (see fig. 6.7). There are multiple peaks (e.g. at 30m, 80-100m), and quite a gap around 50m (verified with other plots). This gap seems to be an unusual length, falling between two more common size groups. The width distribution (fig. 6.8) has a noticeable peak at 32m, just before a clear split around 34m. Also, around 8 or 9 the same kind of dip similar to the dip in the length distribution at 50m. These gaps or dips indicate a possible means to distinguish between different intents (e.g. two vessels coming from the same place, but belonging in different distinct size classes may generally go different ways).

| | | | Samples (%) | Journey (%) |
|---|---|---|---|---|
| State | Acceleration | ACC | 100 | 100 |
| | Course | COURSE | 100 | 100 |
| | Speed | D | 88 | 90 |
| | Turn Rate | TURN | 16 | 87 |
| Dimensions | Length | L | 98 | 98 |
| | Width | W | 98 | 99 |
| | Draught | D | 88 | 90 |
| Category | Vessel Type | TYPE | 44 | 30 |
| Other | Destination (coded) | DST_CODE | 47 | 40 |
| | Navigation Status | STAT | 85 | 91 |
| | Time of day | HOUR | 100 | 100 |
| | | *Count* | *1,858,756* | *20,515* |

**Table 6.2:** Availability of selected variables in dataset in percentage. The samples column indicates the percentage of data samples that contain the variable, while the journeys column shows the percentage of journeys that does. At the bottom the total number of data samples and journeys are given (for context).

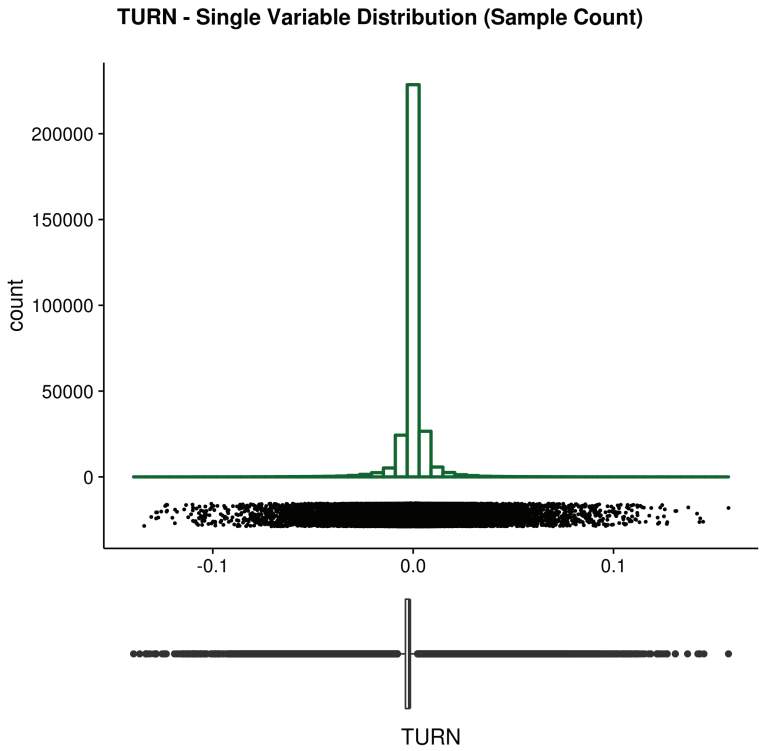



**Figure 6.3:** Turn rate distribution (rad/s).

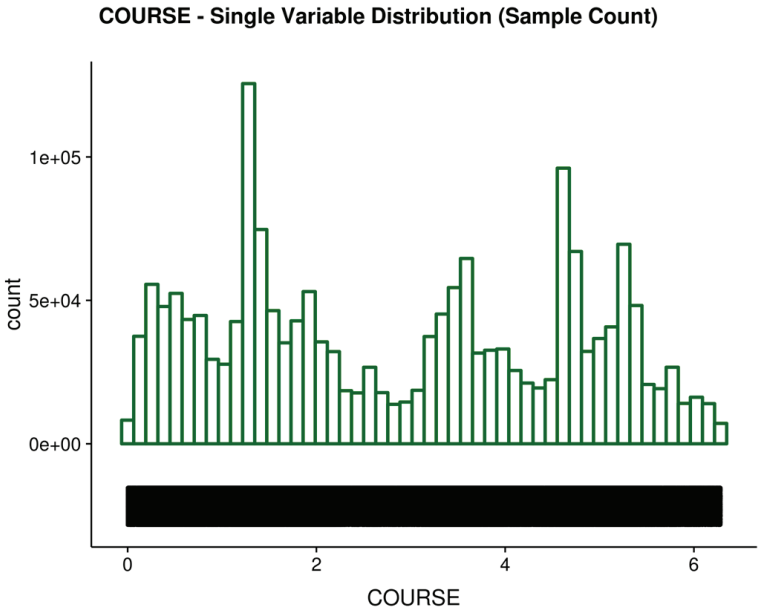



**Figure 6.4:** Standard course distribution (rad) with clear peaks at roughly $\frac{1}{2}\pi$ and $1\frac{1}{2}\pi$.

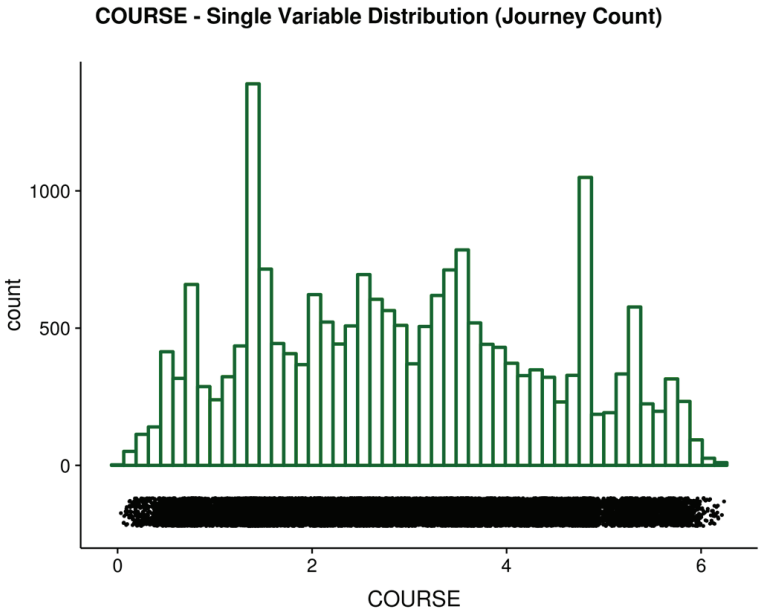



**Figure 6.5:** Journey course distribution (rad) with clear peaks at roughly $\frac{1}{2}\pi$ and $1\frac{1}{2}\pi$.

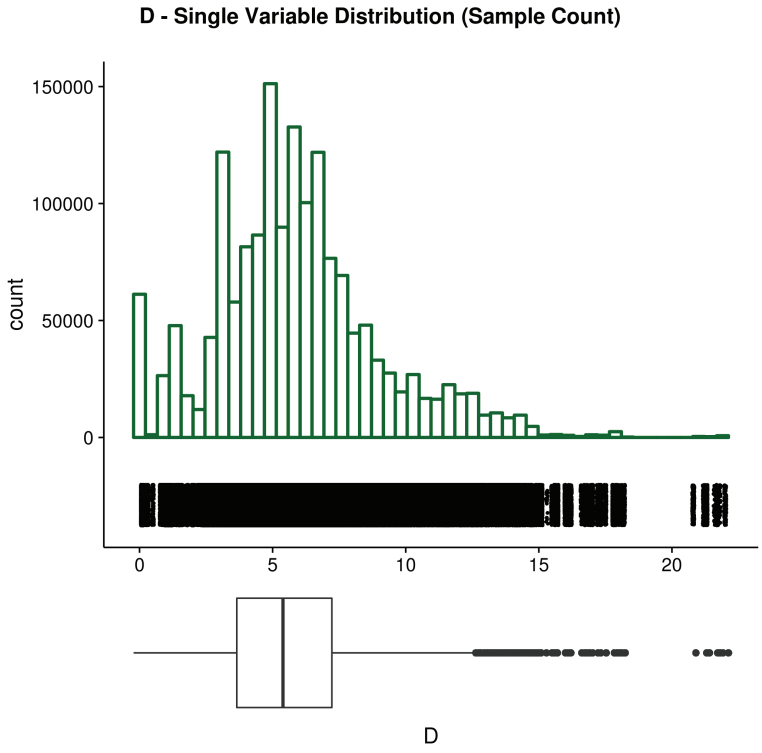



**Figure 6.6:** Draught distribution (m).

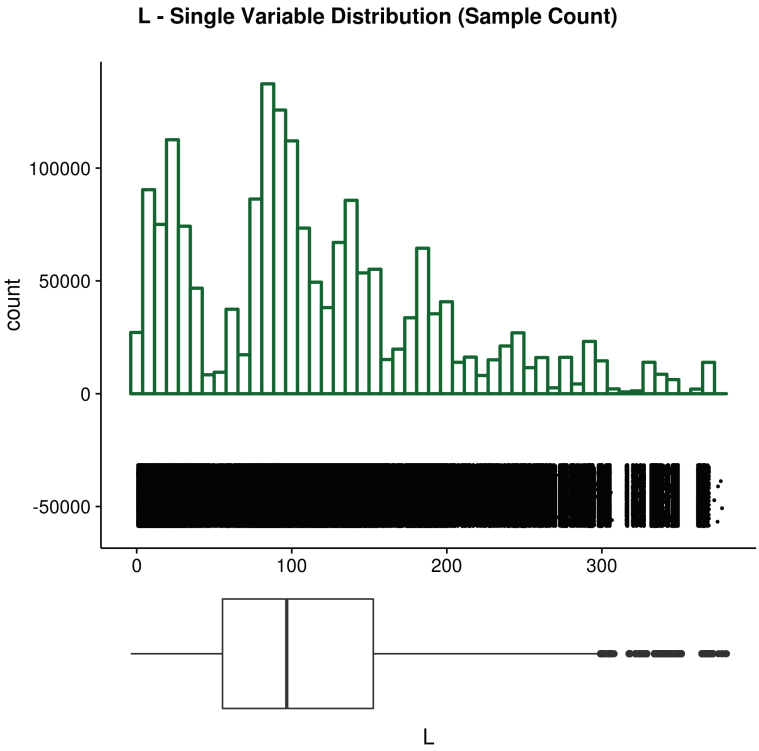



**Figure 6.7:** Length distribution (m).

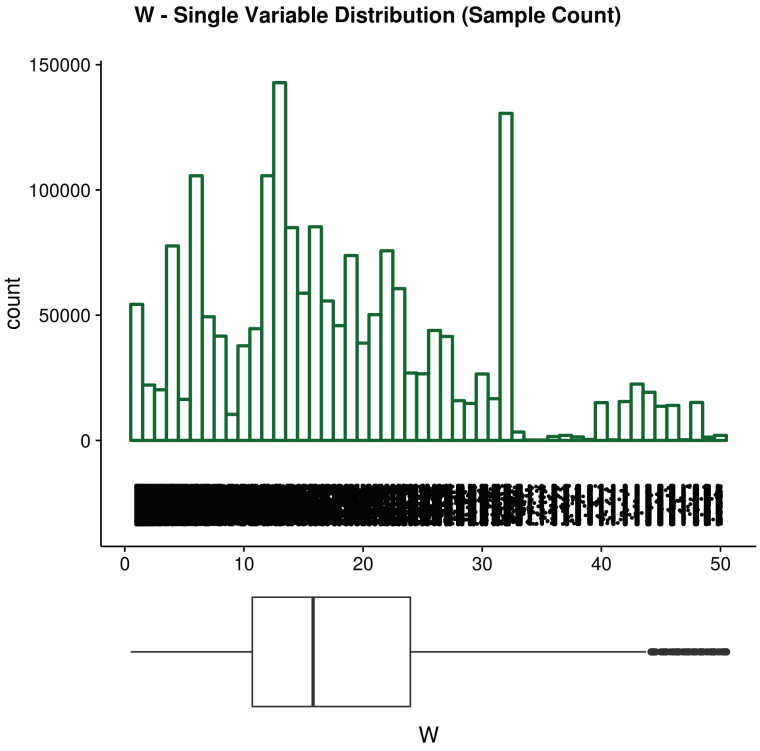



**Figure 6.8:** Width distribution (m).

### 6.3.3 Category variables

The vessel type (fig. 6.9) makes it clear that the vast majority of vessels is a sea ship (G) as opposed to the river ships (R). Few vessels sail both inland and at sea. The presence of pilots on board (P) is also quite significant. From domain knowledge it is clear that vessels with pilots on board are either entering into the harbour, or leaving it. This gives a limited range of possible intents for these vessels.

### 6.3.4 Other variables

In fig. 6.10 the most popular destination is -as expected- Rotterdam, followed by the south of the Netherlands. This makes sense because most reliable destinations come from the harbour database, which is mostly concerned with vessels headed toward the harbour, not outward. However, England and (northern) Germany do have a significant portion. Destination seems a promising variable, because it implies intent, but it also has not one, but multiple significantly prominent values.
Traffic is not entirely evenly distributed throughout the day (fig. 6.11). The highest peak is around noon and the lowest count is around midnight. The waving pattern is probably not influenced by the tidal patterns. These have a 6-hour period, where the tidal peaks shift around an hour per day, reaching a full cycle in six days. The size of the dataset is 34 days, which includes 5 full periods, outweighing the remainder of 4 days. A more likely cause seems to be that during the day, a higher throughput of boats can be handled, due to more visibility and thus more safety. Unfortunately the differences are not significant enough to be of much help when it comes to distinguish between vessels and their intents.
The navigation status of most vessels is most of the time 'underway using engine' (see fig. 6.12). This does not come as a surprise, but may not be helpful when trying to distinguish between vessels. On the other hand, especially the exceptions may prove useful.

### 6.3.5 Two-dimensional distributions

Plotting the variables against each other, together with correlation calculations (table 6.3), resulted in several additional insights. This section shortly highlights the most striking of these insights. The two-variable relationships may prove useful in deriving combined effects of vessel behaviour.
The dimension variables are all strongly positively ($> 0.80$) correlated to each other, especially length and width. Acceleration and absolute turn rate (both manoeuvring variables) are also positively correlated, possibly an indicator of local behaviour. All dimensions are slightly negatively correlated with acceleration and with absolute turn rate. Therefore the larger the ship, the less a vessel manoeuvres. However, speed is positively correlated with the dimension variables.
In fig. 6.13, acceleration and speed are not strongly correlated, though this was expected.

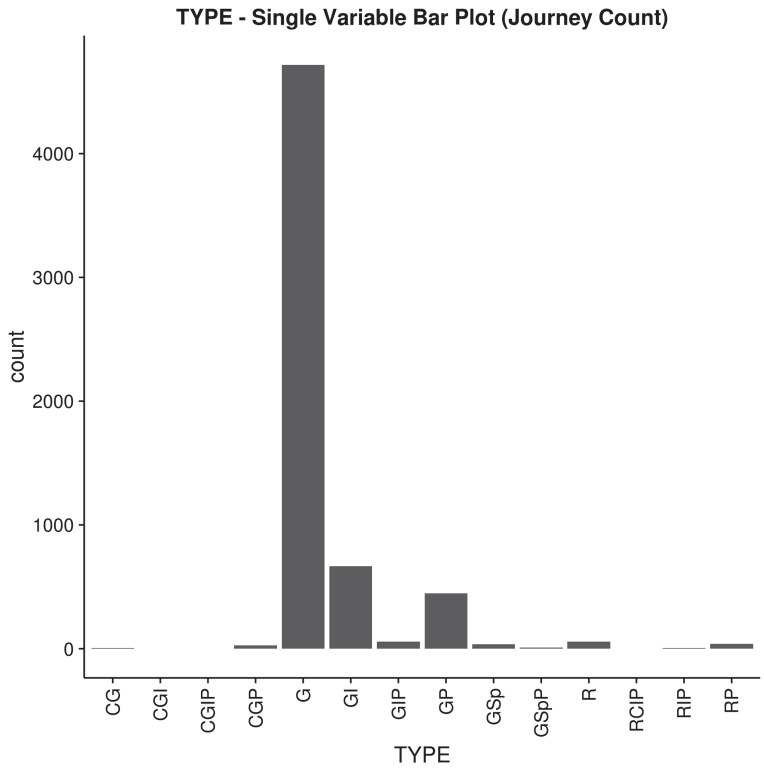



**Figure 6.9:** Vessel type distribution (by journey). See text and table 5.3 for details.

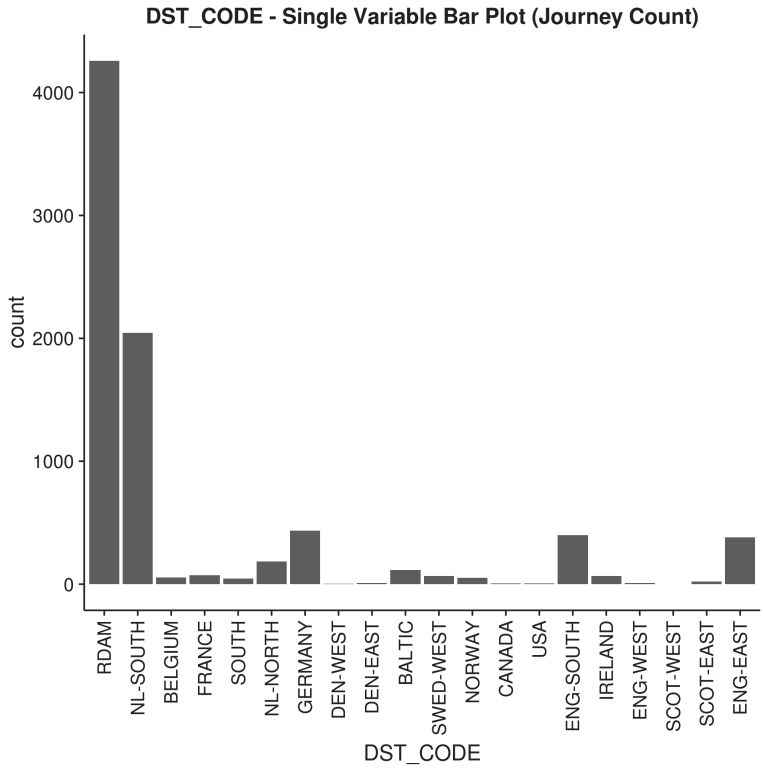



**Figure 6.10:** Destination distribution (by journey).

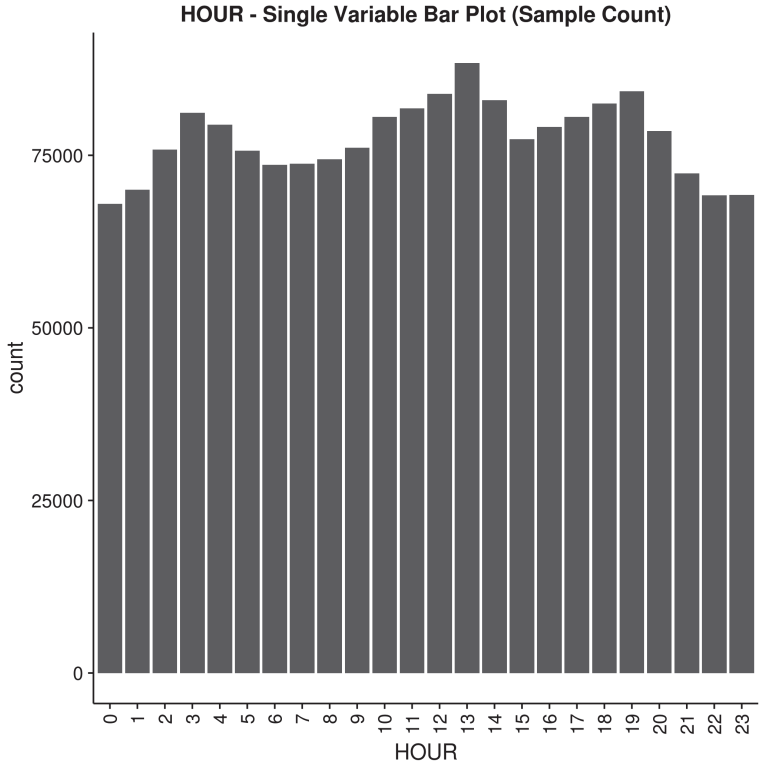



**Figure 6.11:** Time of day distribution.

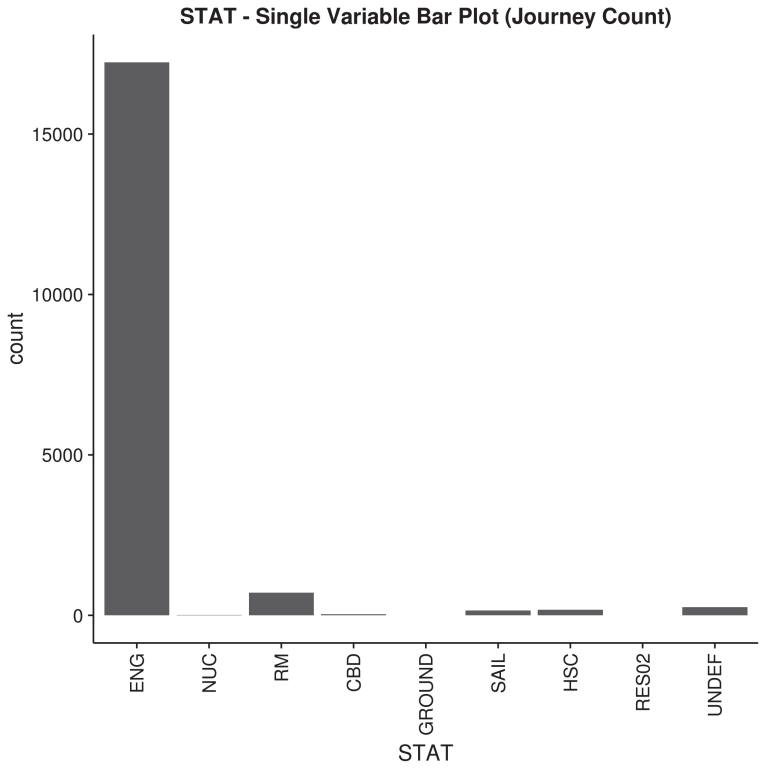



**Figure 6.12:** Navigation status distribution (by journey). See text and table 5.2 for details.

| | Length | Width | Draught | Speed | Acceleration | Turn Rate |
|---|---|---|---|---|---|---|
| Length | - | +0.96 | +0.81 | +0.16 | -0.14 | -0.18 |
| Width | +0.96 | - | +0.80 | +0.14 | -0.13 | -0.17 |
| Draught | +0.81 | +0.80 | - | +0.00 | -0.16 | -0.17 |
| Speed | +0.16 | +0.14 | +0.00 | - | +0.16 | -0.03 |
| Acceleration | -0.14 | -0.13 | -0.16 | +0.16 | - | +0.60 |
| Turn Rate | -0.18 | -0.17 | -0.17 | -0.03 | +0.60 | - |

**Table 6.3:** Pearson correlations between numeric variables. The absolute value of the turn rate is used here.

However, there does seem to be a lower limit line increasing with speed. This suggests a minimum acceleration for a given speed, which increases with speed.

