UiO : Department of Mathematics
University of Oslo

Arnstein Vestre

Algorithms for AIS-based maritime collision-avoidance analysis

A data science project on maritime safety

STK-MAT2011 — Project work in finance, insurance, risk and data analysis

Supervisors:
Azzeddine Bakdi and Ingrid Kristine Glad

2019

Abstract

In crowded waters, distinct and well planned evasion manoeuvres must be executed by vessels entering into near-collision situations. International Maritime Organization (IMO) regulations oblige ships above a minimum size to report their navigation data through the Automatic Identification System (AIS). This paper presents algorithms for processing and analysing raw AIS data, with identification of near-collision situations of various types, and aggregation of core statistics on manoeuvres. The algorithms are applied to high temporal-resolution, real world data. The accurate analysis of near-collision situations may help determine safe limits for navigation speed and distance, and describe the necessary manoeuvres expected to be implemented in situations. The results may be used to amend current rules for preventing collisions at sea, and may also provide guidelines for testing navigation risk, and in particular for autonomous vessel collision avoidance.

Acknowledgements

DNV-GL has helpfully provided a set of temporally high-resolution AIS data. Without this, this project would not have been possible.

Abbreviations, nomenclature and concepts

Abbreviations

AIS	Automatic Identification System
COLREGs	Convention on the International Regulations for Preventing Collisions at Sea
COG	Course over ground
CPA	Closest Point of Approach
IMO	International Maritime Organization
LAT/LON	Latitude and Longitude
MMSI	Maritime Mobile Service Identity
RAM	Random Access Memory
SOG	Speed over ground
WGS	World Geodetic System

Nomenclature

t_{CPA}	Time to Closest Point of Approach
d_{CPA}	Distance between vessels at Closest Point of Approach
$d_{\rightarrow CPA}$	Distance from current position to Closest Point of Approach given constant travel at current speed and direction
$\mathbf{v}_i, \mathbf{v}, \mathbf{u}$	Velocity vector for vessels
$\lvert\mathbf{v}_i\rvert, \lvert\mathbf{v}\rvert, \lvert\mathbf{u}\rvert$	Speed over ground for vessels
α_i	Course over ground for vessel i
T_A	A threshold for statistic A
v_{LAT}	Speed in longitudinal direction
v_{LON}	Speed in latitudinal direction
$\mathbf{v}_{app}$	Approach velocity
$\lvert\mathbf{v}_{app}\rvert$	Approach speed
α_{app}	Angle of approach
$\Delta\alpha_M$	Maximum course change
$\Delta\lvert\mathbf{v}_M\rvert$	Maximum speed change

Main concepts

Closest point of approach	The closest point two vessels approaching each other will arrive to at current course and speed.
Near-collision situation	A situation where two approaching vessels will come within an unsafe distance of each other in the near future if no actions are taken.
Yield time	Length of interval during which an evasive manoeuvre is implemented in a near-collision situation.
Yield distance	The distance to CPA at the time which an evasive manoeuvre is initiated (t_1).
Passing distance	The d_{CPA} at the time of exiting from the evasive manoeuvre (t_F).
Approach speed	The speed at which the two vessels in the near-collisions situation is approaching each other.
Max course change	The maximum course change resulting from the implemented evasive manoeuvre.
Max speed change	The maximum speed change resulting from the implemented evasive manoeuvre.
Manoeuvre interactions	The combination of manoeuvre types implemented by the two vessels in a near-collision situation.

Contents

1 Introduction

Approximately 90% of world trade is carried out by sea [14]. Despite the decreasing trend in the number of very serious casualties in areas such as European Union (EU) waters, the number of accidents and near-accidents is still high both in the EU and elsewhere [3, 7, 14], and presents risks for workers and companies. Due to the rising number and size of vessels, collision avoidance in congested waters has evolved as one of the most important concerns for maritime navigational safety [10].