In fig. 6.14 it seems clear that destination quite influences the course of a vessel. The gaps are promising in terms of being able to distinguish the intent of a vessel from another. From other plots, a few noticeable points are that vessels with destination NL_NORTH are small in size (length and draught), and those to CANADA/USA are very large. Also the speeds of vessels headed to SCOT-WEST are very low, while to DEN-WEST travel fast.

When comparing navigation status to vessel dimensions, vessels that are constrained by draught (CBD) are very large in all three dimensions, vessels that are not under command (NUC) are also relatively long and wide, but do not lie deep. High speed crafts (HSC) and undefined ships (UNDEF) are short.
When comparing the vessel type to other variables, the sea going, piloted sea ships (CGP/CGIP) are very large in all dimensions. The special ships (Sp) are all short in length and not deep, but relatively wide and move very slowly. All river vessels (R) are not wide, accelerate slowly, and have low turn rates.
To summarise, understanding the (cor-) relations between variables can prove useful. If one variable is useful for prediction on its own, its relationship with other variables can be used to understand the effect that the other variables have on vessel behaviour.

### 6.3.6 Spatial dependence

An important note on the state variables, is that they very much depend on position, as can be seen in figs. 6.15 to 6.20, where the spatial distributions of each numeric variable are shown. In these figures it becomes quite clear that it will be useful or even necessary to analyse in detail the spatial dependence of these variables.

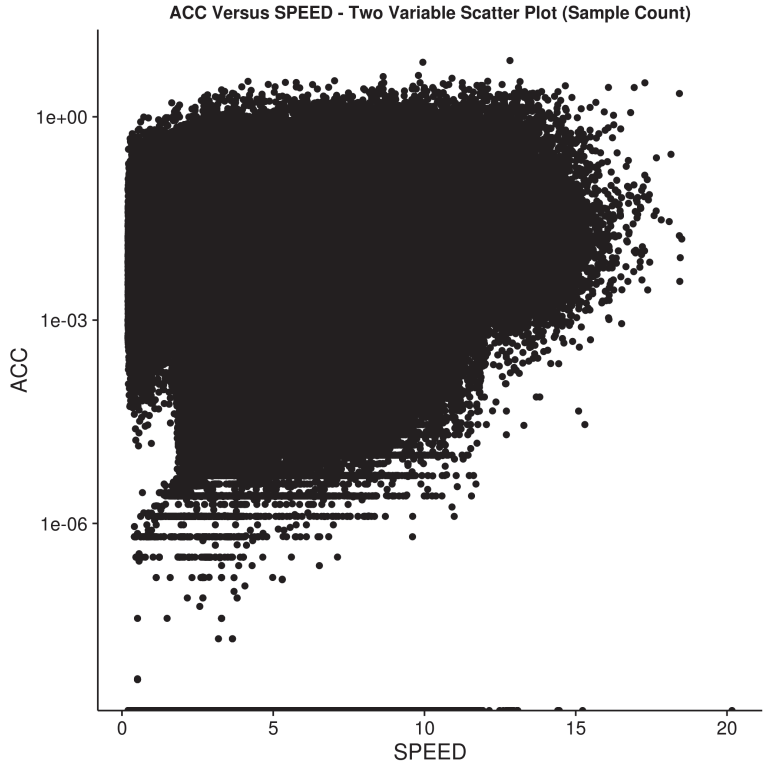



**Figure 6.13:** Acceleration vs. Speed.

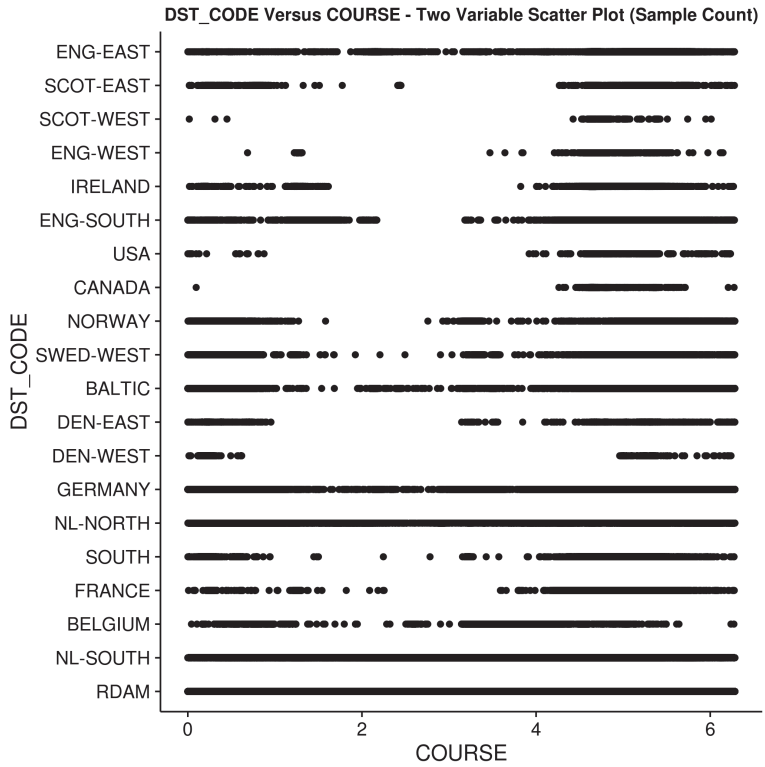



**Figure 6.14:** Destination vs. Course.

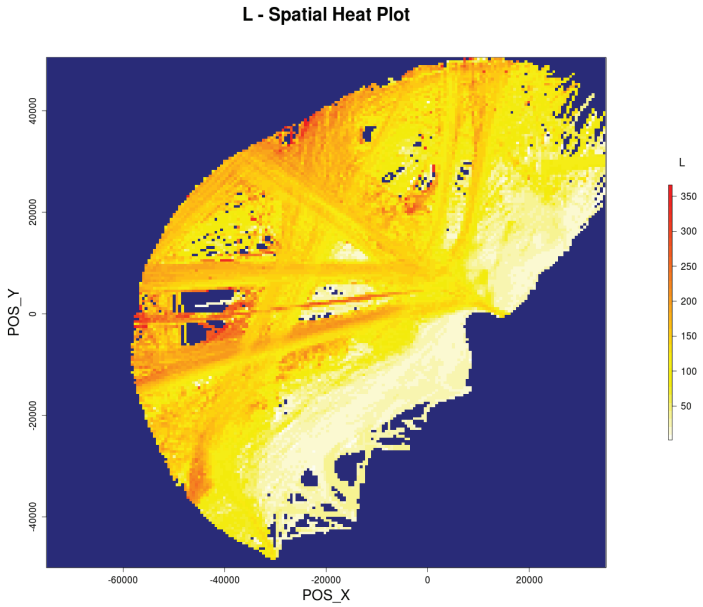



**Figure 6.15:** Spatial distribution of length (m).

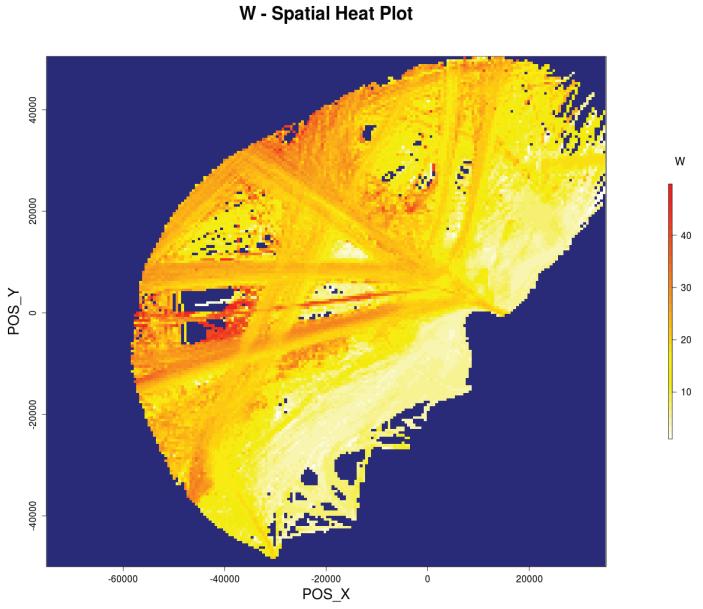



**Figure 6.16:** Spatial distribution of width (m).

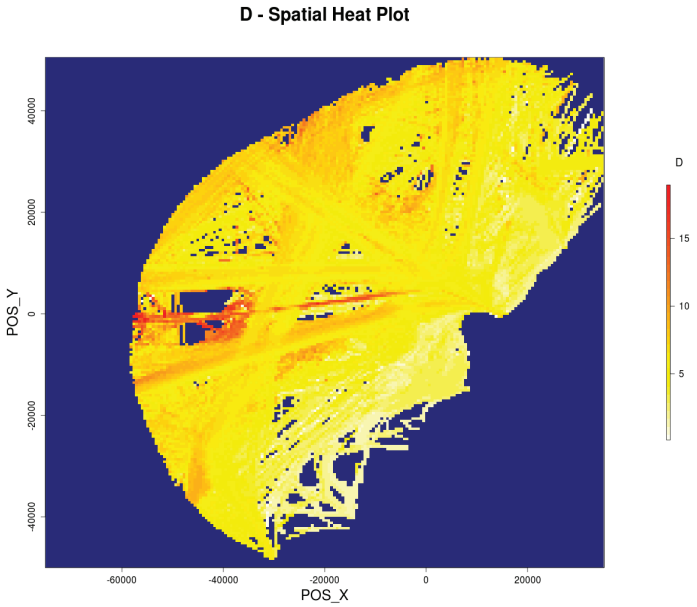



**Figure 6.17:** Spatial distribution of draught (m).

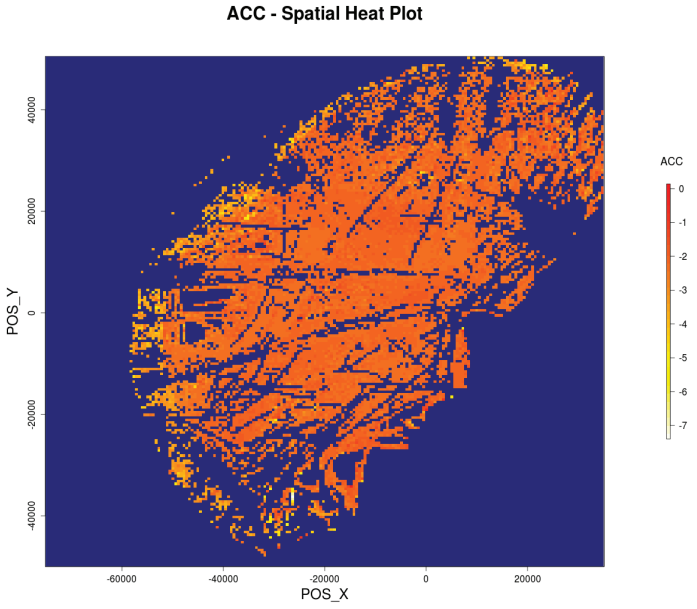



**Figure 6.18:** Spatial distribution of acceleration ($log(m/s^2)$).

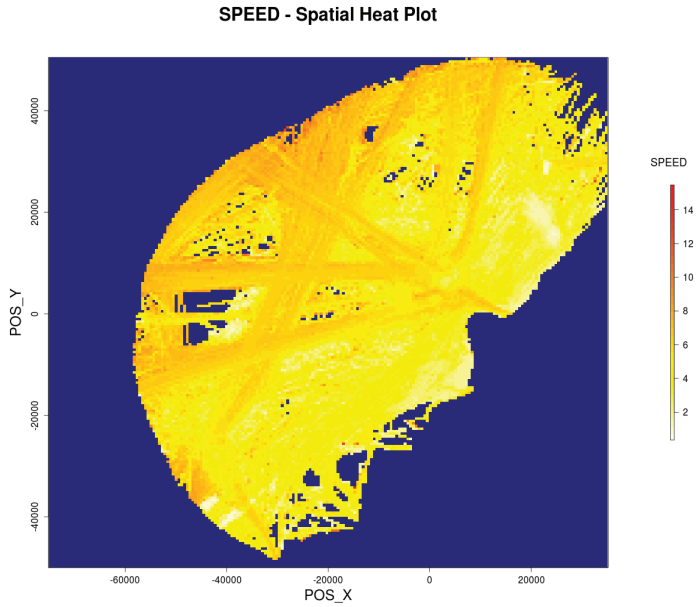



**Figure 6.19:** Spatial distribution of speed (m/s).

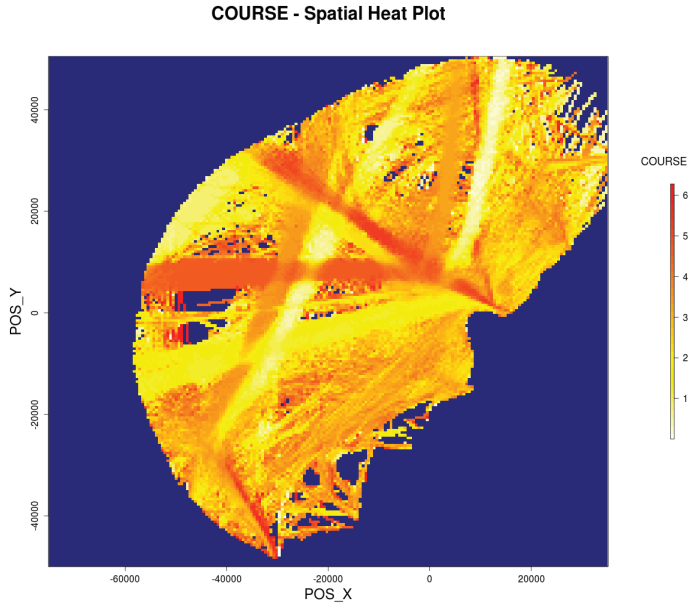



**Figure 6.20:** Spatial distribution of course (rad w.r.t. North).

# Chapter 7

# Data Analysis

This chapter discusses the steps taken in the process of analysing the data, and the applied methods. It also discusses the decisions made and their motivations.
Section 7.1 describes how routes (origin-destination pairs) were derived from the dataset. This was done by clustering entry point and exit points of journeys using the DBSCAN (Ester et al., 1996) clustering technique. Section 7.1.3 explains how the data was plotted in order to investigate which variables can distinguish between different routes (in order to predict which route a vessel will take).
Then section 7.2 describes the approach to derive planned paths (series of connected waypoints) from the dataset, to describe vessel intent in more detail for better prediction. First, waypoints were derived by detecting change-points in the course of a vessel using binary segmentation (Scott and Knott, 1974). Then these waypoints were clustered, after which journeys with the same sequence were grouped.

## 7.1 Route Clustering and Analysis

Each vessel (object) goes through a different process than each other vessel, changing its position in time and space, as well as changing other variables throughout time. This is a different process as opposed to a spatio-temporal point process, where variables change through space and time, according to one or more processes. In the moving object field however, each process is unique and never repeats itself exactly. Therefore it is considered sensible to group the processes based on their trajectories. This can be as detailed as to group them by their positions throughout time, their speed distribution, and other parameters.
However for the purpose of the data analysis - investigating what variables are suitable to predict vessel intent - it is desirable to perform grouping with as few assumptions regarding vessel intent as possible at first, but still capturing the intent of a vessel. Therefore the journeys are grouped by their routes; a route is defined as as a pair of spatial positions: the starting point and ending point of a journey. This means that journeys that have a similar starting point and also a similar ending point belong to the same group. The significance of this grouping is that it roughly captures the intent of a vessel (where it comes from and where it is going), with one single assumption that each vessel has a route in mind to travel. This contrasts with a full trajectory based grouping, which at least assumes a series of points intended, or even an entire path.
In this section, first the used clustering methods are discussed (section 7.1.1), then how these methods are applied to the route clustering in section 7.1.2, and finally what plots were generated (also section 7.1.3).

### 7.1.1 DBSCAN

Density Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) is a data clustering algorithm which clusters data based on the density of data points (in a euclidean space). This means that data points in high density areas are considered to belong to the same cluster as nearby points. Points that are in low density areas are considered to be noise. The advantage of this method is that it can handle clusters of any shape, including non-convex shapes. "The key idea", according to Ester et al. (1996), "is that for each point of a cluster the neighbourhood of a given radius has to contain at least a minimum number of points, i.e. the density in the neighbourhood has to exceed some threshold."
The given radius is denoted by $\epsilon$ or *eps*, while the minimum number of points is denoted by $MinPts$. These are the two parameters used by the algorithm. Core points are defined as points which contain $>= MinPts$ within their *eps* radius. Border points are defined as points that lie within the *eps* radius of a core point, but are not a core point themselves. This is illustrated in figure 7.1. All core points that lie within reach (*eps*) of each other are considered to be of the same cluster. All border points that belong to these core points also are part of that cluster. Points that are neither core points nor

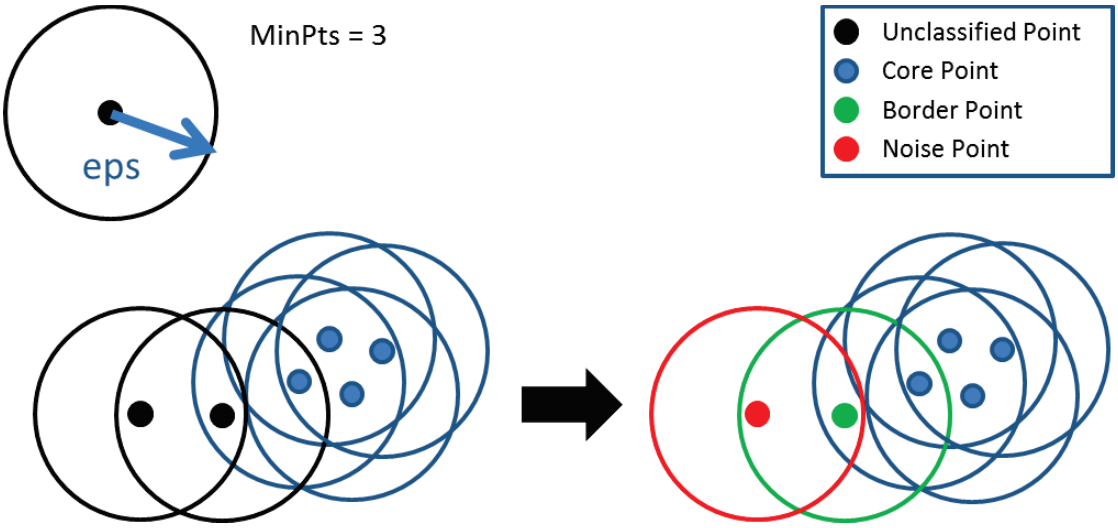



**Figure 7.1:** Illustration of the DBSCAN Ester et al. (1996) clustering principles. The left side shows the first step of the algorithm (determining core points), and the right side shows the second step (determining border points and noise points). The colours are explained in the figure.

border points are considered noise points. The algorithm first determines for each data point if it is a core point, ans subsequently determines which points (core points and other points) lie within range of each core point, which determines the clusters.
Figure 7.1 is used as an illustration of the concepts explained here. The minimum number of points used here is three, and the size of *eps* is indicated. On the left hand of the figure the first part of the algorithm is shown, where each of the blue points is determined to be a core point, since for each of them, there are 3 or more points within *eps* radius. In the second step (right), the green point is found to be a border point, since it is within reach of a core point. The red point, however is not within reach of a core point, and is therefore considered noise. All core points connected to each other are considered one cluster, together with their corresponding border points.
The now widely applied method (by Ester et al. (1996)) for determining the best *eps* value when $MinPts$ is given, is by determining the 'elbow' in a K Nearest Neighbour (kNN) distance plot fig. 7.2. This plot shows for each datapoint how large a radius (*eps*) is necessary to have $minPts$ number of points lie within this radius. The elbow in the plot determines the point where an increase in radius would result in relatively less information gain. All the data points to the left of the elbow are part of the clusters, and all the data points to the right are considered noise (they are too far away from the nearest neighbouring point in euclidean space).
In this research, the R package 'dbscan' (Hahsler, 2016) was used to implement this algorithm.

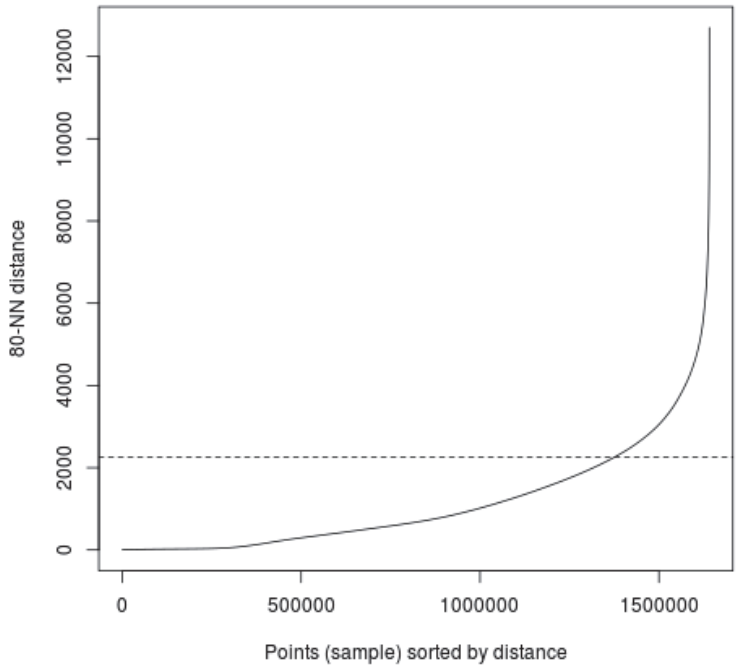



**Figure 7.2:** K Nearest Neighbour distribution plot.

### 7.1.2 Route Clustering

The grouping is done using the DBSCAN method described in 7.1.1. There are various methods to cluster data, but this algorithm has the advantages that it can handle clusters of arbitrary shapes, does not need a priori knowledge of the number of clusters, and is built for spatial data. The entry points (x and y position when a journey starts) of the journeys were clustered separately from the exit points (x and y position when a journey ends). The search for the best clustering parameters was done by ranging the *minPts* parameter from 60 to 120 in steps of 10, and determining *eps* for each *minPts* according to the method described in section 7.1.1. The elbow in the kNN-distance plot was determined by finding the point in the graph where the area under the plot was equal on either side of the point. The correctness of the value was verified visually and so the *eps* value was derived from this point.
For each of these parameter combinations, clustering was performed, resulting in a number of entry point clusters and exit point clusters. These were plotted on a map (e.g. figs. 8.1 and 8.2), to establish which parameter setting results in the most distinct entry and exit point clusters. The two criteria used for this were the relative amount of noise (20%) and if the points were representative of the TSS structure (thus meaningful).
To confirm the method of choosing *eps*, the above procedure was also performed for other *eps* values. These other values were obtained by multiplying the original *eps* value by 0.5, 0.8, 0.9, 1.1, 1.2 and 1.5. From this it was concluded that within a ten percent range of the original *eps* value, the TSS representation was very much the same, but the amount of noise varied significantly.

### 7.1.3 Plotting for Analysis

Routes were established by combining each entry cluster with each exit cluster, and for each route a spatial plot was created (e.g. fig. 7.3). Routes where less than 1 journey per two days pass (so less than 17 for the 34 day dataset), were regarded as noise.
The same plots as in chapter 6 were generated, but then coloured in based on the route, and grouped by entry point, in order to examine the differences between the routes, variable by variable. This was done in order to determine how well each variable can distinguish between routes, so as to be useful for the prediction of the vessel intent.

## 7.2 Waypoint clustering and Trajectory Analysis

After investigating the routes, the next step in the data analysis process was to dig deeper by separating planned path within route clusters. A planned path is here defined as a series of waypoints which a vessel has followed. Therefore, within one route (same starting point and end point) there are multiple planned paths that can be followed. The underlying assumption is that the traversed trajectory is a good approximation for the planned path. Different kinds of vessels, circumstances or other variables may be

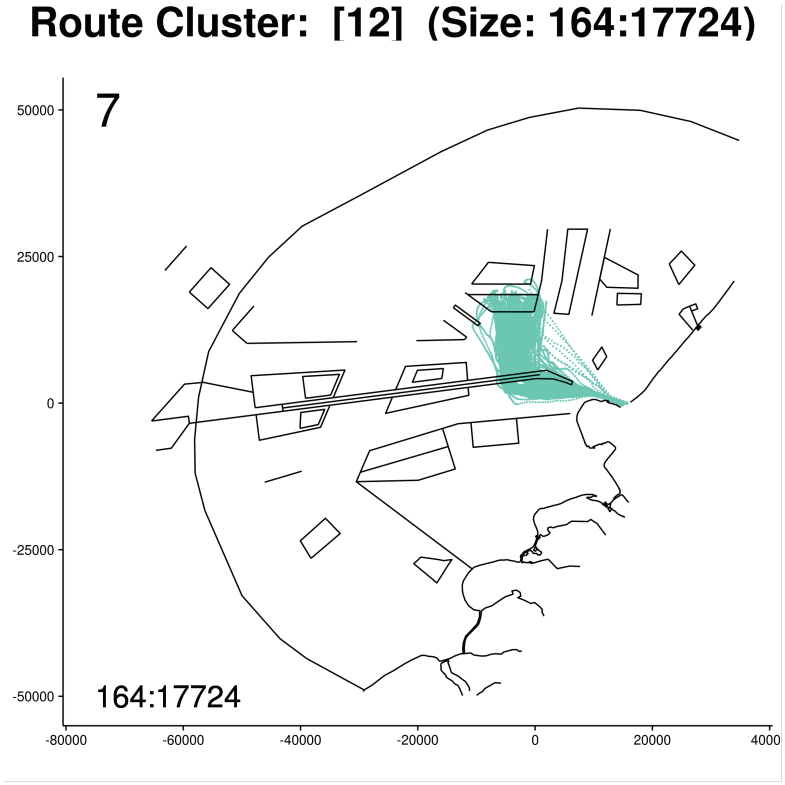



**Figure 7.3:** Spatial plot of route from entry point 12 to exit point 7. The size of this route is 164 journeys with 17,724 data samples.

the cause of this. The reason why this sub-grouping is done, is because it describes the intent of a vessel to more detail than that the route definition (section 7.1.2) does.
The method used to separate different trajectory groups is inspired by the approach used in Lancia et al. (2014). It may be noted, however, that within this thesis the approach is used solely to group data for the purpose of data analysis, not for the purpose of building a prediction model directly. The approach is to derive waypoints from vessel trajectories, to subsequently group trajectories by the series of waypoints they follow. First, the method used to derive the waypoints is described (section 7.2.1), after which its application in this research is described in section 7.2.2.

### 7.2.1 Change-point detection by binary segmentation

Change-point detection is an area of clustering algorithms that aims to detect behavioural changes in sequential data (i.e. clustering based on behavioural characteristics). Often this is applied to time-series, since this is a common application of sequential data. Changes of behaviour can be, for example, changes in mean, changes in variance, or changes in both. The most widely applied approach in this area is that of binary segmentation by Scott and Knott (1974).
To explain this algorithm, it is best to start at how a single change-point is determined in a data series. Consider a series of values $y_1, y_2, \ldots, y_n$ which may or may not have two distinct means. Figure 7.4, for example, is a series of sequential data which seems to have two distinct means. For one point $k \in \{1..n\}$, the mean $\mu_1$ and standard deviation $\sigma_1$ of the values before $y_k$ are calculated, as also for the values after $y_k$ ($\mu_2$ and $\sigma_2$). Based on these values, the likelihood $L$ is determined that the two means $(\mu_1, \mu_2)$ are distinct ($L$ increases if $\sigma$ decreases and $\mu_1 - \mu_2$ increases). This is done for every data point $k \in \{1..n\}$. Then the data point with the highest likelihood is regarded as the change-point IF the likelihood exceeds the likelihood that there is only one mean.
In the case of multiple change-point detection, binary segmentation first applies the above approach to the entire dataset. Then, IF it detects a change point, the dataset is split into two parts at this point, and for both parts, the same approach is applied again. This is done recursively until each likelihood comparison test determines that there is no change-point. This can be seen as analogous to the well-known bisection method.
In this research, the R package 'changepoint' (Killick and Eckley, 2014) was used to implement this algorithm.

### 7.2.2 Procedure

All the journeys within one route cluster are considered. For each journey, the change-point detection method described in section 7.2.1 is used to detect changes in the course of a vessel. The change-points detected are considered the journey's waypoints. The positions of these waypoints are clustered using the DBSCAN method described in section 7.1.1. Then, for each journey the sequence of waypoint clusters that the vessel follows is registered. All journeys with the same sequence are considered to be of the

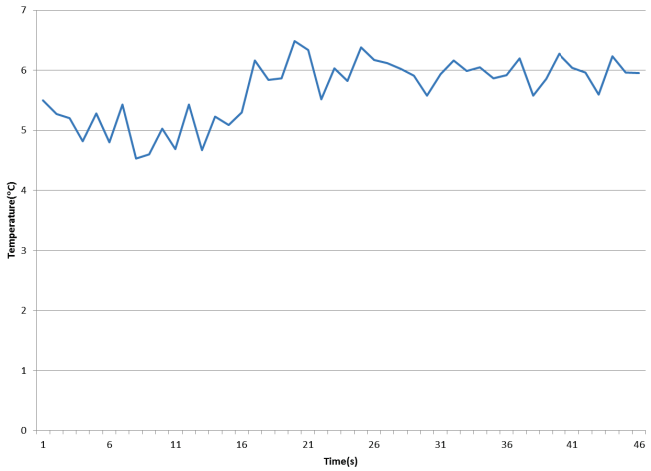



**Figure 7.4:** Sequential dataset with (possibly) two means.

same trajectory group.
The reasons for using course change as a waypoint -change in behaviour- indicator are twofold. First, when a ship-master plans his trip from harbour to harbour, this is done with a series of waypoints in space, which is where he plans to change course to head towards the next waypoint. After this he plans at what time to be at each waypoint, from which he builds a rough speed plan. Therefore there is an assumed underlying intent of the vessel captured in the (discrete) course profile and the speed profile of a journey, though the former is the stronger. Second, the change in course is much more distinct than the change in speed. Throughout journeys the change in course is more concentrated short periods of time, while the change of speed is spread out over longer periods of time (slow acceleration/deceleration). This can also be seen in fig. 7.5, where the change in course is more distinct (bottom left) than the gradual change in speed (bottom right.) The figure also displays the trajectory (top right), and the means estimated by the change-point detection method (bottom left). Naturally, the course variable has been treated as a continuous circular domain.

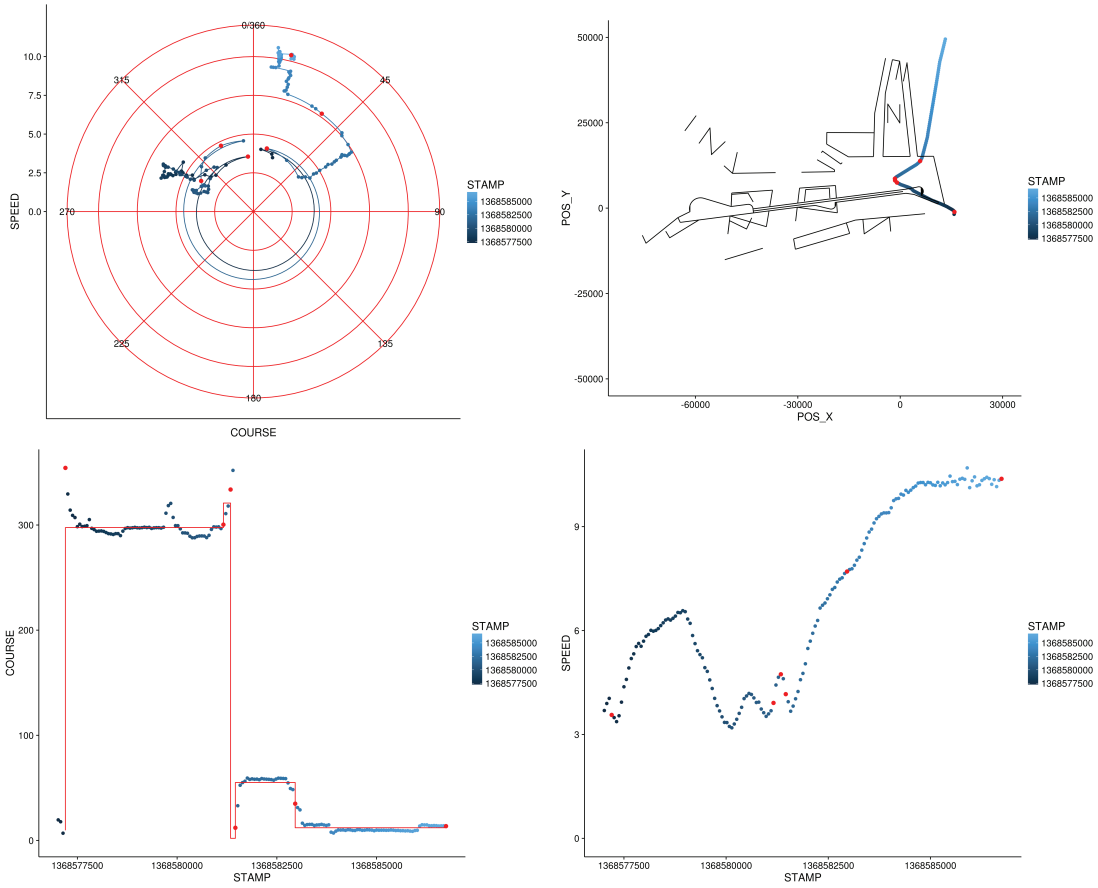



**Figure 7.5:** Illustration of waypoint detection. The red points indicate the datapoint corresponding to the waypoint. The red lines (bottom left) indicate the mean course estimated by the change-point detection method.

# Chapter 8

# Results

In this chapter the results of the data analysis are displayed, explained and discussed. First it is shown how well each variable can distinguish between routes, possibly in combination with other variables. The results show that the variables 'course' and 'destination' can distinguish between routes sufficiently enough to investigate them further. Other variables may add value also, but then in combination with the two mentioned variables. Then the waypoint determination (to derive planned paths) is shown to not be successful.

## 8.1 Route Clustering

The resulting parameters used are $minPts = 80$, $eps = 2197m$ for entry points, and $eps = 2252m$ for exit points. The entry point clusters and exit point clusters are shown on the map in figs. 8.1 and 8.2 respectively. Each of the clusters has been named according to its approximate location. There are 15 entry point clusters and 14 exit point clusters in total. It can be seen that the noise points concentrate themselves around the edge of the map, and around anchorage areas. The reason why some anchorage areas have not been identified as an entry/exit cluster is because of the relatively low number of points within that area.
The routes that connect the entry points to the exit points are shown in table 8.1. The table shows the number of journeys for each route. Each entry/exit point cluster is numbered as obtained from the clustering algorithm. From the number of journeys in the noise routes it can be concluded that in most groups, the noise route does not dominate. However, it can also be seen that for some entry points the vast majority of journeys heads towards a specific exit point. This means that if the entry point is known, the predictability for where the vessel will exit is high. On the whole, the largest routes account for 56% of the journeys.
The spatial plots, grouped by entry point, can be seen in figs. 8.3 to 8.18. For all these plots, the exit point clusters are colour coded (e.g. exit point 7 (Rotterdam) is cyan). The red journeys belong to the noise exit cluster, which is why they do not follow a clear pattern. All other routes, however, mostly have a very clear spatial pattern (behaviour). However, sometimes distinctly different paths are taken by different vessels.

### Route relevance

Not all the found routes are as relevant to the research. Basically any route that crosses through one of the two main intersections is of interest, since the most near-misses are encountered there. Routes that remain at the edges of the coverage area are not of interest, e.g. from entry point 2 (North Hinder S) to exit point 4 (Brugge/Westkapelle) in fig. 8.4. However other routes may originate from the same entry point, which is why such routes cannot be excluded from the analysis. Only if all the routes from the same entry point remain on the edges of the coverage area, can they be excluded. This is the case for entry point clusters 6 (North Hinder N, fig. 8.8), 15 (Katwijk, fig. 8.17) and -except for one noise journey- 14 (Southwest, fig. 8.16).

## 8.2 Variable Suitability

For each group of routes (grouped by entry points), a plot was created for each variable, and for each pair of variables, as described in section 7.1.2. Here, however, the datapoints are coloured by route. For each of these plots, it was evaluated if and how well the variable plotted made a distinction between the routes within the group. As a useful

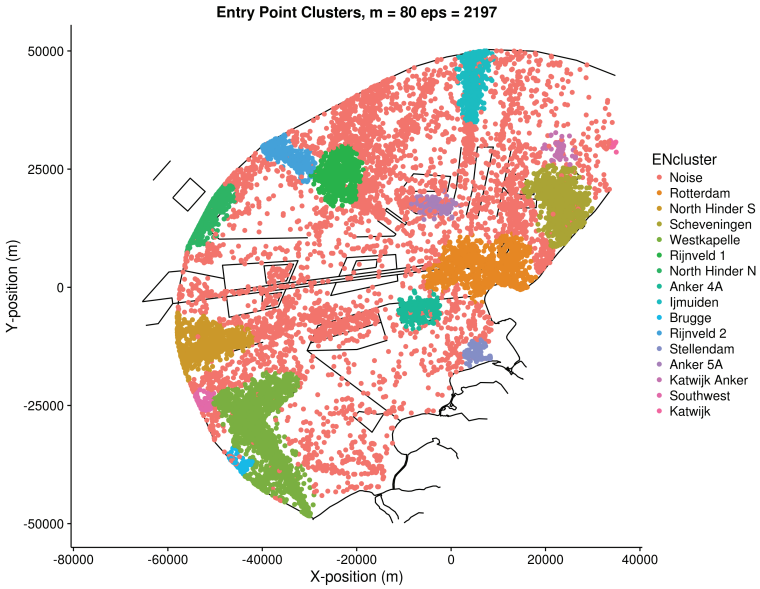



**Figure 8.1:** Resulting locations of the entry point clusters. Each dot represents the entry point of one journey.

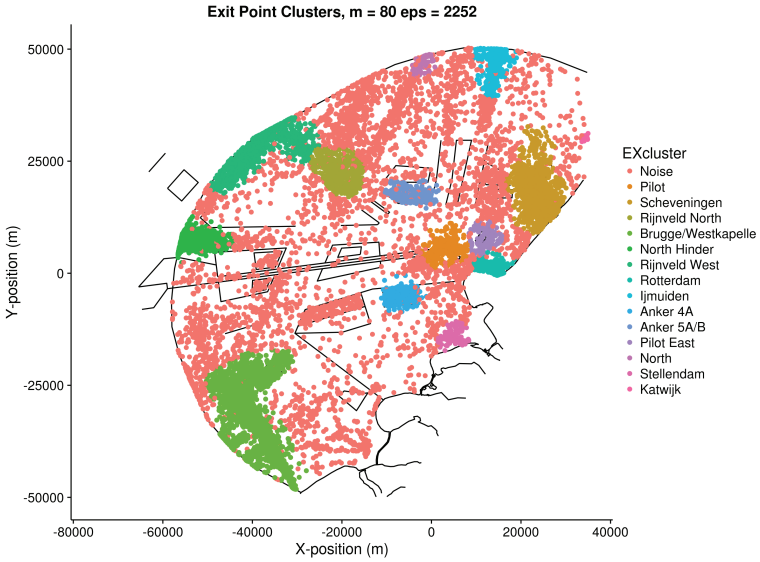



**Figure 8.2:** Resulting locations clusters of the exit points. Each dot represents the exit point of one journey.

| | Exit Point | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Entry Point | Name | Noise | Pilot | Scheveningen | Rijnveld North | Brugge/Westkapelle | North Hinder | Rijnveld West | Rotterdam | Ijmuiden | Anker 4A | Anker 5A/B | Pilot East | North | Stellendam | Katwijk | # Routes | # Journeys | # Samples |
| 0 | Noise | 2223 | 24 | 200 | 77 | 399 | 103 | 140 | 807 | 94 | 49 | 85 | 25 | 60 | 100 | | 14 | 4386 | 391325 |
| 1 | Rotterdam | 790 | 863 | 144 | 129 | 151 | 912 | 323 | 1414 | 540 | 25 | 75 | 170 | | 58 | | 13 | 5594 | 650443 |
| 2 | North Hinder S | 356 | | | | 39 | | | 756 | | 154 | | | | | | 4 | 1305 | 199641 |
| 3 | Scheveningen | 208 | | 1940 | | | | | 29 | | | | | | 34 | | 4 | 2211 | 115124 |
| 4 | Westkapelle | 396 | | | 177 | 3100 | | | 76 | | 34 | | | 40 | | | 6 | 3823 | 214781 |
| 5 | Rijnveld 1 | 107 | | | 56 | 269 | | 22 | 84 | | | 20 | | | | | 6 | 558 | 69180 |
| 6 | North Hinder N | 91 | | | | | | 452 | | | | | | | | | 2 | 543 | 18643 |
| 7 | Anker 4A | 21 | | | | | | | 240 | | 30 | | | | | | 3 | 291 | 26264 |
| 8 | Ijmuiden | 174 | | | | | | | 266 | | | 86 | | | | | 3 | 526 | 62741 |
| 9 | Brugge | 55 | | | | 111 | | | | | | | | | | | 2 | 166 | 11694 |
| 10 | Rijnveld 2 | 29 | | | 21 | | | 65 | 194 | | | 17 | | | | | 5 | 326 | 31268 |
| 11 | Stellendam | 97 | | 18 | | | | | | | | | | | 111 | | 3 | 226 | 30159 |
| 12 | Anker 5A | | | | | | | | 164 | | | | | | | | 1 | 164 | 17724 |
| 13 | Katwijk Anker | 22 | | 108 | | | | | | | | | | | | 86 | 3 | 216 | 10868 |
| 14 | Southwest | 21 | | | | 72 | | | | | | | | | | | 2 | 93 | 4102 |
| 15 | Katwijk | | 70 | | | | | | | | | | | | | 17 | 2 | 87 | 4799 |

**Table 8.1:** Number of journeys per route. Entry points are per row, exit points per column.

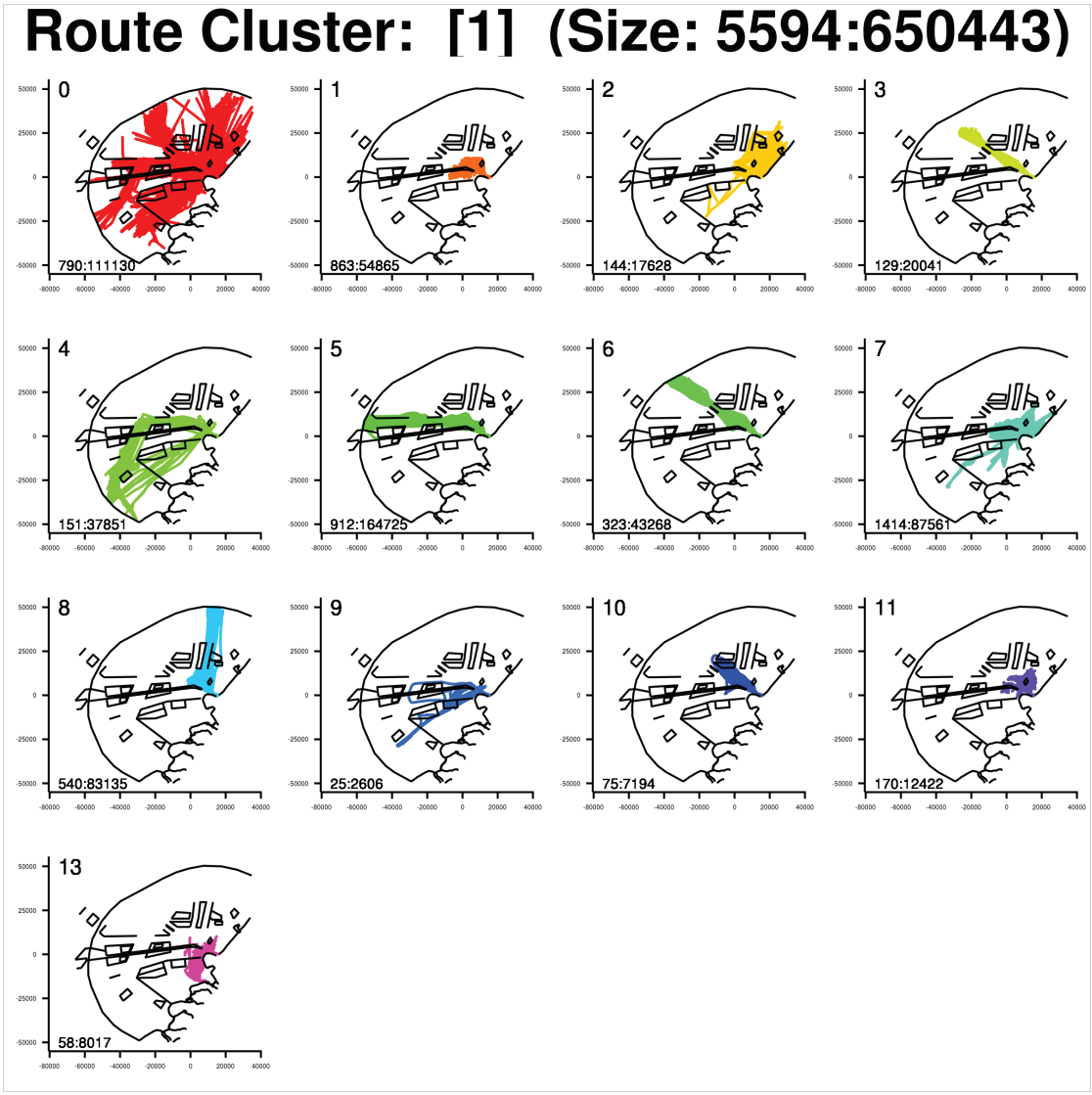



**Figure 8.3:** Spatial plot of all routes originating from entry cluster 1 (Rotterdam). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