The introduction of Automated Identification System (AIS) monitoring in the shipping industry has presented an inflow of positional data. AIS is a maritime information and safety system which automatically gathers and delivers information between ships and shore [6]. AIS equipment is in high use in the maritime industry, and has since 2002 been mandated by the International Maritime Organization (IMO) for large vessels, as well as for passenger vessels and cargo vessels in international traffic.

1.1 Near-collision statistics: Use and applications

Travel at sea is governed by The International Regulations for Preventing Collisions at Sea (COLREGs) [12]. Still, the rules are somewhat ambiguous, and 56% of maritime collisions are caused by violation of COLREGs rules [2]. Producing accurate estimates of parameters describing near-collision situations at sea may help amend current regulations, in order to increase compliance and safety.

There is extensive effort put into the development and improvement of collision-avoidance decision support systems [2, 15, 9]. These are reliant on assumptions about the situations in question. These assumptions could benefit from real-world parameter estimates in order to refine procedures.

Research suggests that risk perception by officers and crew is crucial for the prevention of marine accidents [13]. Awareness by officers on watch and perception of risk and situations are formed by training, and recommendations for guidelines are stipulated by COLREGs [12]. Improving training procedures for officers on watch, informed by real world estimates of near-collision situation parameters may help increase awareness and risk perception by officers in congested waters.

Human error contributes to more than 80% of ship collision accidents [4], and there is a surge of research and development seeking to automate sea transport. In order to implement automation at sea, robust collision avoidance systems are necessary. Approaches include multi-objective optimization algorithms [5], path-planning using bearing for local obstacle avoidance [1] or potential movement

by other vessels [8], as well as combined approaches [2]. A commonality, and inherent in the concept of autonomousness, is the reliance on algorithms to process data streams in order to make good decisions. These are reliant on assumptions about the environment and other vessel actions, in particular in congested waters. Producing parameter estimates to inform such decisions may help further the development of autonomous vessels at sea.

The present paper sets out to produce estimates on a range of core statistics for near-collision accidents (see Chapter 2), and stratify these by the type of situation and vessels involved. The paper presents methods for calculations of these estimates from identified near-collision situations, based on AIS data.

1.2 Large data sets: Challenges to analysis

Large-size data sets, such as AIS data pose difficulties both in terms of processing power for calculations, and consumption of Random Access Memory (RAM). Large and complex data sets also increase the need for algorithm robustness and testing, as analysts cannot easily verify results naively.

AIS data files can be split and analysed in parts by time and distance. The situations of interest are interactions between pairs or smaller groups of ships within subregions of the geospatial data set. The challenge lies in finding ways of efficiently singling out situations of interest by some relevant metric or condition.

Standard geometry offer tools for studying parametrized lines in two-dimensional space. Abstracting from physical conditions, weather and geographical constraints, application of quite naive equations offer useful tools for analysing data in real-world situations. Applying sufficient tolerances to account for such abstractions, these tools lend themselves to the analysis by offering two useful metrics for near-collision event analysis [9, 10, 11, 14]. This paper uses calculations of Closest point of approach distance (d_{CPA}) and Time to closest point of approach (t_{CPA}) to separate out situations of interest, implementing these as algorithms in optimized Python code.

1.3 Data quality: Removal of noise and error

The main issue for the spatio-temporal analysis of AIS data is determining the distribution of the irregular transmission intervals of transmitters. This constitutes the "sampling time" for the data, which by nature consists of observations that are discrete and asynchronous.

In order to do calculations, data needs to be sorted into near-synchronous intervals. Within these intervals a single record representing each unique vessel is chosen. The size of intervals are chosen as a percentile of the sampling distribution of transmission intervals for the vessels in the data set. The choice of percentile constitutes a trade-off between grouping enough observations in each interval for all unique vessels to be included, and reducing the temporal quality of the data by choosing too large intervals. This paper set the 90 percentile as the threshold determining the interval size, giving 20 second intervals for this data.