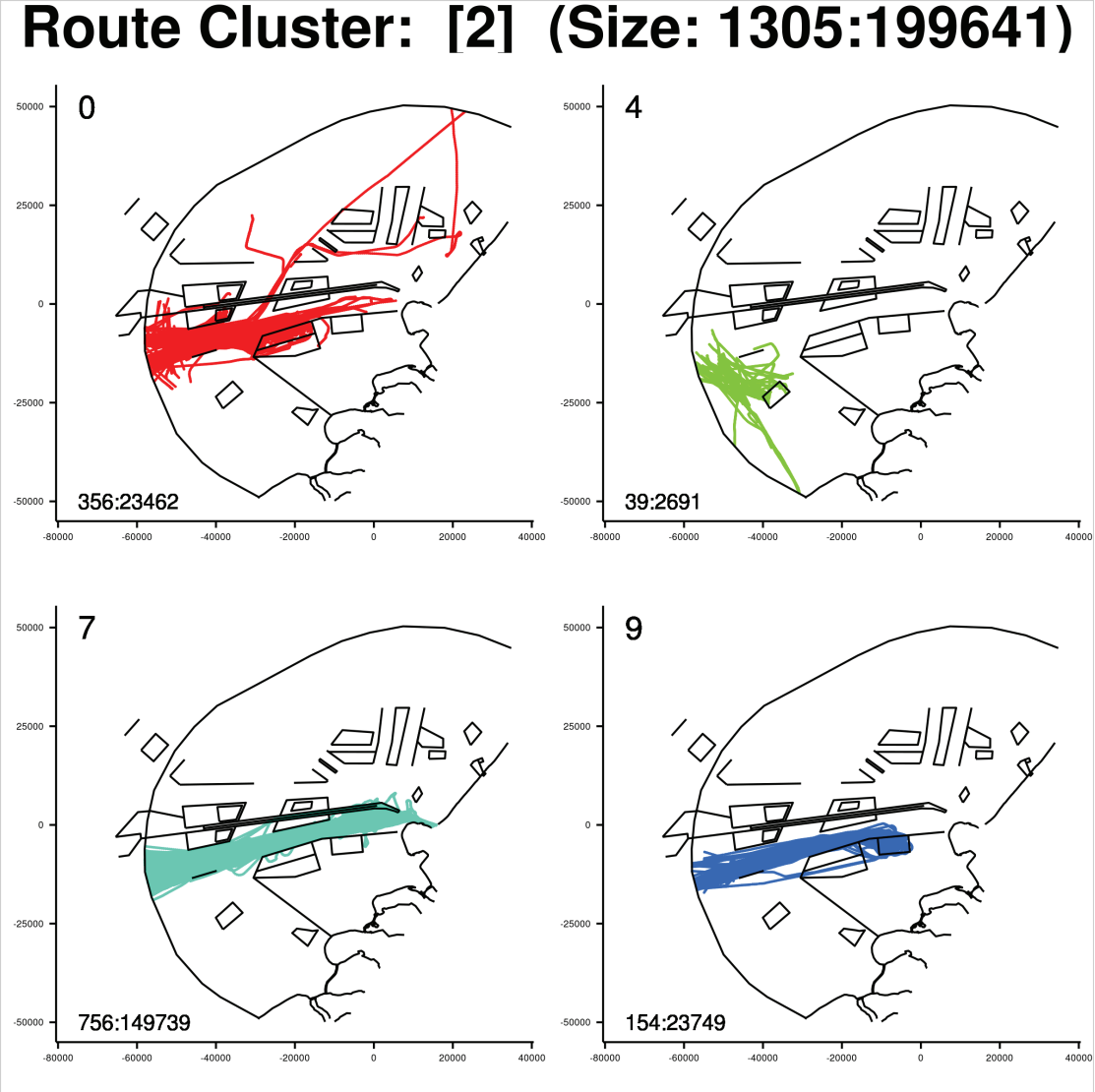



**Figure 8.4:** Spatial plot of all routes originating from entry cluster 2 (North Hinder S). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

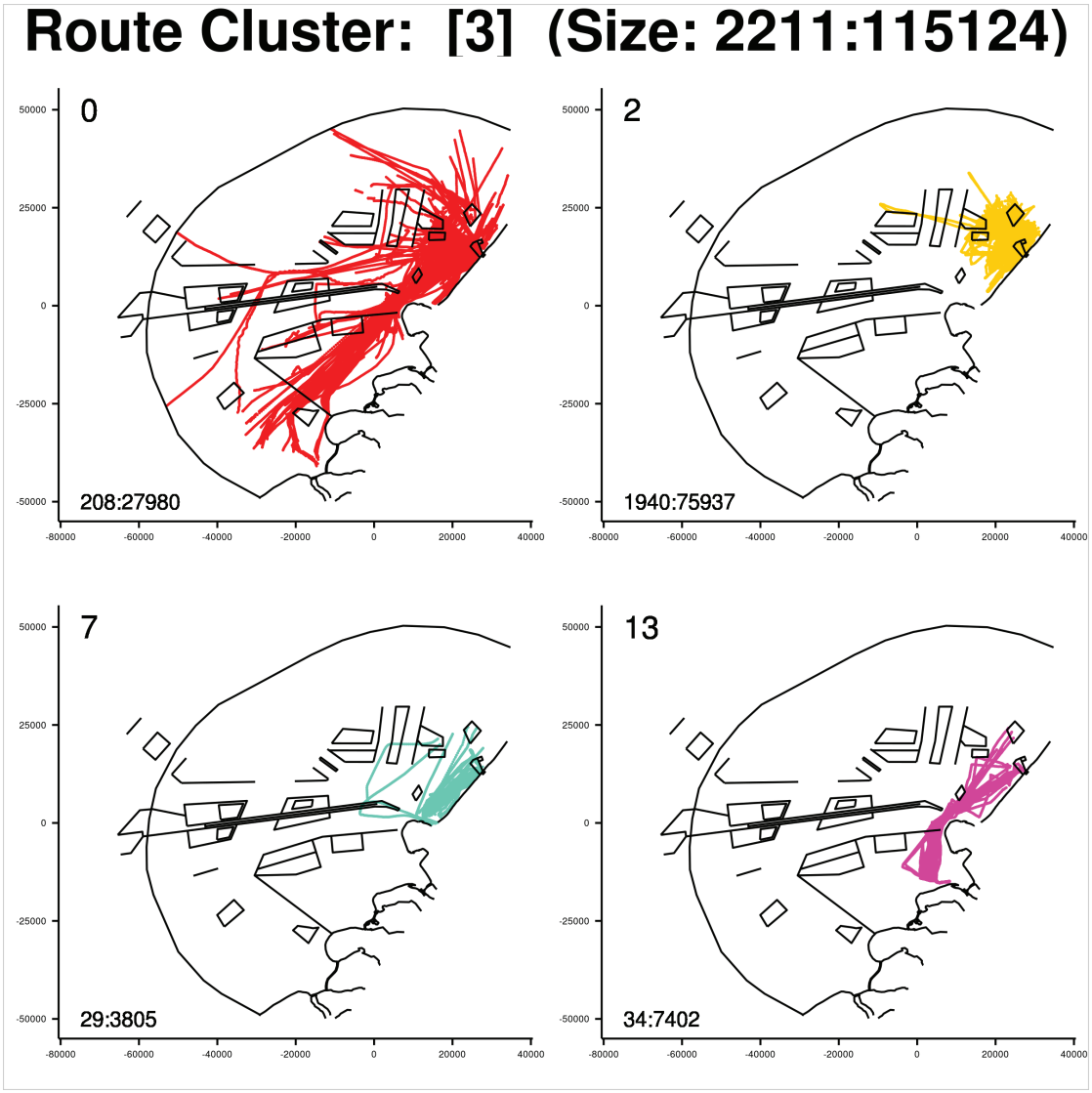



**Figure 8.5:** Spatial plot of all routes originating from entry cluster 3 (Scheveningen). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

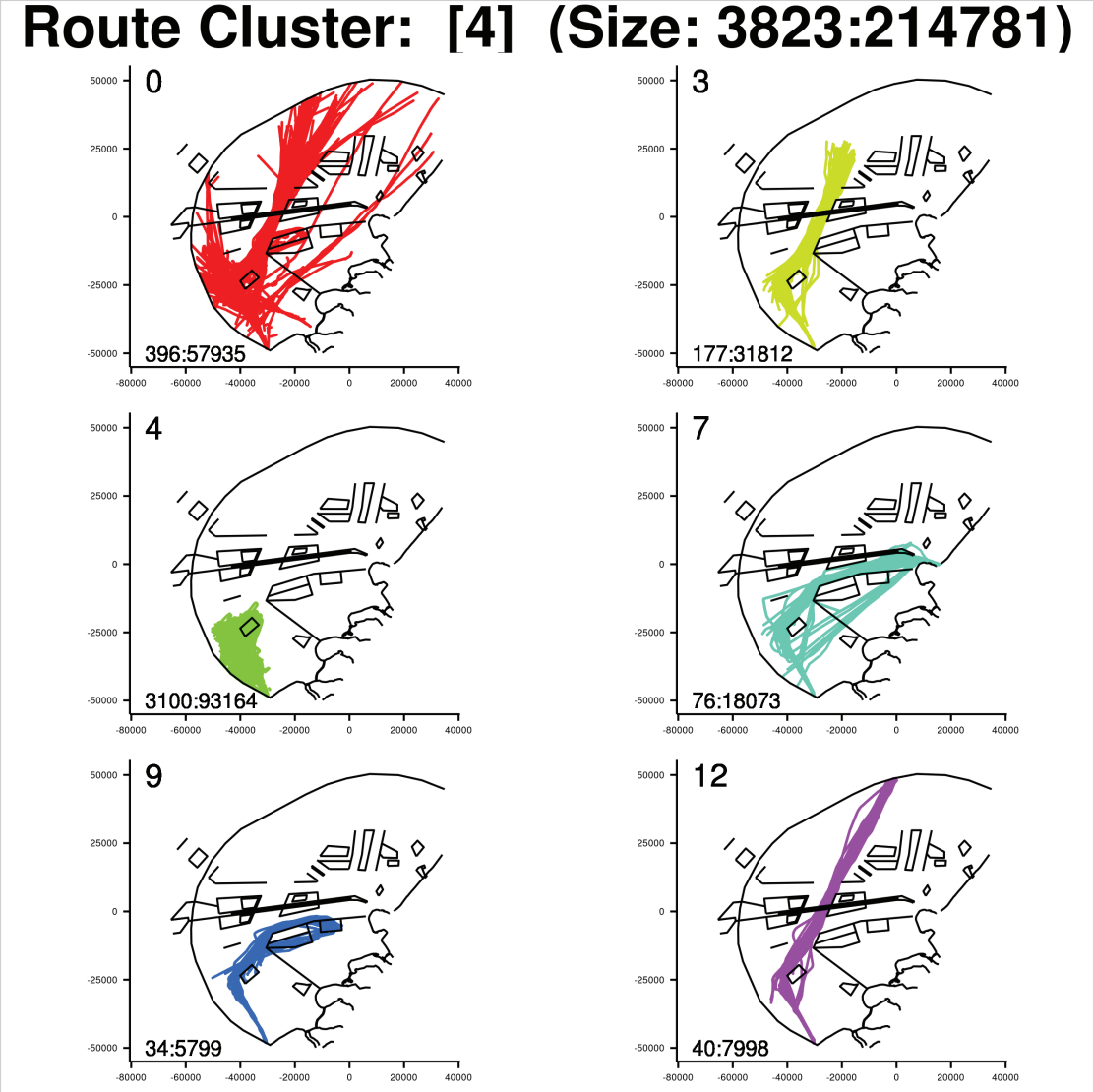



**Figure 8.6:** Spatial plot of all routes originating from entry cluster 4 (Westkapelle). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

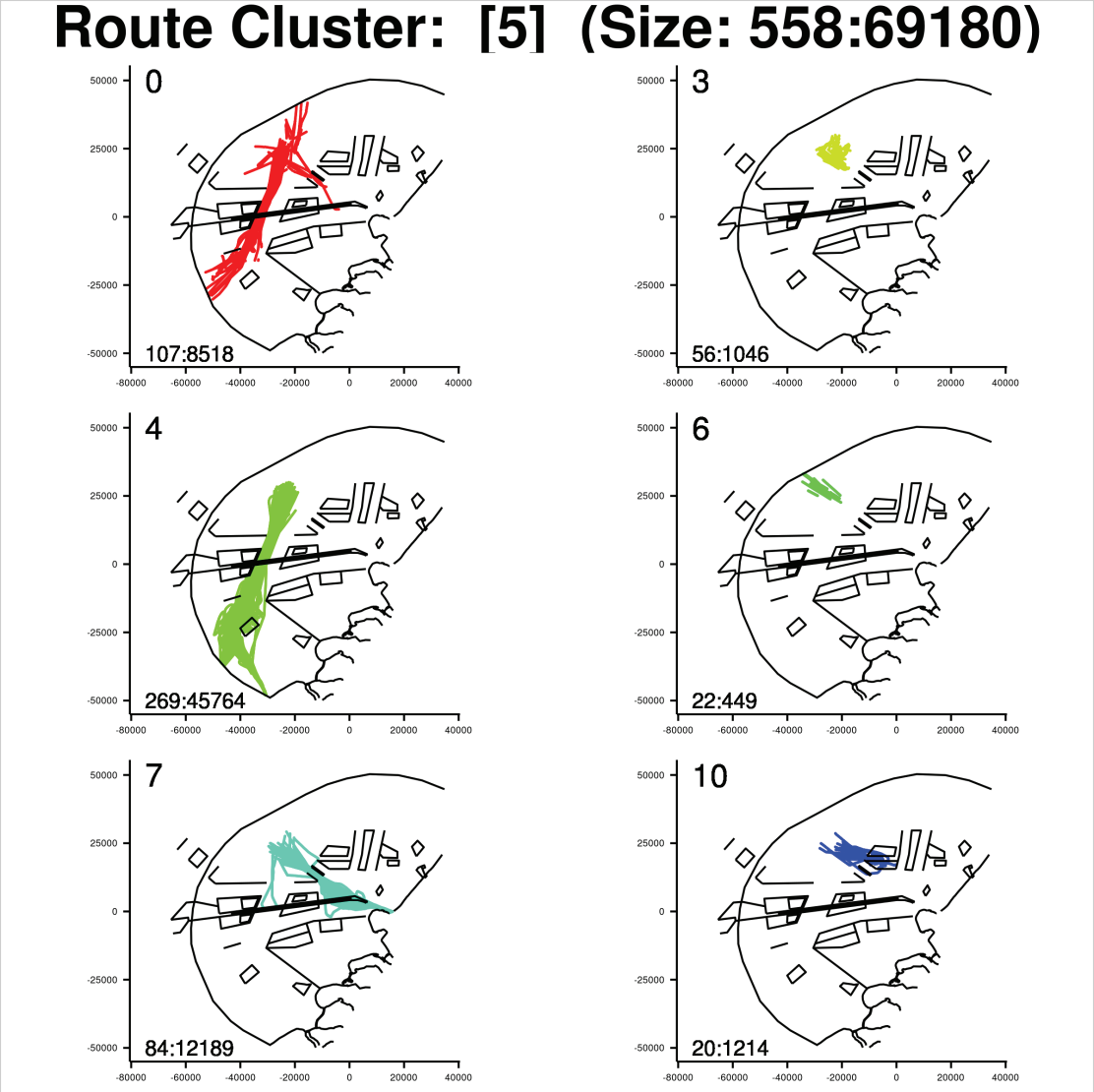



**Figure 8.7:** Spatial plot of all routes originating from entry cluster 5 (Rijnveld 1). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

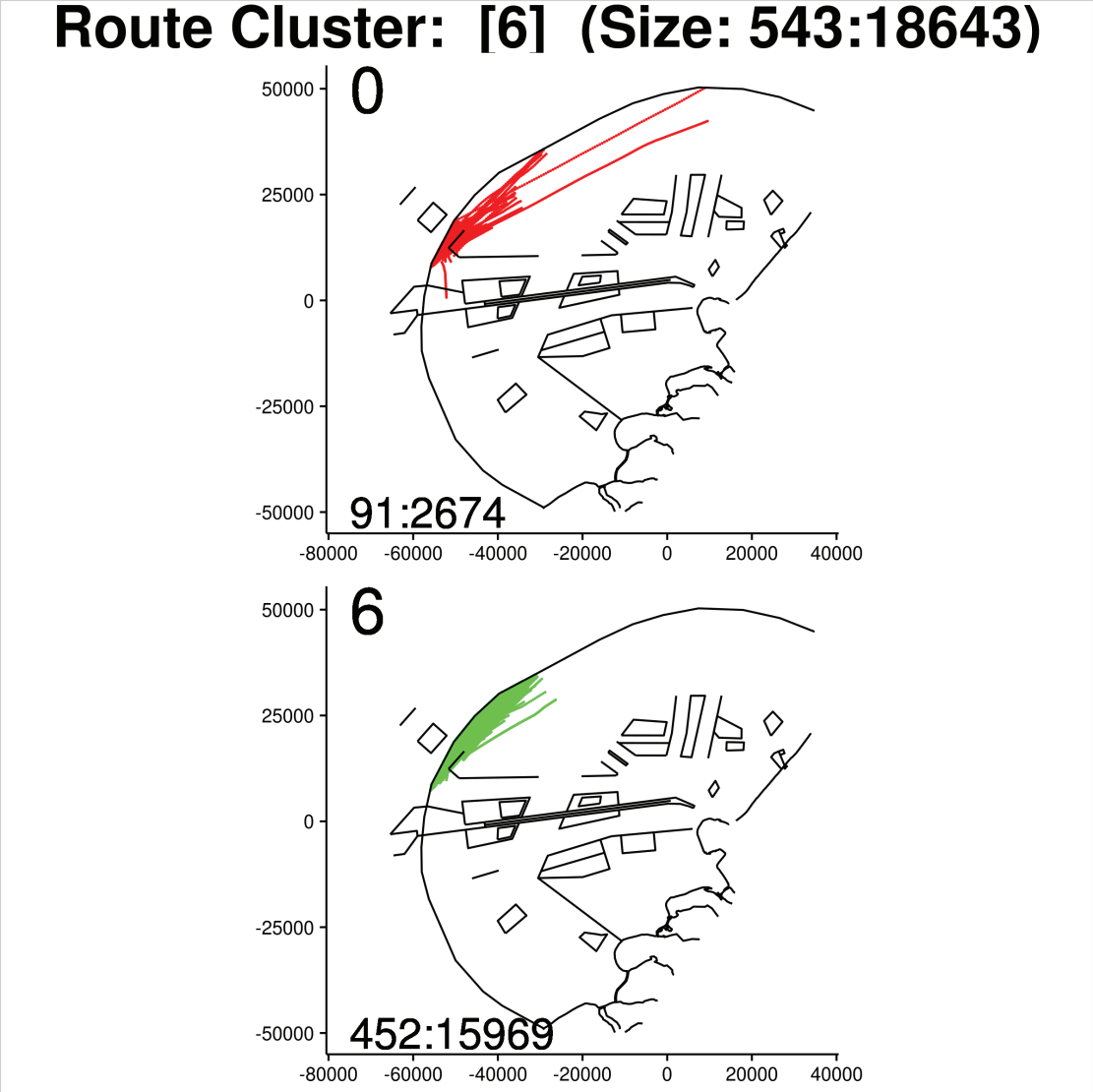



**Figure 8.8:** Spatial plot of all routes originating from entry cluster 6 (North Hinder N). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

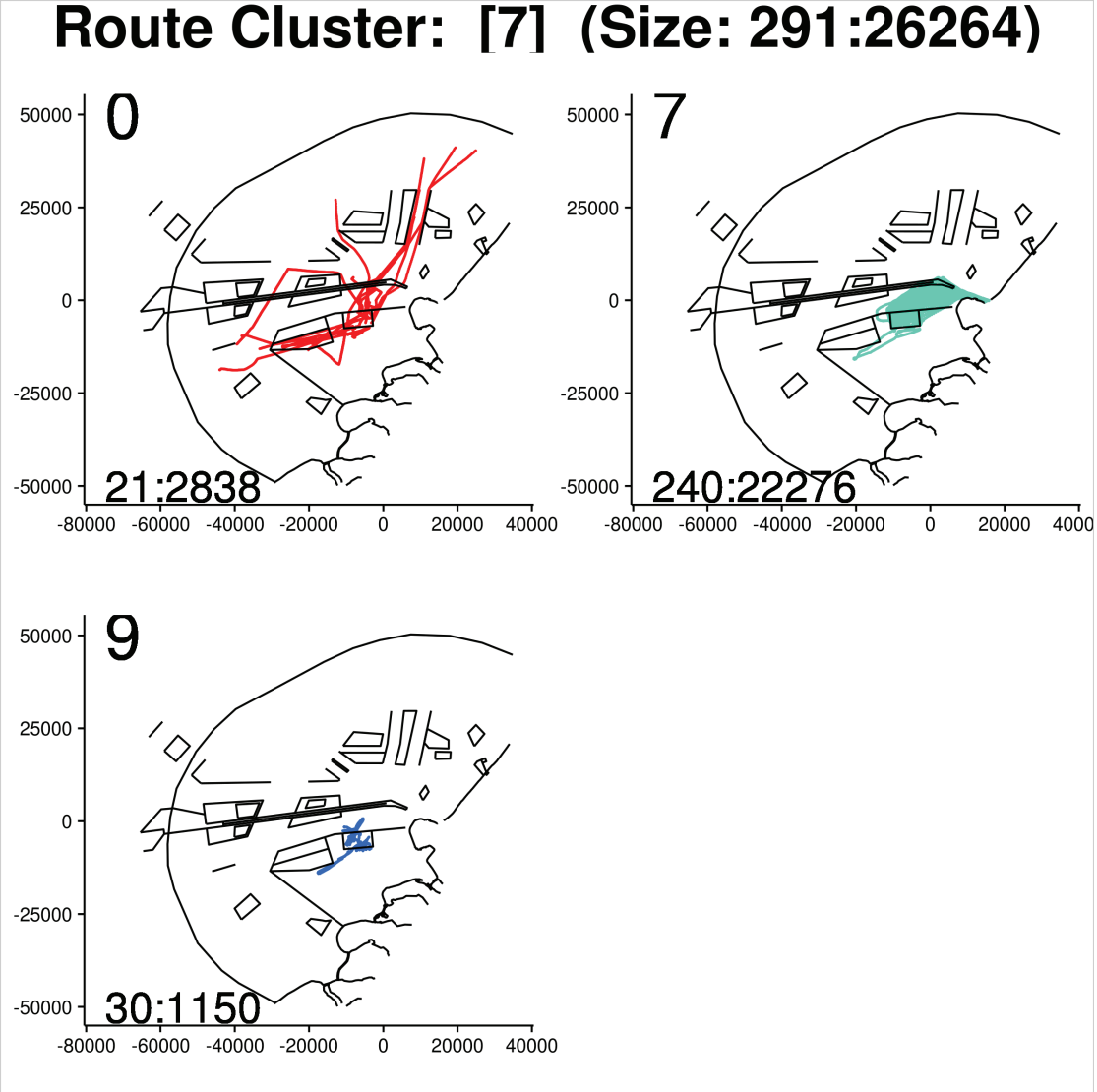



**Figure 8.9:** Spatial plot of all routes originating from entry cluster 7 (Anker 4A). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

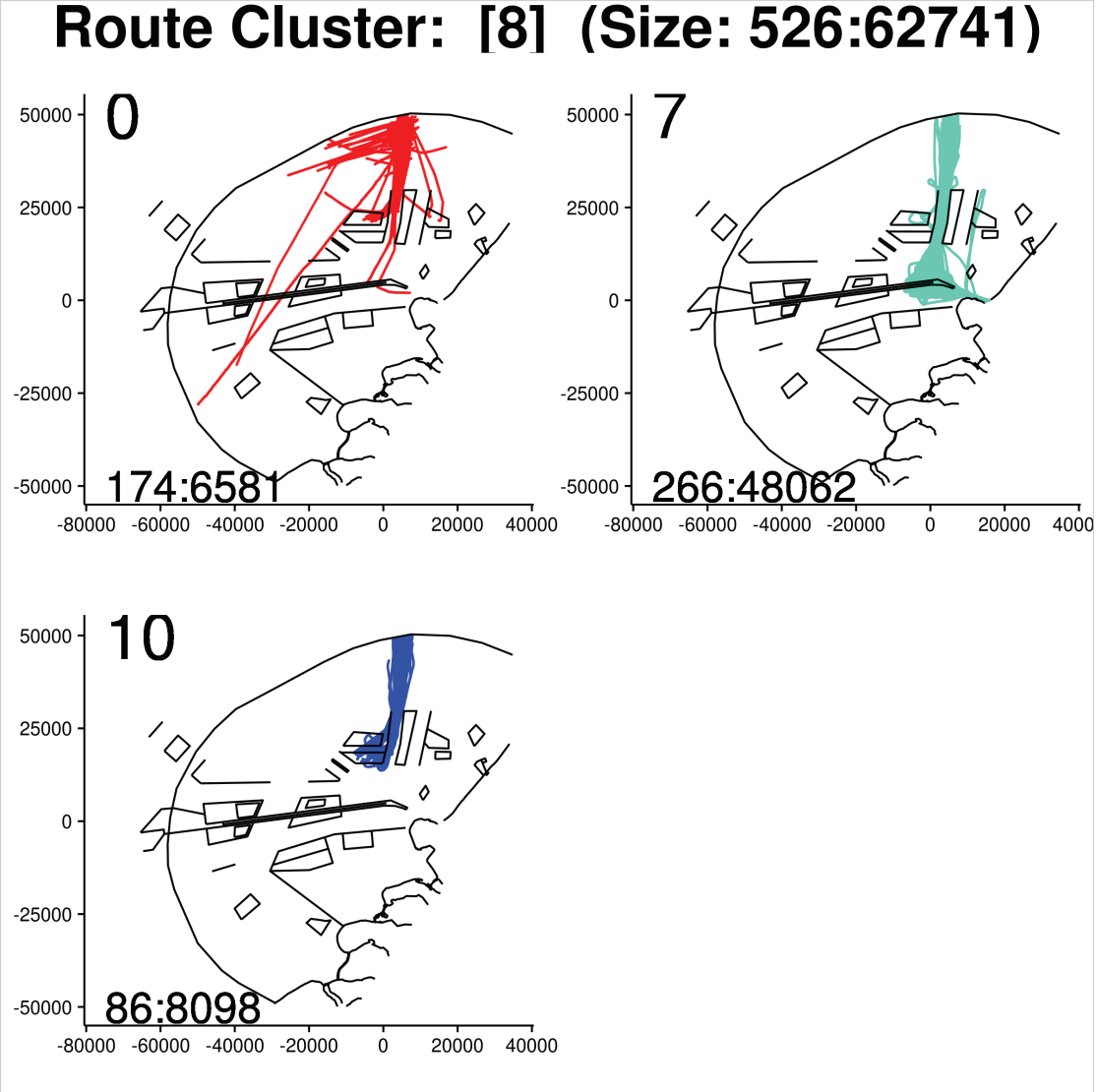



**Figure 8.10:** Spatial plot of all routes originating from entry cluster 8 (Ijmuiden). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

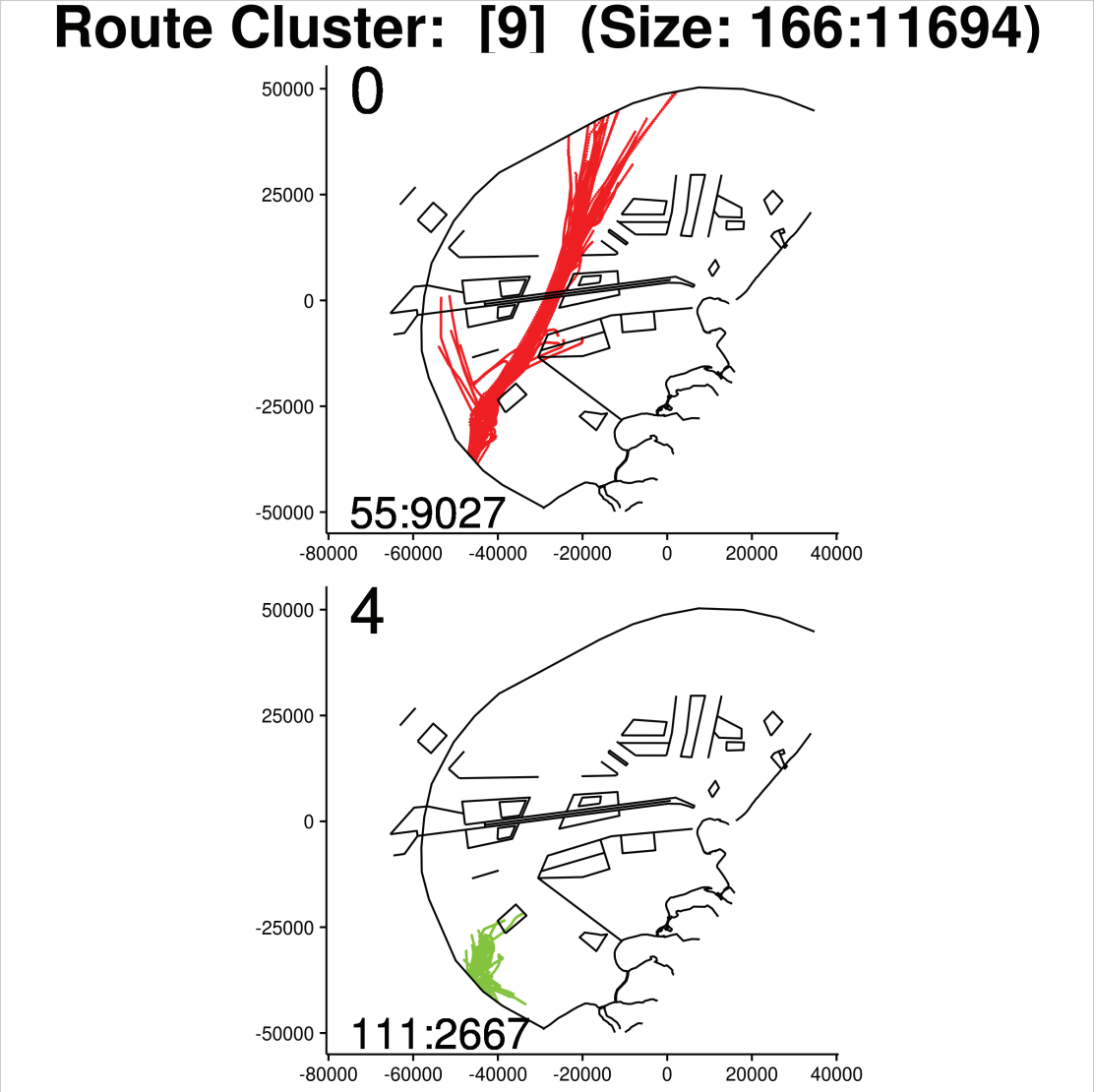



**Figure 8.11:** Spatial plot of all routes originating from entry cluster 9 (Brugge). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

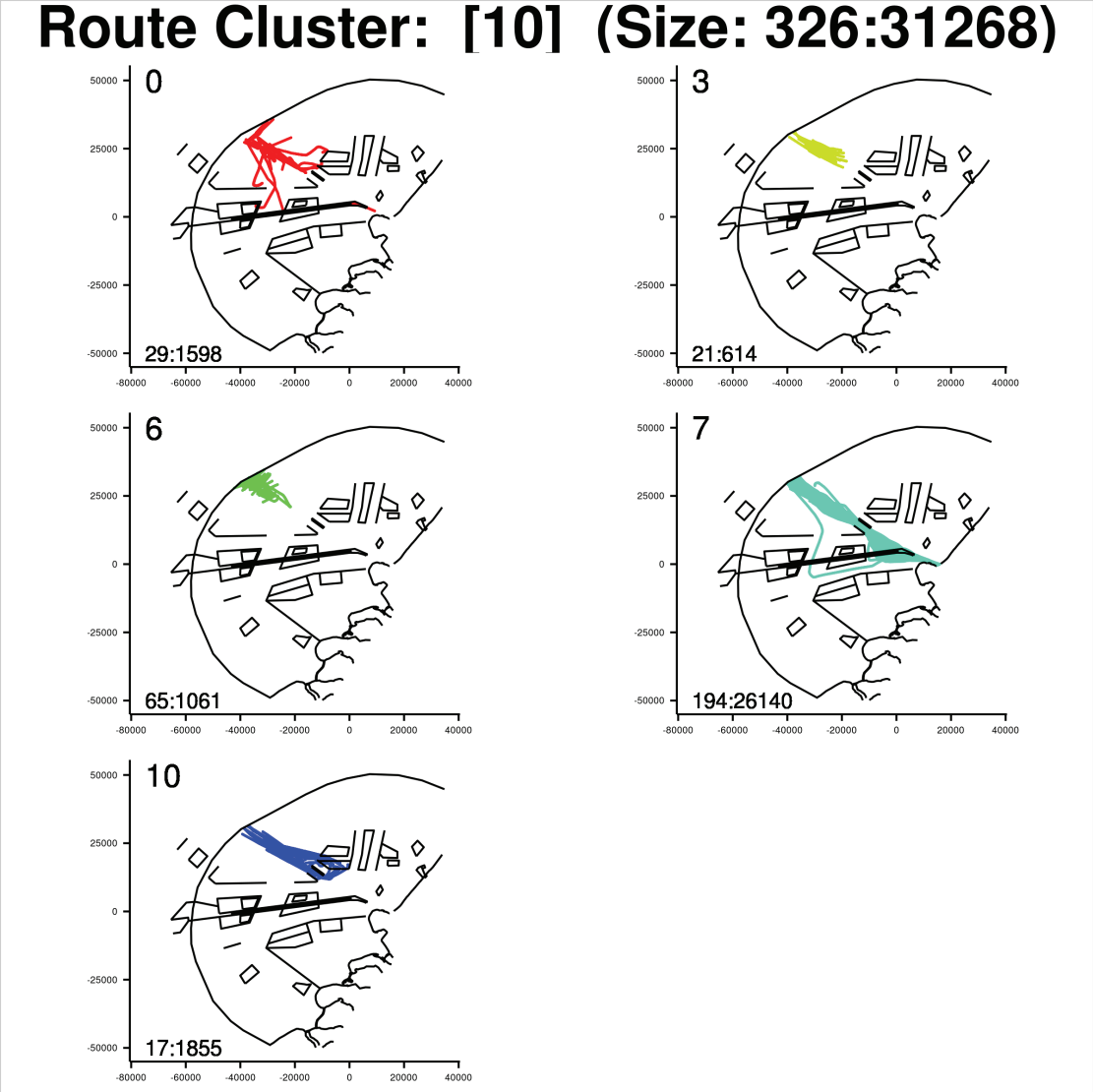



**Figure 8.12:** Spatial plot of all routes originating from entry cluster 10 (Rijnveld 2). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

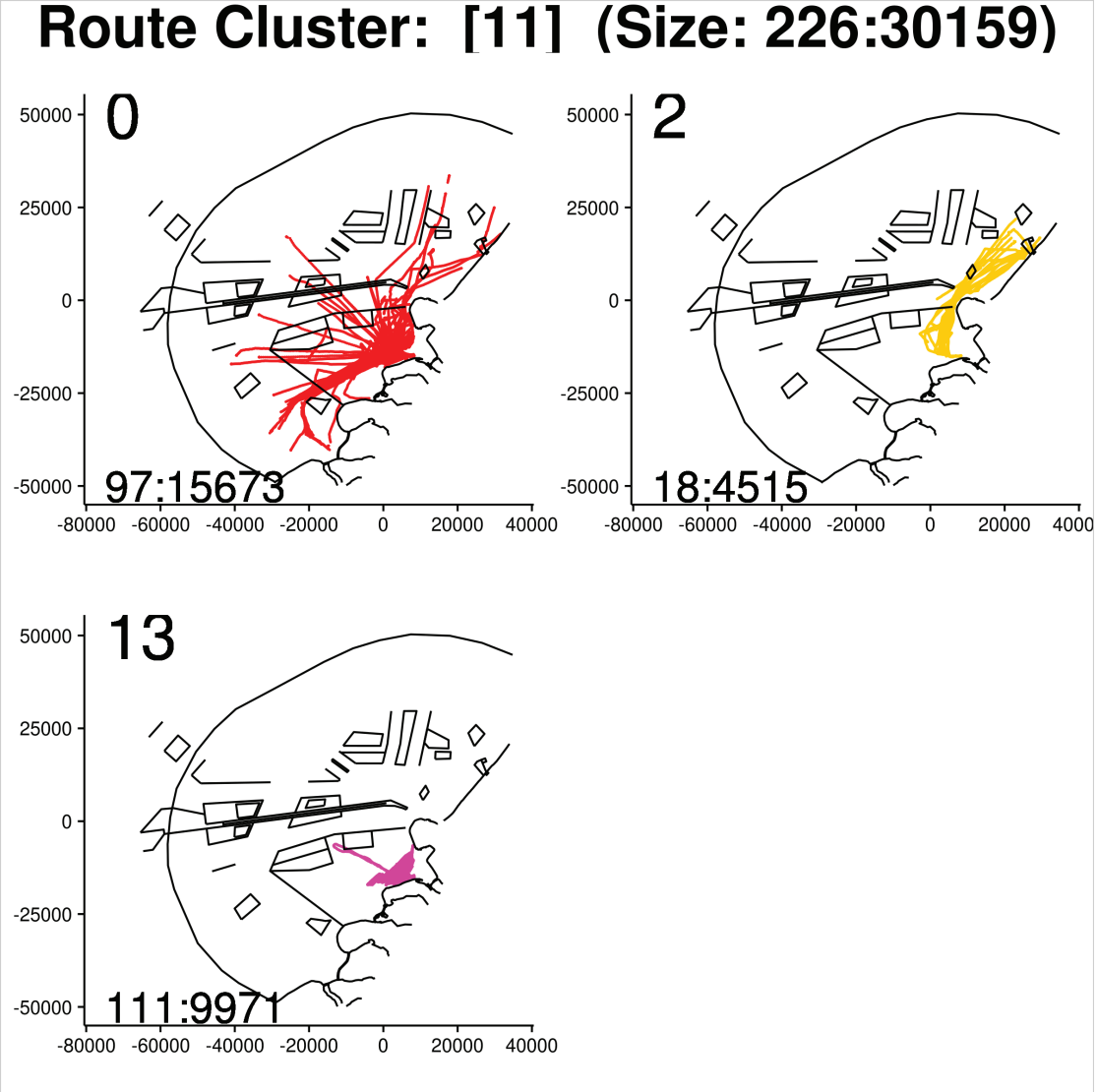



**Figure 8.13:** Spatial plot of all routes originating from entry cluster 11 (Stellendam). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

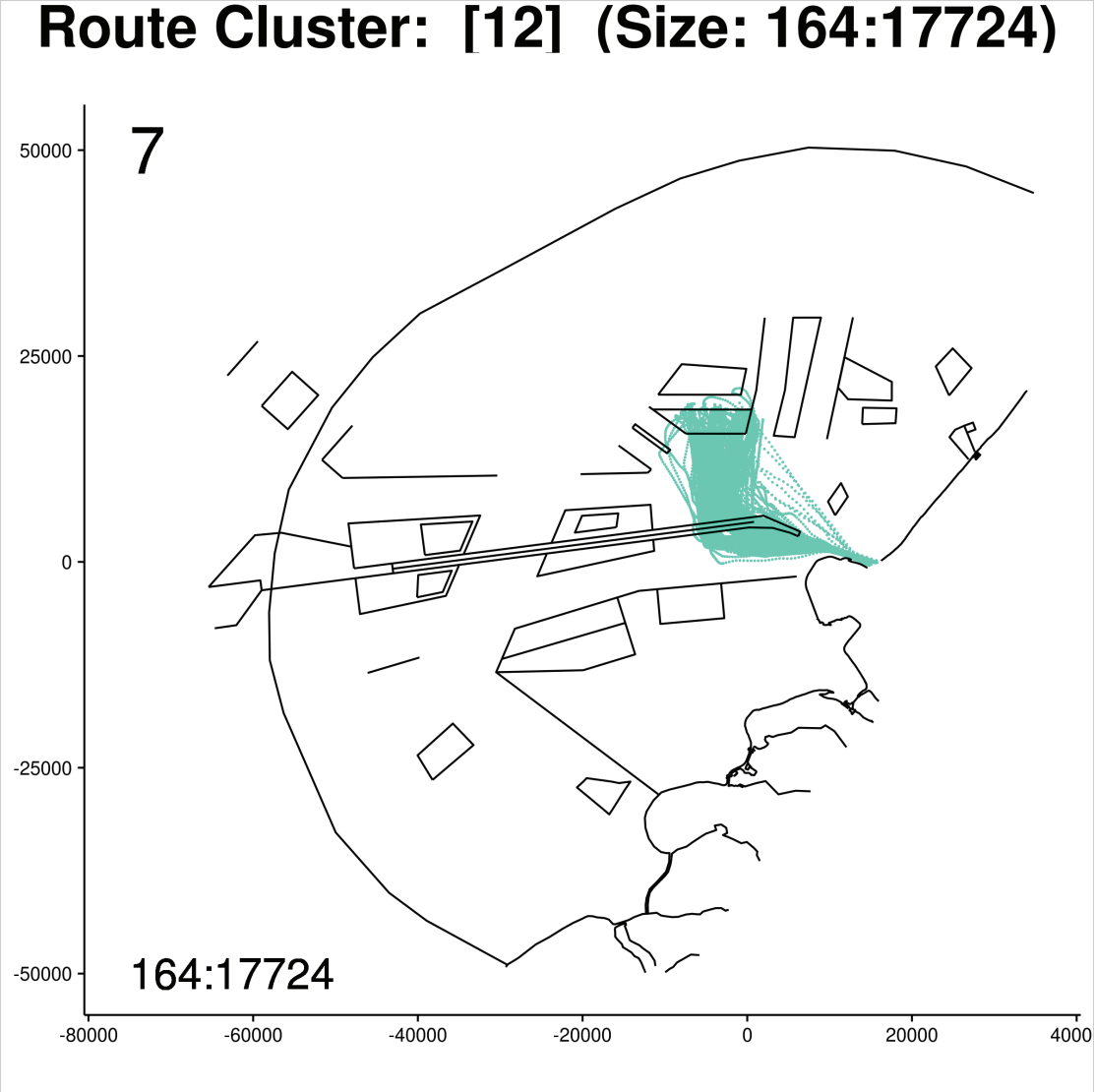



**Figure 8.14:** Spatial plot of all routes originating from entry cluster 12 (Anker 5A). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

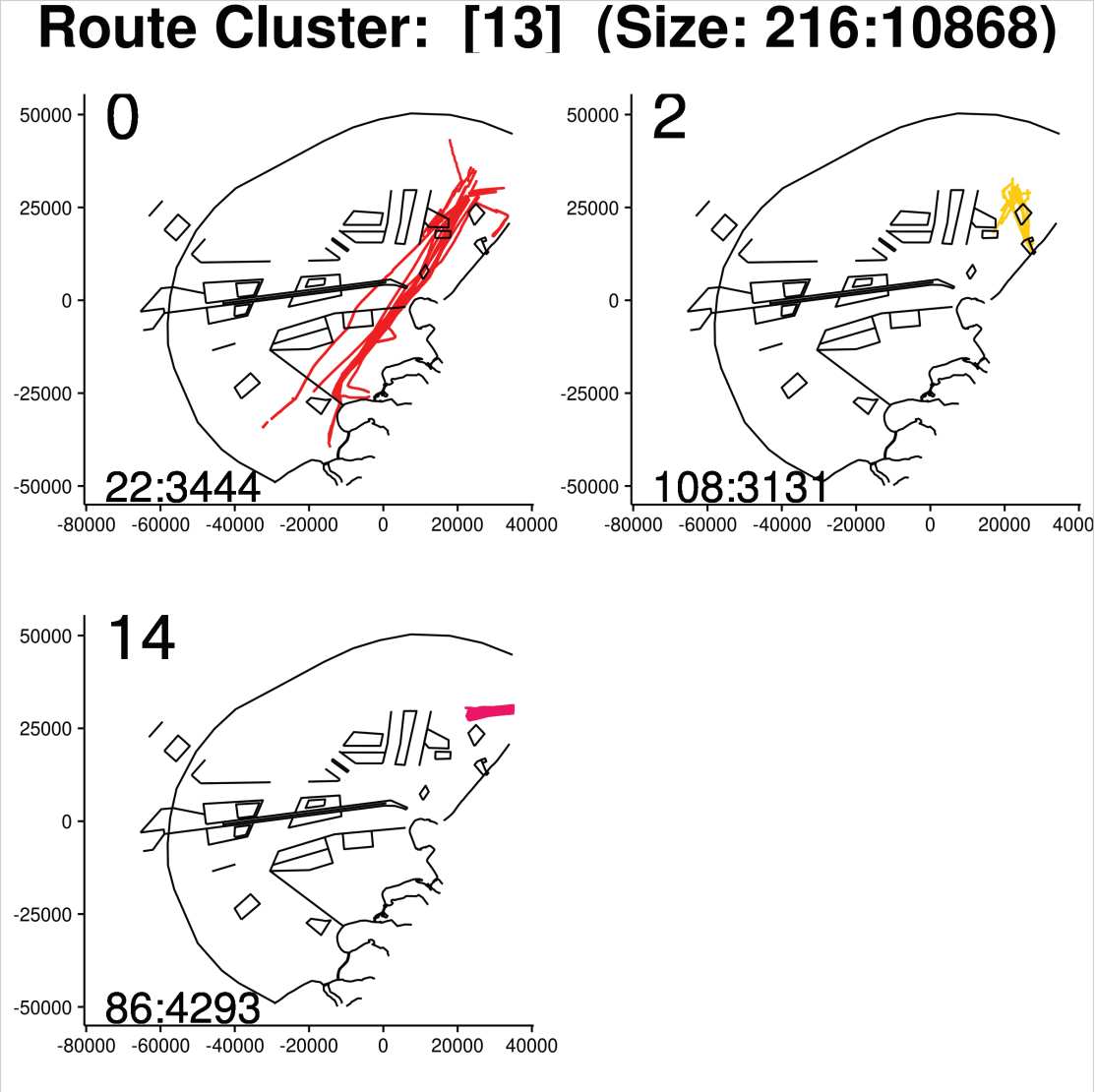



**Figure 8.15:** Spatial plot of all routes originating from entry cluster 13 (Katwijk Anker). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