In general, AIS data sets are not necessarily of high quality. The differing requirements for transmitter categories, devices of different quality, the presence

of missing, noisy and erroneous data as well as the manual-setting nature of some data fields provide challenges to aggregated analysis. The underlying conditions creating these imperfections are inherent, but deficiencies can still be mitigated. The present data set was of high quality, but data quality analysis, and possible substitution or deletion of records is key to analysis robustness.

Additionally, different AIS data providers give a different range of data fields, and the choice of data fields on which to rely is important. A part of the work in the present paper has been to determine on which fields to rely, and thus which methods are also robust for data sets of lesser quality.

1.4 Outline of paper

The rest of the text is organised as follows:

Chapter 2 presents the methodology used for narrowing down the dataset, analysing data quality and calculating statistics, as well as comments on the software implementation of the methods.

Chapter 3 presents the sample data set, as well as two other similar sample data sets for comparison. The methods for data quality analysis are applied, and a sampling interval length is chosen.

Chapter 4 presents the results of the analysis, in particular estimates of statistics, stratified by situation type and category of vessels involved in interactions. Distributions are also presented. A brief discussion of the findings follow.

Chapter 5 presents conclusions on the material.

Appendix A attaches code written in `Python` for narrowing down AIS data set, computations on the set of unique situations, collection of a subset of records for each situation and computation of statistics for each situation.

Appendix B attaches code written in `R` for the exploratory analysis of the data and data quality analysis, as well as plotting of statistics distributions.

2 Methodology

2.1 Introduction

The following section outlines the main methods and algorithms applied in this paper. The process of filtering, narrowing down and analysing the data is broadly presented in the flowchart in Figure 1.

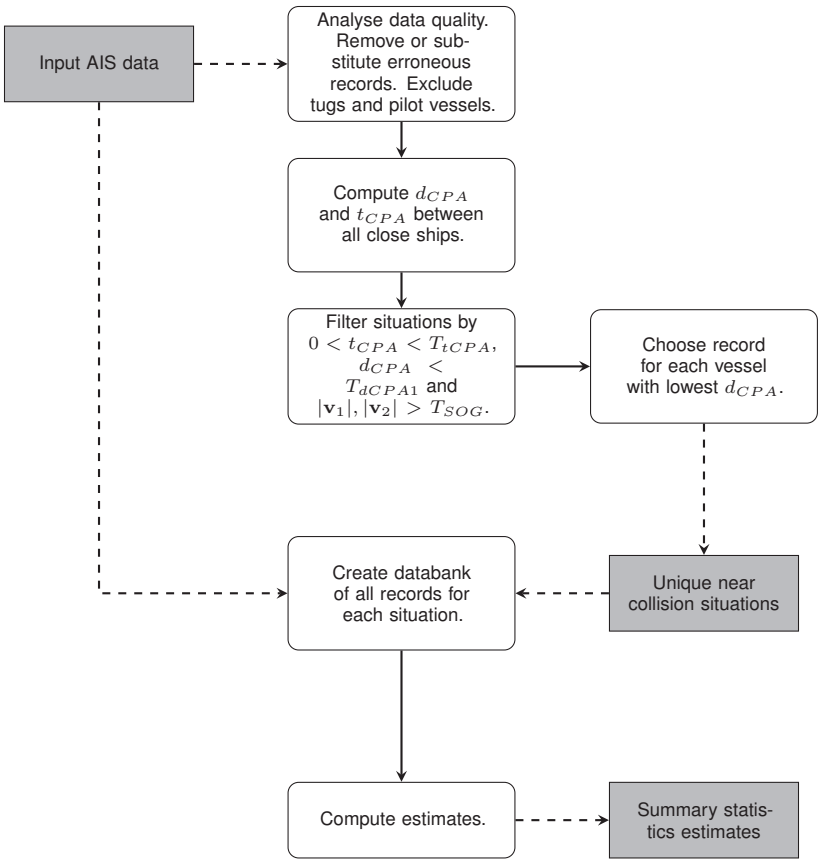

Figure 1: Overview of the process of filtering, narrowing down and analysing the AIS data.

Finding "near-collision" situations, requires a standard for what is to be defined as "near-collision", and how to discern such situations from the AIS

data at hand. For a presentation of the data fields given in standard AIS data, see Chapter 3. The following discusses how to find "near-collision" situations, and how to create algorithms using these expressions to find the situations of interest from a data set of standard AIS data.

2.2 Finding "near-collision" situations

Definitions

The "Closest point of approach" (CPA) between two vessels at any given time is defined as the closest point where the two vessels will arrive to, if the two vessels stay on their current course and keep their current speed. The distance between the vessels at CPA is denoted d_{CPA}.

"Time to closest point of approach" (t_{CPA}) is defined as the remaining time for the two vessels to reach the CPA if no changes in speed (SOG) and/or course (COG) are implemented. If time to CPA is negative, $t_{CPA} < 0$, the two vessels are moving away from each other.

Considering a pair of vessels, a "Near-collision situation" is defined as a situation where if neither of the two vessels implement evasive manoeuvres, the vessels will, in the near future, come within an unsafe distance of each other. Near-collision situations are identified if the CPA distance below a certain threshold ($d_{CPA} < T_{dCPA1}$), in the near future ($0 < t_{CPA} < T_{tCPA}$).

"Sampling time" is defined as the time between two transmissions of AIS data from a single vessel, and "Sampling interval length" as denoting a statistic indicating an assumed fixed sampling time, in which to create "sampling intervals".

Finding CPA using line parametrization

Sunday [11] presents an outline for applying the framework of parametrized lines in two-dimensional space to finding the closest point of approach between two points in motion with a given speed and direction. A brief summary is presented here.

We consider the two lines $\mathbf{L}_1, \mathbf{L}_2$ in their parametric form:

$$\begin{aligned} \mathbf{L}_1 &: P(s) = P_0 + s\mathbf{u} \\ \mathbf{L}_2 &: Q(t) = Q_0 + t\mathbf{v} \end{aligned} \tag{1}$$

The distance between two lines $\mathbf{L}_1, \mathbf{L}_2$ in euclidean two-dimensional space is calculated as:

$$d(\mathbf{L}_1, \mathbf{L}_2) = \min_{P \in L_1, Q \in L_2} d(P, Q) \tag{2}$$

This gives the distance between the two lines as:

$$d(\mathbf{L}_1, \mathbf{L}_2) = \min_{s,t} \mathbf{w}(s, t) \tag{3}$$

where $\mathbf{w}(s, t) = P(s) - Q(t)$.

Consider two maritime vessels, at locations P_0 and Q_0, with velocity vectors $\mathbf{u}, \mathbf{v}$. We set $s = t$, and let t represent time. This is a parametrization of the

two "tracks" on which the vessels are headed, and allows for computation of the distance between the vessels.

We write the position of the two vessels $(P(t), Q(t))$ as:

$$\begin{aligned} P(t) &= P_0 + t\mathbf{u} \\ Q(t) &= Q_0 + t\mathbf{v} \end{aligned} \tag{4}$$

This gives the distance between the vessels at any point in time as:

$$d(t) = |P(t) - Q(t)| = |\mathbf{w}(t)|$$

where $\mathbf{w}(t) = \mathbf{w}_0 + t(\mathbf{u} - \mathbf{v})$, with $\mathbf{w}_0 = P_0 - Q_0$. We find the minimum value of $d(t)$ by finding the t which gives the minimum of the squared distance $D(t) = d(t)^2$. We have:

$$\begin{aligned} D(t) = d(t)^2 &= |\mathbf{w}(t)|^2 = \mathbf{w}(t) \cdot \mathbf{w}(t) \\ &= (\mathbf{u} - \mathbf{v})^\top (\mathbf{u} - \mathbf{v})t^2 + (\mathbf{w}_0)^\top (\mathbf{u} - \mathbf{v})t + (\mathbf{w}_0)^\top (\mathbf{w}_0) \end{aligned} \tag{5}$$