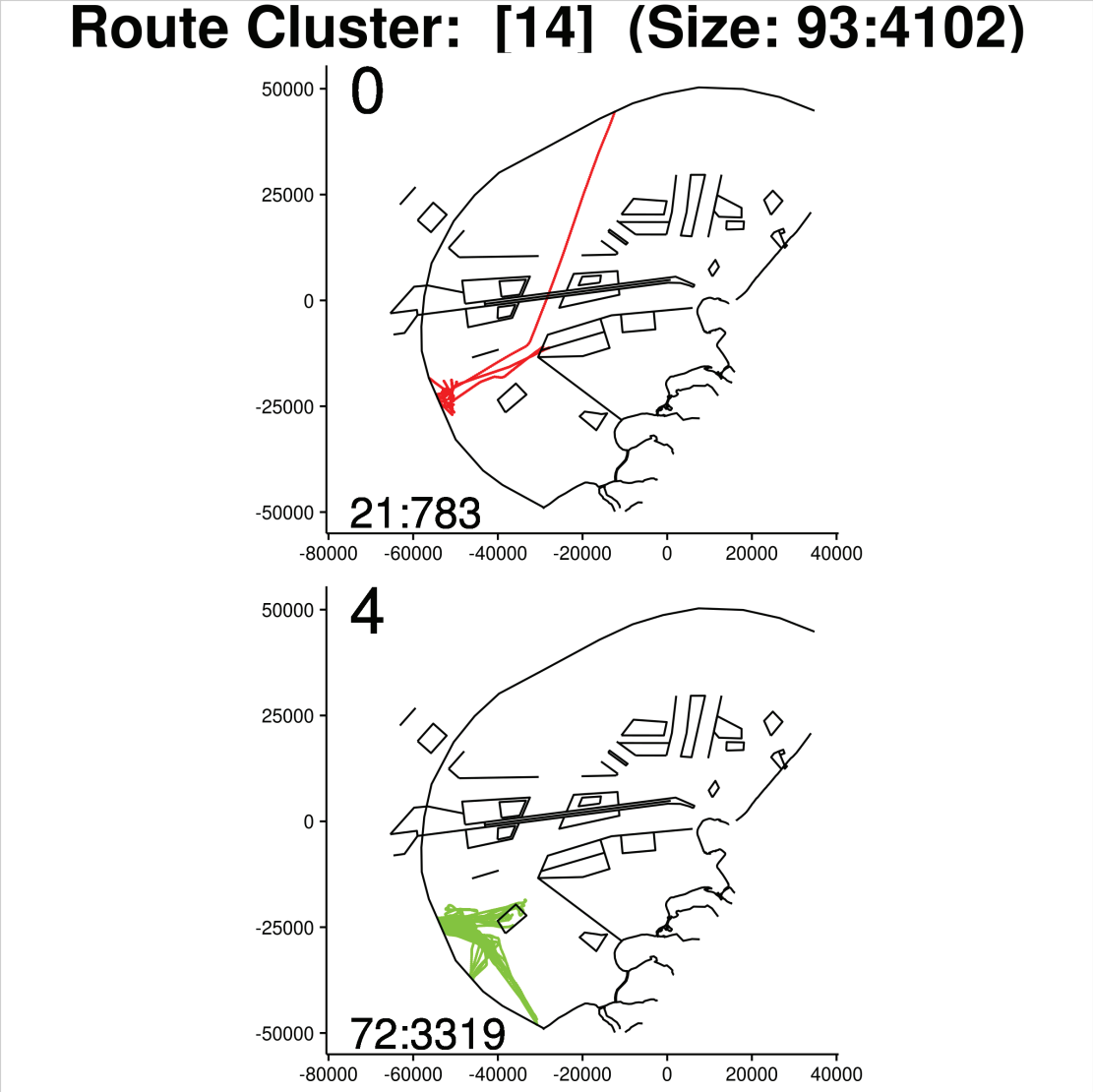



**Figure 8.16:** Spatial plot of all routes originating from entry cluster 14 (Southwest). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

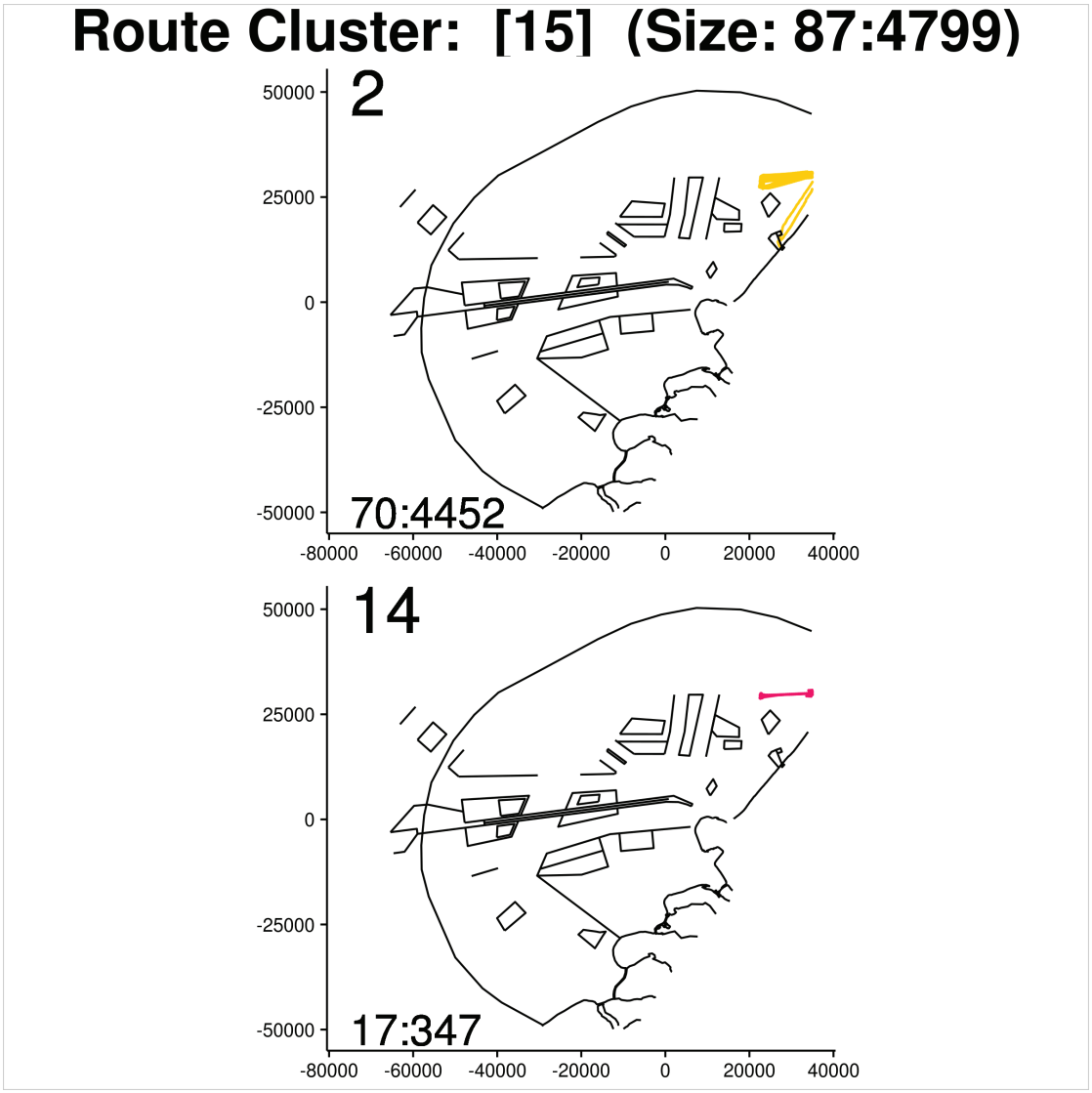



**Figure 8.17:** Spatial plot of all routes originating from entry cluster 15 (Katwijk). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

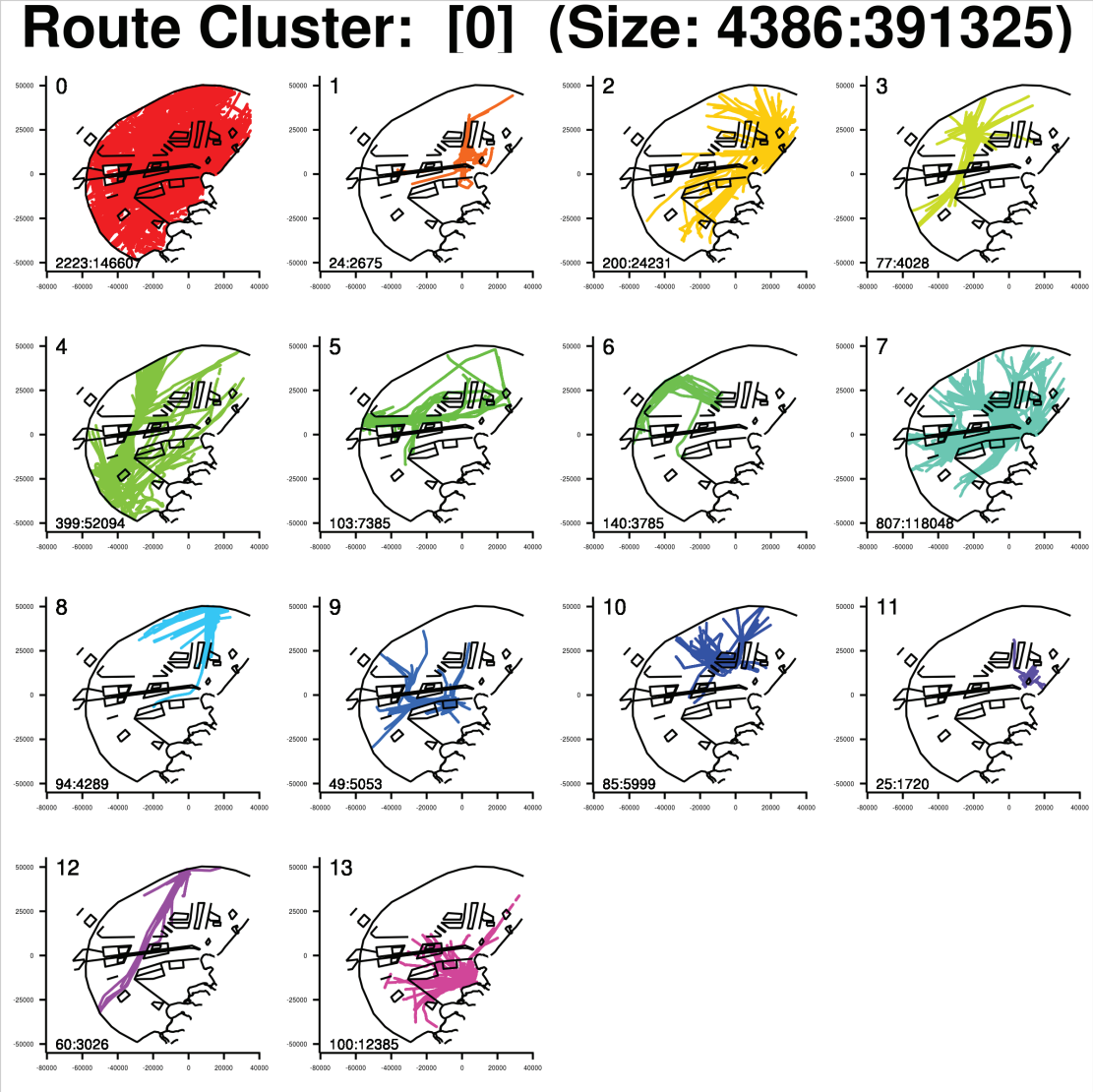



**Figure 8.18:** Spatial plot of all routes originating from entry cluster 0 (Noise). For each route (subplot), the exit point number is shown in the top left corner. In the bottom left corner of each subplot, the 'size' of the route is given (#journeys : # data samples). The total size of all the routes is shown in the title above the subplots.

reference, the routes are displayed in figs. 8.3 to 8.18. In table 8.4 an overview is displayed of how well each variable distinguished between the routes of each group. For each group, the number of journeys, the number of samples, and number of routes is shown, to give an indication of how significant the distinguishing is. For each variable, the four possible values are 'None', 'Weak', 'Moderate' and 'Strong'. The criteria for these values is different for numeric variables (states and dimensions) than for nominal variables (category and other). These values are explained using example figures in sections 8.2.1 and 8.2.2.

### 8.2.1 Numeric variables

A numeric variable is classified as strongly distinguishing when most routes have very little overlap with each other route. Unfortunately no single variable discerns strongly, so no clear example can be given.
A numeric variable is classified as moderately distinguishing when some routes overlap partially, or certain values can cause grouping. Causing grouping means that for one value, it becomes likely that the vessel belongs to one of a few routes. In fig. 8.19 (entry point 5 - Rijnveld 1), multiple values cause grouping. At a course around 0rad, only exit points 3, 4 and noise are possible, while for a course of 1.7rad, exit points 4 and 7 are dominant, with a small possibility of 10 and 3. Exit point 6 is limited to a very small range of courses, barely overlapping with 10 and 7. When considering the same variable, but then for journey averages (fig. 8.20), even more distinct bands become apparent. These tend to being classified as strong distinction, but the overlap of 10 and 7, together with the overlap of 4 and Noise fail the requirement.
Numeric variable is defined as weakly discerning when all routes overlap for a range of values, but that for a certain (significant) range of values, routes with smaller ranges can be excluded. A good example of a weak numeric variable classification is shown in fig. 8.21, where the journey distribution of accelerations of all routes originating in entry point 4 (Westkapelle) are depicted. Here it can be seen that all the routes overlap significantly, which makes distinction difficult for a range of acceleration values. However, for values below $10^{-3}$, it is very likely that the route the vessel will take is towards exit point 4 (Brugge/Westkapelle).
If even the requirement for weakly discerning is not met, it is classified as not discerning ('None'). Figure 8.22 shows the standard distribution of the same group and variable as in fig. 8.21. Here it can be seen that all routes overlap for the entire range of accelerations covered by routes 9 and 12. Excluding these smaller range groups is barely possible, and when it is (e.g. below $10^{-5.5}$), the presence of the noise group is still high.
An overview of the classification scheme is given in table 8.2.

### 8.2.2 Nominal variables

A nominal variable is classified as strongly distinguishing when for each possible given value it is quite certain (80% or more) which route is the result. In fig. 8.23 the desti-

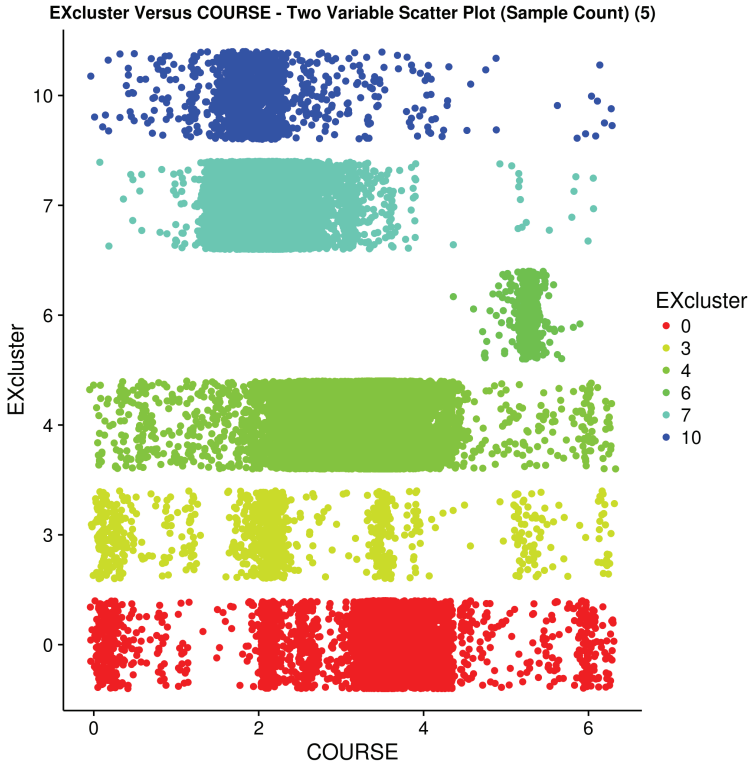



**Figure 8.19:** Route vs. course - Data sample distribution. Course distinguishes moderately between the routes originating from entry point 5 (Rijnveld 1).

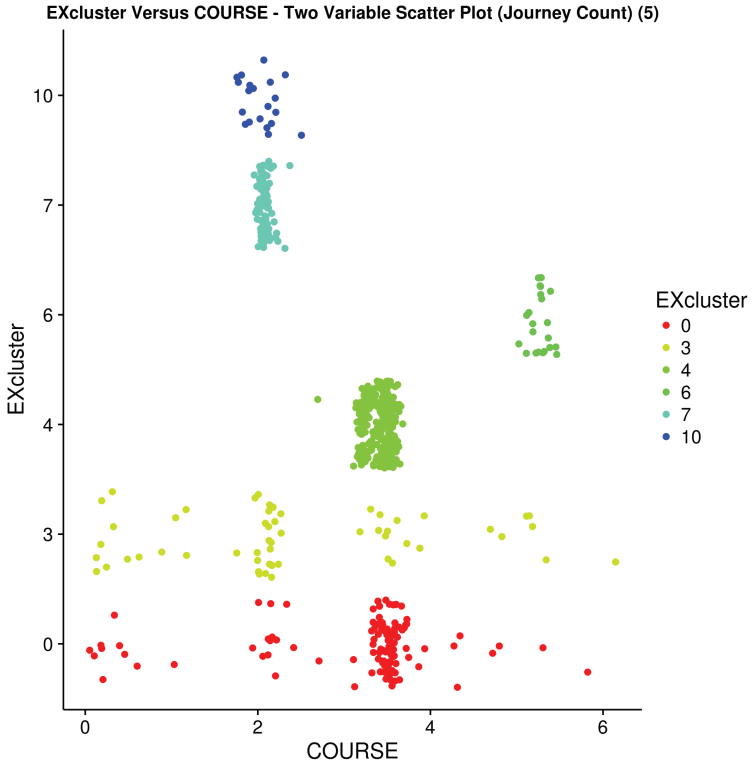



**Figure 8.20:** Route vs. course - Journey distribution. Course distinguishes moderately between the routes originating from entry point 5 (Rijnveld 1).

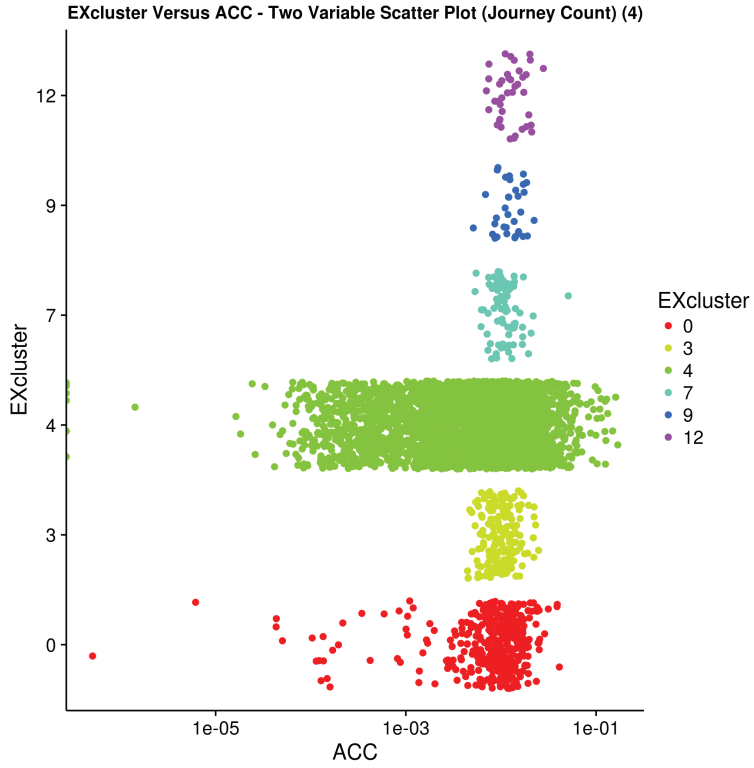



**Figure 8.21:** Route vs. acceleration - Journey distribution. Acceleration distinguishes weakly between the routes originating from entry point 4 (Westkapelle).

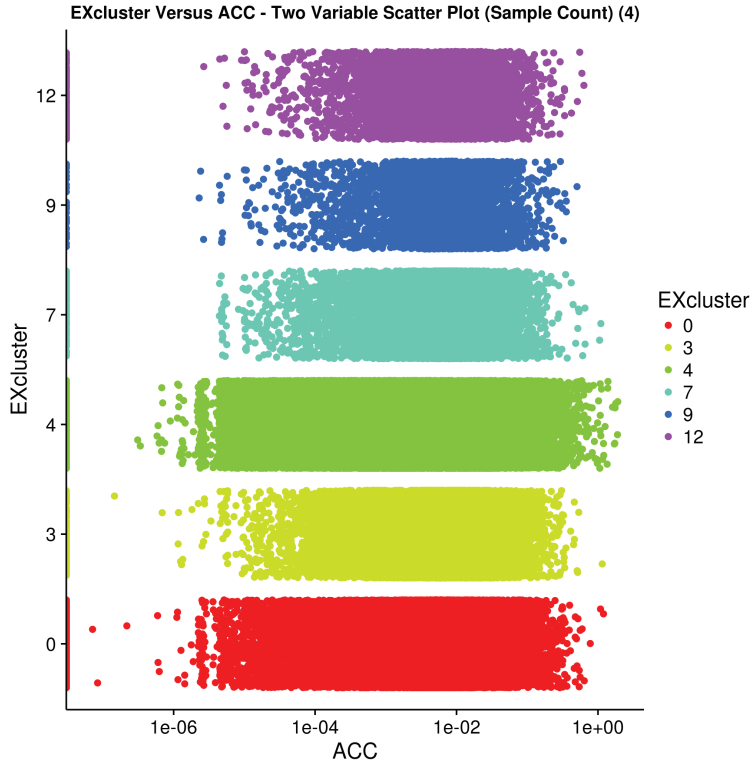



**Figure 8.22:** Route vs. acceleration - Data sample distribution. Acceleration cannot distinguish between the routes originating from entry point 4 (Westkapelle)

| Distinction | Description |
|---|---|
| Strong | Most routes have little overlap with each other route. |
| Moderate | Some routes overlap partially, or certain values cause grouping. |
| Weak | All routes overlap, but small-range groups can be excluded for certain values. |
| None | All routes overlap, no value can exclude any group. |

**Table 8.2:** Summary of distinction classes of numeric variables.

| Distinction | Description |
|---|---|
| Strong | For each (significant) value, 80% certainty of one route. |
| Moderate | One or more (significant) values have an 80% certainty of one route. |
| Weak | One (significant) value has an 80% collective certainty of 2-3 routes (non-noise). |
| None | For each value 4 or more routes are possible or noise is more than 20%. |

**Table 8.3:** Summary of distinction classes of nominal variables.

nations for each route are displayed (entry point 10 - Rijnveld 2). Here it is clear that given value 'RDAM', it is certain that the vessel will head for exit point 7 (Rotterdam). Also, for 'ENG-EAST','NL-SOUTH' and for 'GERMANY', the exit point is quite clear. The remaining destinations, however, also include much noise or several destinations. Due to the significant portion of clear samples, the variable destination is classified as strong in this plot.
A nominal variable is classified as moderately distinguishing when for one or more values it is clear (80%) what route the vessel belongs to. Considering the journey values for a different entry point (fig. 8.24), where for four destinations it is certain that the vessel is on the way to exit point 4 (Brugge/Westkapelle). However, the other two destinations have much noise present, which classifies the variable destination as moderately distinguishing for this group of routes.
A nominal variable is classified as weakly distinguishing when one value can at least result in 2-3 routes collectively accounting for 80% certainty (non-noise). Figure 8.25 (entry point 13 - Katwijk Anker) shows for the status of 'RM' - restricted manoeuvrability, two routes are possible, with a small chance of noise. Since two routes are possible, it does not distinguish moderately here, but there is some information available about the distribution of a few groups.
If for each value 4 or more routes are possible, or there is too much noise ($> 20\%$) then the variable is classified as non-distinctive for the entry point. This for example is the case in fig. 8.26 (entry point 11 - Stellendam), where 'ENG - underway using engine' gives a larger likelihood that a vessel is heading towards exit point 13 (Stellendam), but the presence of noise is too large. This also even goes for the other navigation status values.
An overview of the classification scheme is given in table 8.3.

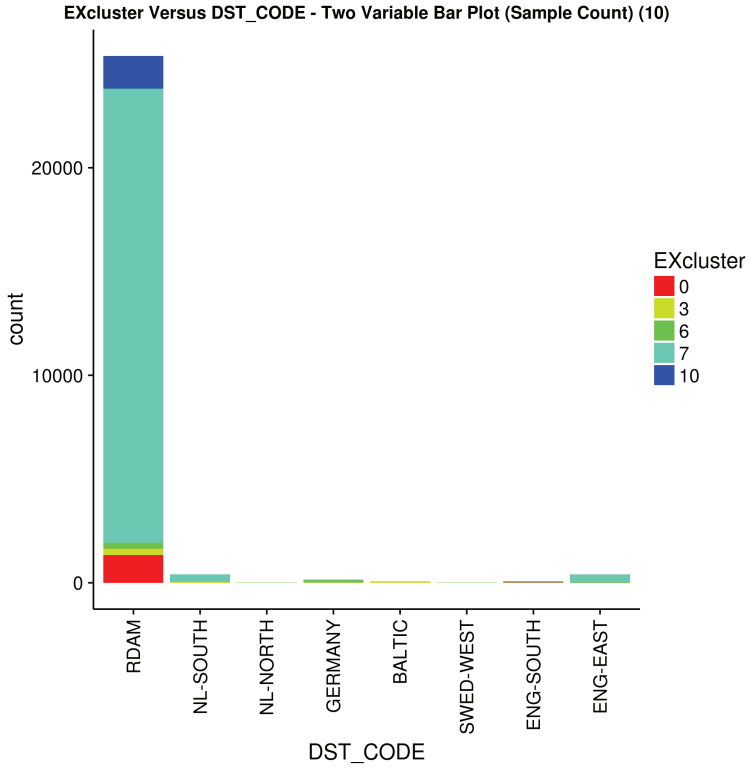



**Figure 8.23:** Route vs. destination - Data sample distribution. Destination distinguishes strongly between the routes originating from entry point 10 (Rijnveld 2).

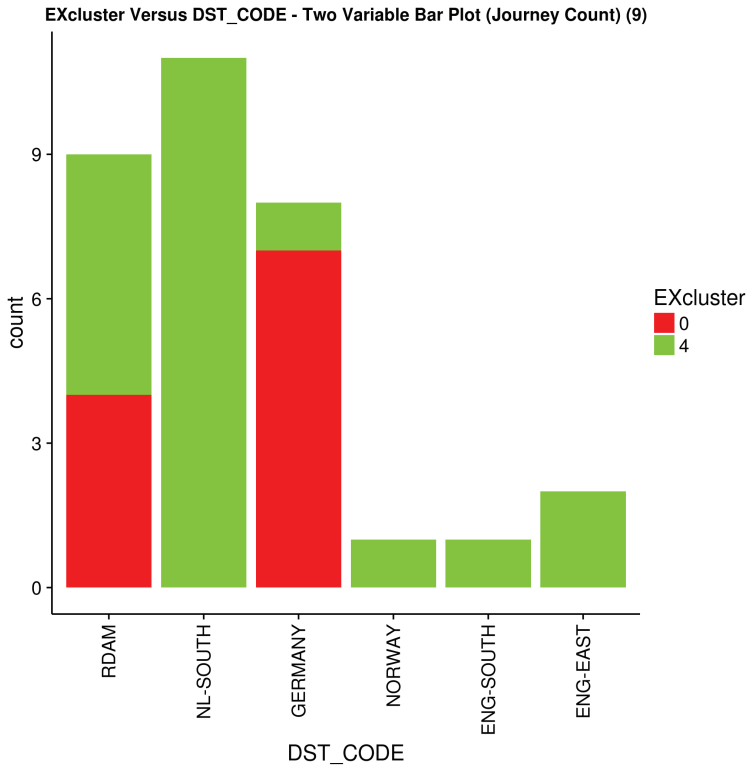



**Figure 8.24:** Route vs. destination - Journey distribution. Destination distinguishes moderately between the routes originating from entry point 9 (Bruggen).

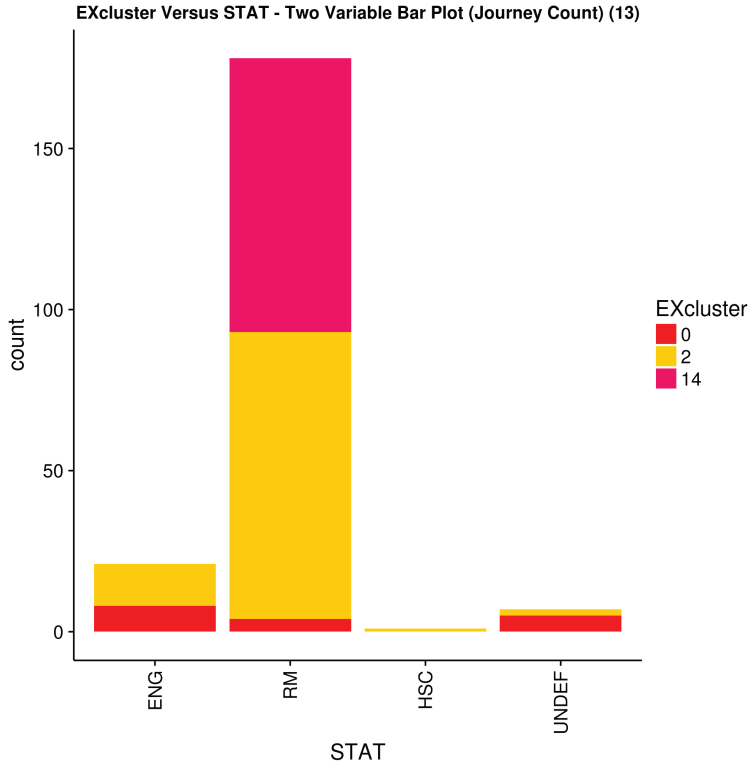



**Figure 8.25:** Route vs. navigation status - Journey distribution. Navigation status distinguishes weakly between the routes originating from entry point 13 (Katwijk Anker).

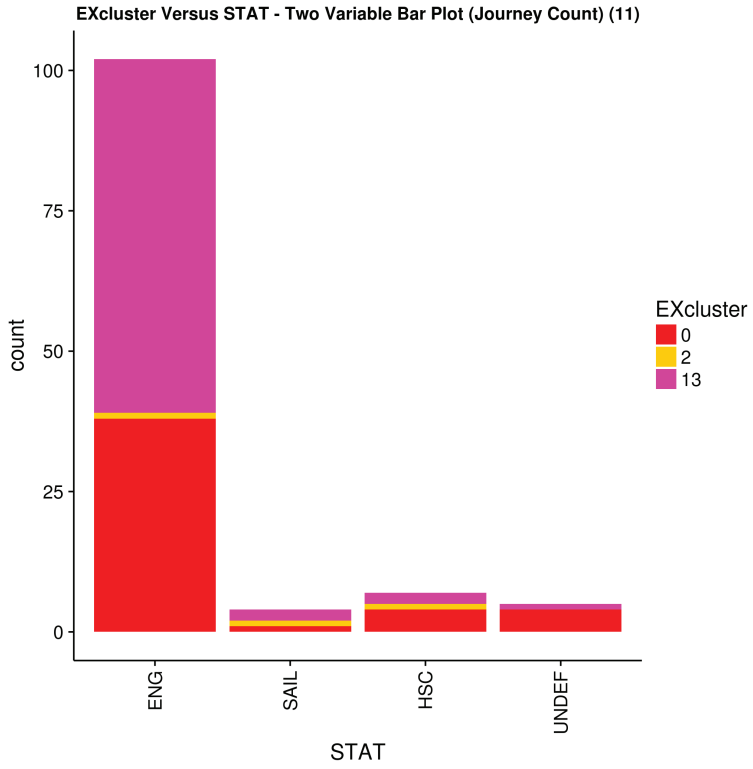



**Figure 8.26:** Route vs. navigation status - Journey distribution. Navigation status cannot distinguish between the routes originating from entry point 11 (Stellendam).

| | | | | | State | | | | Dimensions | | | Category | Other | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Accele-ration | Course | Speed | Turn Rate | Length | Width | Draught | Type | Destination (coded) | Navigation Status | Time of day |
| Entry Point | Name | Routes | Journeys | Samples | ACC | COURSE | SPEED | TURN | L | W | D | TYPE | DST_CODE | STAT | HOUR |
| 0 | Noise | 14 | 4386 | 391325 | None | Moderate | Weak | None | Weak | Weak | Weak | Weak | Moderate | None | None |
| 1 | Rotterdam | 13 | 5594 | 650443 | Weak | Moderate | Weak | Weak | Weak | Weak | Weak | None | Moderate | Weak | None |
| 2 | North Hinder S | 4 | 1305 | 199641 | None | Weak | Weak | None | Weak | Weak | Weak | Weak | Moderate | Weak | Weak |
| 3 | Scheveningen | 4 | 2211 | 115124 | Weak | None | None | None | None | None | None | - | Weak | Moderate | Weak |
| 4 | Westkapelle | 6 | 3823 | 214781 | None | Moderate | Weak | Weak | Weak | Weak | Weak | Weak | Moderate | Moderate | None |
| 5 | Rijnveld 1 | 6 | 558 | 69180 | None | Moderate | Weak | None | Weak | Weak | Weak | Weak | Moderate | Weak | Weak |
| 6 | North Hinder N | 2 | 543 | 18643 | None | None | None | None | Weak | Weak | Weak | - | None | Moderate | Weak |
| 7 | Anker 4A | 3 | 291 | 26264 | None | None | Weak | None | Weak | Weak | Weak | Moderate | Moderate | Moderate | Weak |
| 8 | Ijmuiden | 3 | 526 | 62741 | None | Moderate | Weak | None | None | None | None | Weak | Weak | None | Weak |
| 9 | Bruggen | 2 | 166 | 11694 | None | None | None | None | Weak | None | Weak | - | Moderate | None | None |
| 10 | Rijnveld 2 | 5 | 326 | 31268 | None | Weak | Weak | None | None | None | None | Moderate | Strong | Weak | Weak |
| 11 | Stellendam | 3 | 226 | 30159 | None | Weak | None | None | None | None | Weak | - | None | None | None |
| 12 | Anker 5A | 1 | 164 | 17724 | - | - | - | - | - | - | - | - | - | - | - |
| 13 | Katwijk Anker | 3 | 216 | 10868 | None | Weak | None | None | None | None | Weak | - | None | Weak | None |
| 14 | Southwest | 2 | 93 | 4102 | Weak | Weak | None | None | Weak | Weak | Weak | - | Moderate | Moderate | Weak |
| 15 | Katwijk | 2 | 87 | 4799 | Weak | Weak | Weak | None | Weak | Weak | Weak | - | - | Moderate | Weak |

**Table 8.4:** An overview of how well each variable distinguishes between routes. The routes are grouped by Entry Point. The variable names and codes are at the head of each column. Possible values are 'None', 'Weak', 'Moderate' and 'Strong'.

### 8.2.3 Discussion

This section discusses the results in table 8.4, while also including relevant remarks from two-variable plots. As stated in section 7.1.3, each combination of two variables was plotted for each group of routes, coloured by their exit point. Up front it may be noticed that entry point 12 (Anker 5A) has not been evaluated, which is because the group only contains one route, therefore knowing the entry point is already sufficient to know the route.

#### 8.2.3.1 Course

Of the state variables, course is apparently the most promising variable. For five entry points, it can distinguish between routes in a moderate degree. This is especially clear when looking at the journey plots. Specifically for exit point 8 (Ijmuiden), combining course with other variables such as acceleration and turn rate, results in a visibly strong distinction. For other entry points, combining course with other variables contributes to distinguishing between routes (more than just by using course), because certain combinations of the variables result in a certain route being the sole possibility. The only group for which this does not count is entry point 3 (Scheveningen). This could be explained by the fact that the boats coming from Scheveningen primarily return to Scheveningen (see table 8.1), which means that each journey covers all courses. What is also noticeable about course, is that journey course averages are often quite closely grouped within one route.
Course can be concluded as a promising variable for intent prediction, especially because it also can distinguish between the largest groups (entry points 0,1,4). However it must be taken into account that the differences between courses becomes larger as a journey progresses, and is not necessarily already distinct around the entry point. For example, all vessels originating in Rotterdam start out with a roughly westward course.
It is important to note that these conclusions are based on the entire journey of a vessel. The course of a vessel very much depends on position (see section 6.3.6). A strong example of this is that most journeys with entry point 'Rotterdam' start out with the same westward course, and only part ways after having travelled approximately 10km. This means that the usefulness of course for predicting intent also depends on space. This will need further investigation. This finding is what has formed the incentive to perform waypoint detection by course.

#### 8.2.3.2 Destination

Of the nominal variables, but possibly of all the variables, destination seems to be a promising variable. As can be seen in fig. 8.23, when a value for destination is given, the likelihood of a certain route is high. It distinguishes moderately for most groups, including the larger groups. When combining destination with other variables, the distinction between routes becomes more clear. This especially is true for combinations with state

and dimension variables. Each combination of variables contributes to separating routes in different entry point groups. In fig. 8.27, the distribution of destination is plotted against the course for entry point 'Noise'. Here it can be seen that several combinations result in a clear identification of the 'route' a vessel will take. For (a simple) example, a vessel with destination 'DEN-EAST' with a course of approximately 4.5 rad will very likely be headed towards exit point 5 (Rijnveld 1). Combining the two variables has resulted in isolating the noise traffic with destination 'DEN-EAST'. This accounts for less than 20 journeys (going to DEN-EAST), but at least it becomes clear that combining variables can even help to discern intent from the entry point group 'Noise'. For other entry points, combinations with numeric variables result in very many more of these isolations.
8427 journeys have an entry point for which it is 80% certain which route a vessel will take based on the vessel's destination. This is 41% of all journeys. This number must be set relative to the sum of journeys that follow specific routes. Based on entry point only, for 35% of the journeys the route which they will take is 80% certain. It can therefore be said using the variable destination improves the capacity to predict the intended exit point.

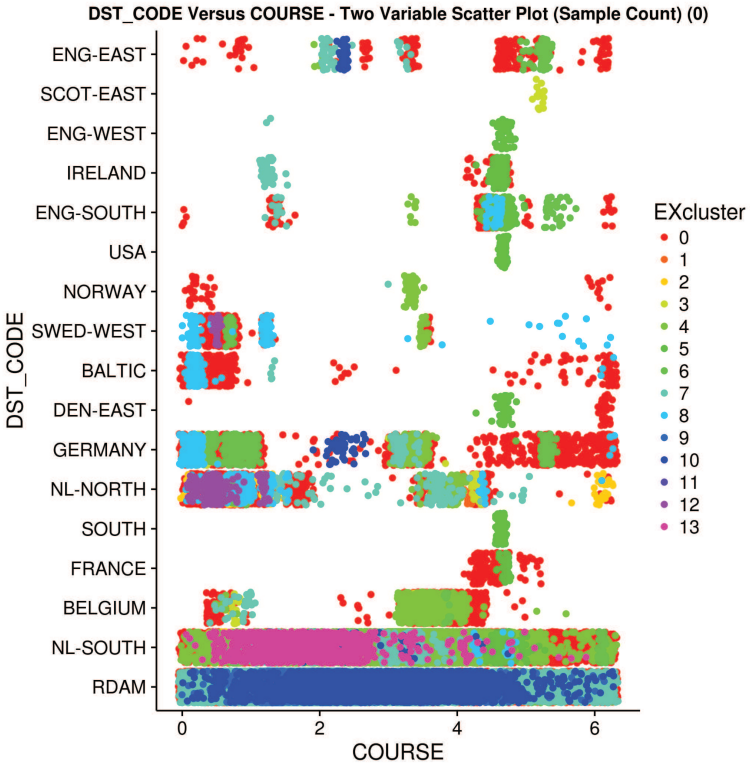



**Figure 8.27:** Destination vs. course coloured by exit point - Sample distribution of routes originating from entry point 0 (Noise)

## 8.3 Waypoint detection

The route from entry point 8 (Ijmuiden) to exit point 7 (Rotterdam) is shown in fig. 8.28. This route has been chosen since for this entry point, because this route has a wide spread of possible trajectories to reach the end point. When the change-point detection and clustering was performed on this route, the best clustering unfortunately did not result in a clear distinction of sub-routes. In fig. 8.29 the best waypoint clusters are displayed. Though two waypoint clusters might separate the routes in the beginning (north, clusters 3 and 4), all vessels traverse both of these clusters in any case, and from there on continue to the large waypoint cluster 1, from there on continuing to Rotterdam from there. The hope was to discover a split in cluster 1, but this cannot be done with this approach, since changing the parameters in favour of achieving a split results in significant growth in the noise cluster. A probable cause for this is that the vast amount of waypoints that is clustered in cluster 1 has a very dense, but reasonably homogeneous distribution. This in turn may be due to the large amount of interacting traffic at this crossing, causing local behaviour to dominate global (intent) behaviour. Due to time constraints, this has not yet been further investigated, and is thus left for future work.

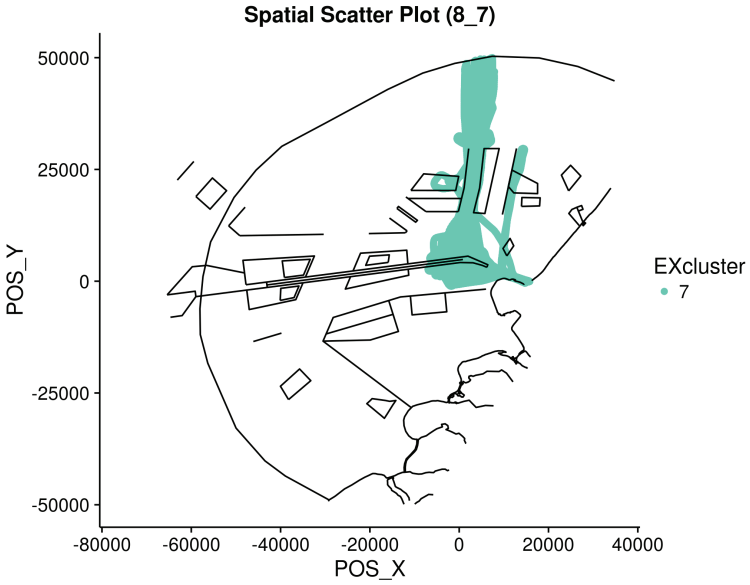



**Figure 8.28:** Spatial distribution of route 'Ijmuiden(8) to Rotterdam(7)'.

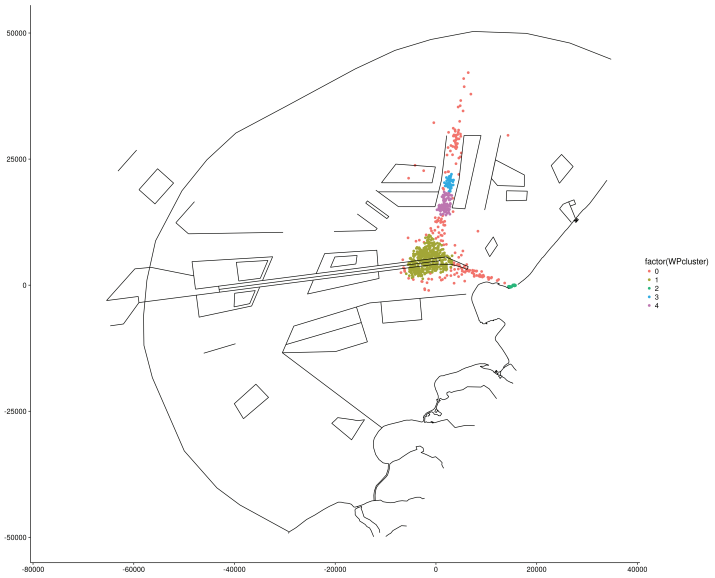



**Figure 8.29:** Waypoint clusters discovered for route 'Ijmuiden(8) to Rotterdam(7)'. Parameters used: minPts=15, eps=865.