This obtains its minimum when

$$\frac{d}{dt} D(t) = 2t[(\mathbf{u} - \mathbf{v})^\top (\mathbf{u} - \mathbf{v})] + 2(\mathbf{w}_0)^\top (\mathbf{u} - \mathbf{v}) = 0 \tag{6}$$

This solves to give the time to CPA as

$$t_{CPA} = \frac{-(\mathbf{w}_0)^\top (\mathbf{u} - \mathbf{v})}{|\mathbf{u} - \mathbf{v}|^2} \tag{7}$$

Situations where neither vessel is moving, or where both vessels are moving at the exact same SOG and COG (i.e. $\mathbf{u} - \mathbf{v} = \mathbf{0}$) are later excluded. In terms of the algorithm, it suffices to set $t_{CPA} = 0$ for the latter instance. The CPA distance is found as:

$$d_{CPA}(P(t), Q(t)) = |P(t_{CPA}) - Q(t_{CPA})| \tag{8}$$

When $t_{CPA} \leq 0$, the closest point of approach has "already occurred" [1] and the situation is excluded.

AIS data provides longitude (LON) and latitude (LAT) for every vessel, as well as speed over ground (SOG) and course over ground (COG). For the purpose of computing d_{CPA} and t_{CPA} between each pair of ship ($i = 1, 2$), LON_i gives the longitudinal (x) and LAT_i gives the latitudinal (y) coordinates, SOG_i gives $|\mathbf{v}|$ and COG_i the velocity vector angle (α_i), relative to the y axis. The velocity vectors are transformed from polar to Cartesian form as follows:

$$P_0 = [\text{LON}_1, \text{LAT}_1] \tag{9}$$

$$Q_0 = [\text{LON}_2, \text{LAT}_2] \tag{10}$$

$$\mathbf{u} = \text{SOG}_1 \cdot [\sin(\text{COG}_1), \cos(\text{COG}_1)] \tag{11}$$

$$\mathbf{v} = \text{SOG}_2 \cdot [\sin(\text{COG}_2), \cos(\text{COG}_2)] \tag{12}$$

This gives the tools to compute the distance between every pair of ship in the dataset.

[1] By this, the vessels in question will not necessarily have been in a near-collision situation in the past, but they would have been if the two vessels had kept their current speed and course for the appropriate amount of time. If a situation did in fact occur previously for these vessels, it will will for some previous point in time have $t_{CPA} > 0$.

Defining sampling time intervals

AIS transmitters transmit data at different and varying time intervals for each vessel. Signals are transmitted from vessel transmitters, and received by a receiving station onshore. Due to saturation at the point of reception, the interval at which data is recorded is different from area to area. Additionally, it is different from vessel to vessel, and also varies with the speed and rate of turn for each single vessel. Data is transmitted as discrete records and, due to transmission traffic, are not synchronized. Thus, an AIS data set does not provide data for all vessels present and moving within the monitored area for every given point in time. In order to compute d_{CPA} and t_{CPA} for the dataset, it is necessary to define what is regarded as happening "at the same time".

In order to do this, a sampling interval length needs to be defined. Further the data must be subdivided into intervals by this length. Each such interval needs to be large enough that if a sample is made of all records in a given interval, a sufficiently large percentage of the population of actual vessels at sea is included with at least one record in the sample. In terms of calculations, one such record is chosen as a representative record for the vessel within the interval.

The sampling times of a data set can be represented by a distribution. To see an example of such a representation, see Figure 4 on page 20. For a given time window of several intervals, close to all pairs of vessels need to be represented within the same interval for one interval, and most vessels need to be represented within every sampling interval. On the contrary, in order to make precise estimates of d_{CPA} and t_{CPA} for a vessel, data with a certain temporal resolution is needed. A balance should be struck between the number of vessels expected to be included in each interval and the allowed analysis resolution allowed by the interval length.

One approach is to decide on a level of the number of unique vessels allowed to be excluded from each distinct interval. Suggestions could be to choose a level ensuring 90, 95 or 99 % of all vessels appear in each interval. For the purpose of this paper a 90% threshold is chosen, giving the empirical 90th percentile of the cumulative sampling time distribution as sampling interval length.