# Chapter 9

# Validation

In this chapter, the small validation procedure that was performed is detailed and discussed. The variable 'destination' is validated for how well it can predict the route that a vessel will take, based on the entry point. Prediction is done solely based on the most likely route (from the original dataset distributions). A sample of 200 journeys is taken from the validation set to confirm these likelihoods.

## 9.1 Procedure

A small validation was performed to check the results of variable usability(section 8.2). Validation of the course variable is not straightforward, since it is very much dependent on position. Since this research has not yet covered the positional dependence of this variable, the conclusions are preliminary and cannot yet be validated. The destination of a vessel, is usually already known at or near the entry point, therefore this variable is validated. The point of the validation is to confirm the certainties of prediction based on the entry point and destination that have been derived.
The procedure for validation started by randomly sampling 200 journeys from the validation set (the second half of the data, see section 5.1). For each of these journeys, the entry and exit point were determined by clustering them along with the original dataset. Since the number of validation journeys is 100 times smaller than the number of journeys in the original dataset, the effect on the clustering method is negligible. The entry points were used as input for the validation, while the exit points were used for validate the output.
For each entry cluster, the largest group of routes was used to predict which exit point each validation journey would take. The same was done for the combination of entry cluster and destination: the most likely route was used to predict the journey's intended route (exit point). These predictions were then compared to the actual exit points, and the number of correctly predicted exit points was registered. The percentages of correctly predicted exit points were then compared to the expected percentages.

## 9.2 Results

In table 9.1 the results of the validation process can be seen. The left half shows the validation of the prediction by entry point, while the right half includes the destination in the prediction. Only the entry clusters and destinations contained in the validation samples are shown. The columns indicated with 'max. likelihood' indicate the predicted likelihood of the largest route in terms of journey count. Rows are marked in bold if te number of sampled journeys is 10 or more.
For the larger entry cluster groups (clusters 0-4), the validated likelihoods are close to that of the predicted likelihoods, considering the size of the validation sample set. Only the noise cluster (number 0) has a large difference with the predicted value. This could be attributed to the fact that noise in the validation set can be very different from noise in the original set. For the smaller groups, the number of validation journeys is too small ($< 10$) to draw even preliminary conclusions.
When using destination to predict the intended route, the predicted likelihoods differ more from the validated likelihoods (up to 13% for entry point 3, destination <empty>). It can be expected that with a larger validation set, more conclusions can be drawn.
In section 8.2.3.2 it was stated that the variable destination improves the capacity to predict the intended exit point, relative to using the entry point only. This was based

on the 80% confidence threshold. The validation results show no large enough groups with a 80% likelihood, which leaves the improvement inconclusive.
The largest conclusion to be drawn from this validation is that the validation set is not large enough to draw significant confirmation or rejection of the results. It is left to future work to extend this validation to a larger set and possibly more variables. Also, it would be best to say that all the results of this research are but preliminary.

| | Predicted | | Validated | | | | Predicted | | Validated | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Entry Cluster | Journeys | Max. Likeli-hood | Journeys | Correctly predicted | Likelihood | Destination | Journeys | Max. Likeli-hood | Journeys | Correctly predicted | Likelihood |
| **0** | **4404** | **51%** | **49** | **31** | **63%** | **RDAM** | **1099** | **67%** | **10** | **6** | **60%** |
| | | | | | | NL-SOUTH | 449 | 48% | 6 | 2 | 33% |
| | | | | | | NL-NORTH | 47 | 60% | 1 | 1 | 100% |
| | | | | | | **<empty>** | **2613** | **63%** | **32** | **24** | **75%** |
| **1** | **5594** | **25%** | **66** | **14** | **21%** | RDAM | 613 | 81% | 6 | 5 | 83% |
| | | | | | | NL-SOUTH | 183 | 63% | 3 | 2 | 67% |
| | | | | | | FRANCE | 59 | 85% | 1 | 1 | 100% |
| | | | | | | SOUTH | 44 | 89% | 1 | 1 | 100% |
| | | | | | | GERMANY | 216 | 67% | 2 | 1 | 50% |
| | | | | | | ENG-SOUTH | 341 | 94% | 5 | 5 | 100% |
| | | | | | | ENG-EAST | 294 | 71% | 1 | 1 | 100% |
| | | | | | | **<empty>** | **3486** | **25%** | **46** | **9** | **20%** |
| **2** | **1304** | **58%** | **13** | **8** | **62%** | **RDAM** | **1140** | **66%** | **12** | **8** | **67%** |
| | | | | | | NL-SOUTH | 46 | 52% | 1 | 1 | 100% |
| **3** | **2208** | **88%** | **21** | **16** | **76%** | **<empty>** | **2174** | **89%** | **21** | **16** | **76%** |
| **4** | **3823** | **81%** | **26** | **20** | **77%** | NL-SOUTH | 1083 | 96% | 9 | 9 | 100% |
| | | | | | | GERMANY | 62 | 44% | 1 | 1 | 100% |
| | | | | | | ENG-EAST | 60 | 82% | 1 | 1 | 100% |
| | | | | | | **<empty>** | **2332** | **80%** | **13** | **9** | **69%** |
| 5 | 562 | 48% | 4 | 3 | 75% | NL-SOUTH | 171 | 82% | 2 | 1 | 50% |
| | | | | | | <empty> | 240 | 52% | 2 | 2 | 100% |
| 6 | 541 | 83% | 4 | 4 | 100% | GERMANY | 68 | 91% | 1 | 1 | 100% |
| | | | | | | <empty> | 441 | 83% | 3 | 3 | 100% |
| 8 | 526 | 51% | 5 | 3 | 60% | RDAM | 399 | 65% | 4 | 3 | 75% |
| | | | | | | <empty> | 106 | 88% | 1 | 1 | 100% |
| 9 | 166 | 67% | 2 | 2 | 100% | <empty> | 134 | 67% | 2 | 2 | 100% |
| 10 | 310 | 63% | 4 | 2 | 50% | RDAM | 238 | 81% | 2 | 2 | 100% |
| | | | | | | <empty> | 57 | 65% | 1 | 0 | 0% |
| 11 | 226 | 49% | 4 | 3 | 75% | <empty> | 211 | 50% | 4 | 3 | 75% |
| 12 | 164 | 100% | 1 | 1 | 100% | RDAM | 158 | 100% | 1 | 1 | 100% |
| 14 | 93 | 76% | 1 | 0 | 0% | <empty> | 42 | 79% | 1 | 0 | 0% |

**Table 9.1:** Validation Results.

# Chapter 10

# Conclusion

*Which variables can be used to predict the intent of a vessel?* - is the main research question of this thesis was. This chapter ventures to answer this question through its sub-questions, and will end with answering this main question.

*What is a useful way to group the data based on intent?*
In this thesis two manners of grouping have been used: by route and by planned path. A route is an 'origin-destination' combination of a vessel's journey (entire set of datapoints a vessel traverses within the scope of the area considered). A planned path is a series of connected waypoints between and including the origin and the destination of a journey. The route captures describes a vessel intent in a very minimalistic way. The only assumption is that a vessel has a certain aim, and comes from a certain point. Therefore analysis can be done in a very basic manner if journeys with similar routes are grouped. One step of complexity deeper is the concept of planned path. This thesis has only touched upon this manner of grouping, since a suitable method for finding generic waypoints has not yet been found/applied, though a start has been made. However it does seem promising, since it can capture the intermediate aims of a vessel with a simple description (small series of waypoints).

*Which variables would allow distinguishing between intents?*
The course of a vessel, as well as its destination, are at the least promising variables to distinguish which route a vessel intends to take. The destination variable is conceptually closely tied to the point where a journey exits the scene. For 41% of all journeys, the intended route can be predicted with an 80% certainty using the destination variable, with a given point of entry. This is an improvement upon the 35% of journeys for which the intended route can be predicted based on entry point only with the same level of certainty. Therefore the preliminary conclusion can be formed that destination is useful

in distinguishing between possible vessel intents.
Different routes that have a common entry point can be reasonably distinguished from each other when regarding the course variable. However, this holds for the entire set of course values throughout the entire journey. Course also depends on spatial position, which means that depending on the location of a vessel, it will be easier or less easy to distinguish where the vessel intends to go, based on its course. A strong example of this is that most journeys with entry point 'Rotterdam' start out with the same westward course, and only part ways after having travelled approximately 10km. Further investigation is necessary to exploit the combination of course and position (and possibly other factors).
The other variables that have been investigated in this research (speed, acceleration, turn rate, length, width, draught, type, navigation status, and time of day) do not distinguish well between routes at first sight. However, it can be said that combining these variables with course and destination provides more distinction between routes. Further investigation should reveal the significance of this addition. Also, the variables speed, acceleration and turn rate are dependent on time and space in very much the same way as course.

*To what extent can historical data provide sufficient support for a medium-term prediction method?*
Based on the findings in this thesis only, this question cannot be answered in full. This is mainly due to the fact that actual prediction has not yet been performed, save for a small scale validation which was yet too small to be conclusive. However, in terms of medium-term to long-term behaviour, it can be concluded that most vessels follow similar behavioural patterns (both route as well as planned path) in spatial terms., assuming that the derived routes are a good approximation of vessel intent.

***Which variables can be used to predict the intent of a vessel?***
In summary, it can be concluded that the variable destination is useful in predicting the intent of a vessel in terms of route. Also, the course of a variable is not directly usable to predict a route due to its dependence on vessel position and possibly other variables, but is at least promising in predicting vessel intent in terms of planned path. Other variables are possibly useful in combination with destination and/or course to predict the goal and planned path of a vessel.

# Chapter 11

# Recommendations

In this chapter the reader is pointed towards the future. This is done in two parts. The first part (section 11.1) deals with necessary or potential extensions/improvements of the analysis done in this thesis. The second part (section 11.2) discusses potential future applications, that can build on the analyses done. This purpose is to set a step-up for the sake of later research, but also to provide a context for the future.

## 11.1 Furthering the analysis

Several improvements can be made on the analysis, and several extensions can potentially provide additional information about the longer term behaviour of vessels. This section discusses what can be done further to create a better understanding of vessel intent, for the purpose of prediction.

### 11.1.1 Destination

The most immediate next step is to extend the validation volume (as described in chapter 9) Since the destination variable has proven to be useful in distinguishing where vessels are headed, it is a logical step to bring this a step further. The current coding scheme was only a rough dividing of the wide variety of destinations that have been coded. Also, the destinations have been coded in such a way that only the destination entries from AIS and the harbour database that can for certain be identified unambiguously are coded, leaving all others un-coded. A more scrutinous coding scheme promises to enhance the distinguishing ability of the destination variable. Also, for the (more) ambiguous destination entries, a system can be set up based on likelihood for example, or based on an extensive case-based feed.

### 11.1.2 Exit point refinement

A minor improvement can be made upon the manner of clustering the exit points. A few exit point clusters span a large area, making the exit point quite indistinct. This can be solved in at least two manners. The first is to perform single-recursive clustering on all the exit points that belong to the cluster. The second approach is to cluster exit points only within the routes grouped by entry point cluster. The latter approach does cause many exit point clusters to occur across the entire dataset, and may therefore cause much work in identifying the exit points (of which many will overlap).

### 11.1.3 Waypoints

The potential of distinguishing waypoints is still there, not yet explored much. The fact that the course variable has proven to be distinct and stable for many vessels is additional grounds for furthering this investigation. As mentioned in section 8.3, distinguishing clear waypoints was not successful with the investigated route. The first step would be to investigate other routes (traversing less busy areas) with the same method. The hypothesis that the lack of discerning power is due to the largely local behaviour can be tested in this way. Then, it may be more promising to investigate other clustering methods. Another alternative is to replace the change-point detection technique by a technique similar to that of Ikeda et al. (2013) (see section 2.2.2.2).
If waypoints can be found and clustered well, the variable 'course' becomes useful, since

the spatial dependence is described (in part) by the series of waypoints that a journey follows. Predictions can then be made regarding which value for course a vessel will take directly after the next waypoint. This can then be properly validated.

### 11.1.4 Uncertainty

Another important aspect that would be a valuable addition to the investigation would be an accompanying measure of confidence and/or reliability for all the variables. As discussed in section 3.2, the reliability of state data depends on the distance from the radar, as well as on the quality of on-board sensors. It may be possible to determine a measure of how reliable this information is. This could possibly also take care of the large amounts of clutter on the edges of the coverage. A similar measure could be used for all the static data.

### 11.1.5 Human and environmental factors

Further down the line in the investigation of vessel intent, is the analysis of the human and environmental factors. Extensive interviews and voice recordings can be used to analyse the influence of communication, for example, or weather and tidal measurements.

## 11.2 Potential Uses

The ventures of this thesis have been focused on the spatio-temporal intent of a vessel, but chapter 2 also has shown that human and environmental factors are of significant influence on the behaviour of a vessel. In this case study, input from experts has shown that the level of communication (skills) -between vessels and between vessels and VTSO 's- is a major factor in the behaviour of vessels, as well as the risk of collisions. Therefore any model that predicts medium-term collision risk will have to include these factors in some way.
The most intuitive follow-up of the basis formed by this thesis (besides the extensions discussed in section 11.1), is to derive (empirical) variable distributions per route, coupled to position and possibly orientation. Then for a new journey (not used for the distributions), a likelihood estimation can be made to discern which route the journey belongs to. This can be any form of estimation, such as a particle filter.
This can be detailed even further, by determining these distributions per 'leg' (connection between two waypoints) of a path, and using them to estimate which waypoint is most likely next, or even to predict what will happen directly after the next waypoint. If this type of prediction can be done well, then density/risk estimates can be made on a medium-term or even long-term time horizon, for each relevant crossing in a TSS scheme.
In the long run, it seems best to build a coherent system of models, which can account for the many factors that influence the behaviour of a vessel. Isolating models, each

treating one aspect of the problem, cannot represent the complete socio-technical system. The relations between the components (humans, vessels, environment, long-term, short-term) are just as important as their separate behaviours. Therefore a system of models with carefully designed relations/interactions is more promising.

# References


Althoff, M., O. Stursberg, and M. Buss
2009. Model-based probabilistic collision detection in autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 10(2):299–310.

Best, G. and R. Fitch
2015. Bayesian intention inference for trajectory prediction with an unknown goal destination. In IEEE International Conference on Intelligent Robots and Systems, volume 2015-December, Pp. 5817–5823.

Blom, H. A. P. and E. Bloem
2006. Joint particle filtering of multiple maneuvering targets from possibly unassociated measurements. Journal of Advances in Information Fusion, 1(1):15–34.

Blom, H. A. P. and A. Sharpanskykh
2015. Modelling situation awareness relations in a multiagent system. Applied Intelligence, 43(2):412–423.

Bomberger, N. A., B. J. Rhodes, M. Seibert, and A. M. Waxman
2006. Associative learning of vessel motion patterns for maritime situation awareness. In 2006 9th International Conference on Information Fusion, FUSION, Pp. 1–8.

Bouarfa, S., H. A. P. Blom, R. Curran, and M. H. C. Everdij
2013. Agent-based modeling and simulation of emergent behavior in air transportation. Complex Adaptive Systems Modeling, 1(1):15.

Bukhari, A. C., I. Tusseyeva, B. G. Lee, and Y. G. Kim
2013. An intelligent real-time multi-vessel collision risk assessment system from vts view point based on fuzzy inference system. Expert Systems with Applications, 40(4):1220–1230.

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Rinartz, C. Shearer, and R. Wirth
2000. Crisp-dm 1.0 step-by-step data mining guide. Report, SPSS.

Chauvin, C. and S. Lardjane
2008. Decision making and strategies in an interaction situation: Collision avoidance at sea. Transportation Research Part F: Traffic Psychology and Behaviour, 11(4):259–269.

Delahaye, D. and S. Puechmorel
2000. Air traffic complexity: Towards intrinsic metrics. In 3rd USA/Europe Air Traffic Management Research and Development Seminar, Napoli.

Elfring, J., R. Van De Molengraft, and M. Steinbuch
2014. Learning intentions for improved human motion prediction. Robotics and Autonomous Systems, 62(4):591–602.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu
1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD-96 Proceedings, volume 96, Pp. 226–231.

Fujii, Y., H. Yamanouchi, and N. Mizuki
1974. Some factors affecting the frequency of accidents in marine traffic. Journal of Navigation, 27(2):239–247.

Goerlandt, F., J. Montewka, V. Kuzmin, and P. Kujala
2015. A risk-informed ship collision alert system: Framework and application. Safety Science, 77:182–204.

Hahsler, M.
2016. Density based clustering of applications with noise DBSCAN and related algorithms.

Helbing, D. and P. Molnár
1995. Social force model for pedestrian dynamics. Physical Review E, 51(5):4282–4286.

Houenou, A., P. Bonnifait, V. Cherfaoui, and W. Yao
2013. Vehicle Trajectory Prediction based on Motion Model and Maneuver Recognition, Pp. 4363–4369. IEEE International Conference on Intelligent Robots and Systems.

Ikeda, T., Y. Chigodo, D. Rea, F. Zanlungo, M. Shiomi, and T. Kanda
2013. Modeling and prediction of pedestrian behavior based on the sub-goal concept. In Robotics: Science and Systems, volume 8, Pp. 137–144.

IMO
2001. International Maritime Organization - guidelines for the onboard operational use of shipborne automatic identification systems. Report.

ITU
2010. International Telecommunications Union (Radiocommunication Sector) - technical characteristics for an automatic identification system using time-division multiple access in the vhf maritime mobile band. Report Rec. ITU-R M.1371-4, ITU.

Killick, R. and I. Eckley
2014. changepoint: An r package for changepoint analysis. Journal of Statistical Software, 58(3):1–19.

Krumm, J., R. Gruen, and D. Delling
2013. From destination prediction to route prediction. Journal of Location Based Services, 7(2):98–120.

Krumm, J. and E. Horvitz
2006. Predestination: Inferring destinations from partial trajectories. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 4206 LNCS, Pp. 243–260.

Kuwata, Y., M. T. Wolf, D. Zarzhitsky, and T. L. Huntsberger
2014. Safe maritime autonomous navigation with colregs, using velocity obstacles. IEEE Journal of Oceanic Engineering, 39(1):110–119.

Lancia, C., D. Taurino, G. Frau, J. Verstraeten, and C. L. Tallec
2014. Traffic predictions supporting general aviation. In SIDs 2014 - Proceedings of the SESAR Innovation Days.

Lefèvre, S., C. Laugier, and J. Ibañez Guzmán
2012. Risk assessment at road intersections: Comparing intention and expectation. In IEEE Intelligent Vehicles Symposium, Proceedings, Pp. 165–171.

Lenart, A. S.
1999. Manoeuvring to required approach parameters - cpa distance and time. Annual of Navigation.

Li, B. and F. W. Pang
2013. An approach of vessel collision risk assessment based on the d-s evidence theory. Ocean Engineering, 74:16–21.

Lowe, C. D. and J. P. How
2015. Learning and predicting pilot behavior in uncontrolled airspace. In AIAA Infotech at Aerospace.

Lygeros, J. and M. Prandini
2002. Aircraft and weather models for probabilistic collision avoidance in air traffic control. In Proceedings of the IEEE Conference on Decision and Control, volume 3, Pp. 2427–2432.

Montewka, J., T. Hinz, P. Kujala, and J. Matusiak
2010. Probability modelling of vessel collisions. Reliability Engineering and System Safety, 95(5):573–589.

Pallotta, G., M. Vespe, and K. Bryan
2013. Vessel pattern knowledge discovery from ais data: A framework for anomaly detection and route prediction. Entropy, 15(6):2218–2245.

Prandini, M. and J. Hu
2016. A probabilistic approach to air traffic complexity evaluation. In Proceedings of the IEEE Conference on Decision and Control, Pp. 5207–5212.

Rahman, S. A., C. Borst, M. M., and M. van Paassen
2012. Measuring Sector Complexity: Solution Space-Based Method, book section 2, Pp. 11–34. InTech.

Rhodes, B. J.
2007. Taxonomic knowledge structure discovery from imagery-based data using the neural associative incremental learning (nail) algorithm. Information Fusion, 8(3):295–315.

Rhodes, B. J., N. A. Bomberger, and M. Zandipour
2007. Probabilistic associative learning of vessel motion patterns at multiple spatial scales for maritime situation awareness.

Ristic, B., B. La Scala, M. Morelande, and N. Gordon
2008. Statistical analysis of motion patterns in ais data: Anomaly detection and motion prediction.

Scott, A. J. and M. Knott
1974. A cluster analysis method for grouping means in the analysis of variance. Biometrics, 30(3):507–512.

Simmons, R., B. Browning, Y. Zhang, and V. Sadekar
2006. Learning to predict driver route and destination intent.

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

UNECE
2013. United Nations Economic Commission For Europe - united nations code for trade and transport locations (un/locode). url: http://www.unece.org/cefact/locode/service/location.

Vaněk, O., M. Jakob, O. Hrstka, and M. Pěchouček
2013. Agent-based model of maritime traffic in piracy-affected waters. Transportation Research Part C: Emerging Technologies, 36:157–176.

Vasquez, D., T. Fraichard, and C. Laugier
2009. Incremental learning of statistical motion patterns with growing hidden markov models. IEEE Transactions on Intelligent Transportation Systems, 10(3):403–416.

Wang, H., R. Wen, and Y. Zhao
2015. Topological characteristics of air traffic situation. In Proceedings of the 11th USA/Europe Air Traffic Management Research and Development Seminar, ATM 2015.

Wang, N., X. Meng, Q. Xu, and Z. Wang
2009. A unified analytical framework for ship domains. Journal of Navigation, 62(04):643.

Wen, Y., Y. Huang, C. Zhou, J. Yang, C. Xiao, and X. Wu
2015. Modelling of marine traffic flow complexity. Ocean Engineering, 104:500–510.

Xu, W., X. Liu, and X. Chu
2015. Simulation models of vessel traffic flow in inland multi-bridge waterway. In ICTIS 2015 - 3rd International Conference on Transportation Information and Safety, Proceedings, Pp. 505–511.

Zandipour, M., B. J. Rhodes, and N. A. Bomberger
2008. Probabilistic prediction of vessel motion at multiple spatial scales for maritime situation awareness.

Zhang, W., F. Goerlandt, J. Montewka, and P. Kujala
2015. A method for detecting possible near miss ship collisions from ais data. Ocean Engineering, 107:60–69.