Pre-processing and implementation

In order to speed up computations, all non-unique records are excluded from each interval, leaving a sample of unique vessels within each interval. Applying this to the sample data reduces the number of records from 18.5 million to 8.6 million records, with a 20 second interval length chosen by the method described above (see further Chapter 3).

In order to filter out situations without needing to compute the euclidean distances between vessels, all records more than 2 kilometres apart in latitudinal or longitudinal directions are passed by the algorithm. For the remaining vessel-pairs, t_{CPA} and d_{CPA} are computed between these pairs within each interval. All records with $t_{CPA} < 0$ or $t_{CPA} > T_{tCPA}$ are discarded, and the rest are taken as candidates for records being part of a near-collision situation. The current paper set $T_{tCPA} = 40$ minutes as the threshold. Assuming no two vessels encounter in a near-collision situation more than once, the most severe

situation (defined as the one with lowest d_{CPA}) is then uniquely selected for each vessel-pair.

Different data fields are given in different units. SOG is given in knots, COG in clockwise degrees relative to north, and latitude and longitude in degrees relative to the World Geodetic System (WGS). For the purpose of the computations described above, distance calculations are done in degrees, while output is given in meters. SOG is output in m/s and time in seconds.

The code in full can be accessed in Appendix A.

2.3 Constructing a situation databank

Definitions

A "unique near-collision situation record" is a record identifying one near-collision situation between two identified maritime vessels, not providing more information than the closest point of approach distance at the point when this distance is at its lowest.

A "near-collision situation databank" is constructed as a collection of records containing all AIS data for the two vessels in the time period around the unique near-collision situation record, thus providing all necessary information about each near-collision situation, necessary for estimating statistics.

From unique situations to databank

Having identified near-collision situations by unique records, a databank of all records associated with each situation is created. All records for the two vessels involved in each situation are retrieved from the original data set, and the records preceding and succeeding the unique near-collision situation record by one hour are inserted in the databank.

The code for implementing this can be found in Appendix A.

Synchronizing time scales

As discussed previously, AIS records are transmitted at different time intervals, and are not synchronized between vessels. In order to be able to analyse situations in the databank, the time scale for the records of the pairs of vessels involved in each situation need to be synchronized.

Situations are synchronized starting with the records of the vessel for which the number of records is the smallest. For each record, the closest record in time from the other vessel is chosen, discarding all records with no corresponding observation 5 seconds before or after. The operation outputs two sequences and a pairwise, synchronized set of records for each situation. A joint time sequence is computed as the mean of the two synchronized time sequences.

The code for implementation can be found in Appendix A.

2.4 Cleaning the data

In order to produce meaningful statistics, records and situations which by the given criteria get defined as "near-collision", but which in fact are not, need to be discarded.

Tugs and pilots

Some seemingly "near-collision" situations are the result of normal behaviour. Two examples of such are the operations of tugs and pilot vessels.

Tugs are vessels tasked with towing larger vessels in tight waters and within harbours. Tugging involves operations close to other vessels, and tugs will thus often exhibit behaviour that falls within the near-collision threshold, when in fact the vessels are operating as intended.

Pilot vessels are vessels transporting pilots – officials tasked with guiding large vessels in tight waters – safely to harbour or through challenging waterways. This will by construction always include approaching a larger vessel and aligning in order to transfer the pilot, and will produce a near-collision situation when in fact the vessel is operating as intended.

The ideal way to exclude such ships involves excluding all vessels for which the vessel type is designated as tug or pilot. If this information is not given, a solution is to exclude all vessels with navigational status 11 ("power-driven vessel towing astern") or 12 ("power-driven vessel pushing ahead or towing"). This solution is incomplete, as the status of a vessel is manually recorded, and not always correct.

For the purpose of the present paper, as vessel type is not given in the sample data, the field "Vessel risk category", indicating the type of cargo transported by the vessel, is used to exclude such vessels. By inspection, vessels exhibiting tug- or pilot vessel-like activity consistently fall into the categories "Other activity" or "Missing information", and by excluding all records which fall into this category, effectively excludes such situations.

Missing data

AIS data is transmitted with different levels of data richness. Some transmitters provide a wide range of information, while others only provide core information, such as position, course and speed. For some types of analysis, knowing length and heading of the vessels could provide interesting information, such as for making length corrections on thresholds on d_{CPA}, producing a better proxy for the vessel's manoeuvrability.

Examining the occurrence of missing data is key to assessing the data quality. A short examination of the presence of missing data in the present data set is made in Chapter 3. As a non-negligible percentage of records in this data set is missing length and heading, this paper does not make use of these fields.

Erroneous data

As with missing data, AIS transmitters may also provide erroneous data. This may occur in all data fields, such as positional data outside the geographical area of analysis, SOG above logical limits for the vessel (this paper uses a 50 knot threshold to filter the data), COG or heading above 360° (and in particular at 511°).

Exclusion of erroneous data is ideally implemented as the first action on the original data. The present paper excluded these values at the stage of creating the databank. A summary of erroneous values for the data sample at hand can be found in Chapter 3.

Platforms

The goal of the present work is to estimate various statistics for the population of moving maritime vessels. But AIS transmission systems are also utilized by platforms and other stationary entities. In order to not have data corrupted by such transmissions, it is necessary to exclude all such records. Stationary entities are labelled with Maritime Mobile Service Identity (MMSI) numbers of non-8 digits. Excluding records with MMSI below 10 000 000 or above 99 999 999 expunges these from the data.

2.5 Calculation of relevant statistics

The following statistics are estimated in order to assess properties of collision avoidance manoeuvres for vessels involved in near-collision situations.

- Interaction manoeuvre
- Yield time (Time spent on manoeuvre from start to exit)
- Yield distance (Distance to CPA at point of entering yield manoeuvre)
- Passing distance
- Approach speed
- Max course change (due to evasion manoeuvre)
- Max speed change (due to evasion manoeuvre)

A subset of the estimates are stratified by the type of situation, defined by the angle of approach between the vessels, and type of vessels involved. The following describes the methods for categorization and estimation.

Determining yield manoeuvre interval

In order to compute the listed statistics, it is necessary to determine for which records the relevant information is stored. This is done by identifying the "yield manoeuvre interval". The yield manoeuvre interval is defined by a starting time, t_1 and exit time, t_F. The starting time is defined in one of three ways:

1. If d_{CPA} never passes below 10 meters, the starting time t_1 is taken to be the time when the d_{CPA} is at its minimum.
2. If d_{CPA} passes below 10 meters during the situation, the starting time t_1 is taken to be the first point in time when d_{CPA} passes below this threshold during the interval in which t_{CPA} is positive
3. If d_{CPA} passes below 10 meters, but t_{CPA} is negative at this point, the starting time t_1 is taken to be the point in time when d_{CPA} is at its minimum while t_{CPA} is still positive.
4. If td_{CPA} never passes below a threshold T_{dCPA2}, in the present paper set to $T_{dCPA2} = 50$ meters, the situation is discarded.

The exit time is defined as the first point in time after t_1 when t_{CPA} is negative, indicating that the vessels are now in the clear.

These two time points indicate the yield interval. Furthermore, the "halfway point" is calculated as the mean value of the two interval boundaries, $t_M = \frac{t_1+t_F}{2}$.

Situation categories

Near-collision situations are diverse, both in terms of how and where they occur, and which type of vessels are involved. The following summarizes two important ways of categorizing the situations: Based on the angle of approach, and based on which types of vessels are involved in the situation.

Categorization by angle of approach: Overtaking, crossing and head-on

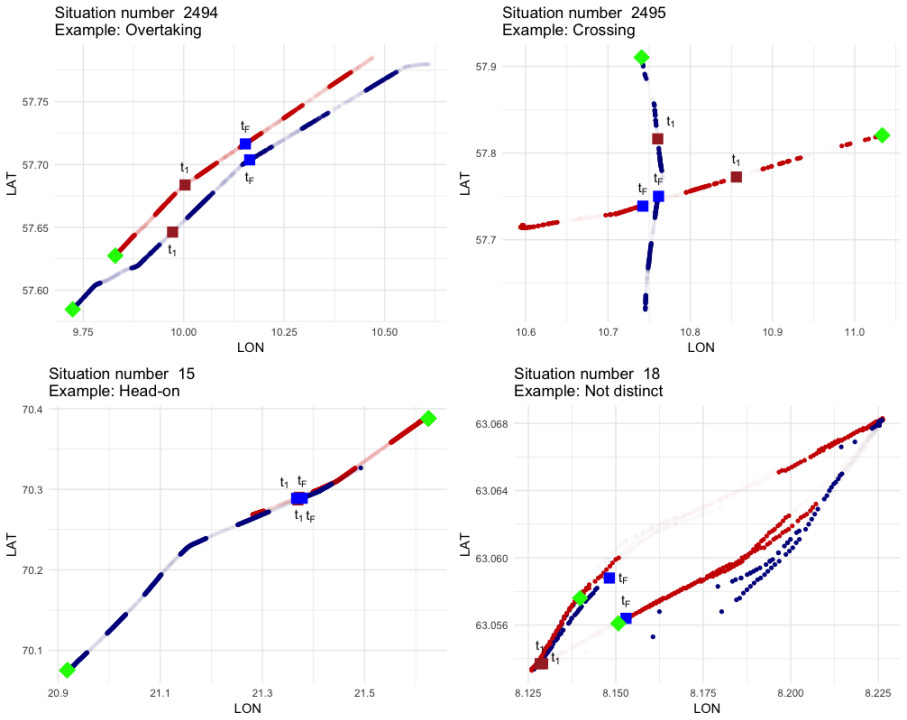

Figure 2: Panels 1-3 show three types of situations by the definitions in COLREGs: Overtaking, Crossing and Head-to-head. Panel 4 shows a situation that is less visually distinct, and where a numerical statistic to classify the situation is necessary. The four illustrations are taken from situations in the data set. Green diamond annotates starting point of vessel, red box position at t_1 and blue box position at t_F. Situation numbers refer to candidates for near-collision situations prior to final filtering, and thus goes from 0 to 2530, whereas the total number of near-collision situations is 1394.

The International Maritime Organization (IMO) in 1972 adopted a convention laying out the rules for travel at sea. In particular rules 10 to 18 govern traffic separation schemes and the actions of vessels in sight of each other, and further in particular rules 13, 14 and 15 govern the actions of vessels "overtaking" another vessel, encountering another vessel in a "head-on" situation or in a "crossing" situation [12].

The convention states that in the event of an overtaking, the overtaking vessel shall "keep out of the way of the vessel being overtaken" [12]. Overtaking is further defined such that two vessels are in an overtaking situation when the angle of approach between the vessels are between 135° and 225°.

The convention further states that in the event of a vessel meeting another in a head-on situation, defined as being on "reciprocal or near-reciprocal course", "shall alter her course to starboard so that each shall pass on the port side of the other." [12]. The angle of approach classifying a situation as head-on is not officially defined, but can be taken to be between 0° and 15°, as well as 345° and 360°.

A crossing situation is defined as all other situations when two power-driven vessels meet, and the convention stipulates that "the vessel which has the other on her own starboard side shall keep out of the way and shall, if the circumstances of the case admit, avoid crossing ahead of the other vessel." [12].

In addition to stipulating the required actions at different angles of approach, the convention states that any action to avoid collision shall be made in ample time, be large enough to be readily apparent to another vessel, that alteration of course alone is preferable if there is sufficient sea-room available, and that if necessary, a speed change action shall be implemented as a reduction of speed, or reversing the means of propulsion [12]. Thus, actions taken to avoid collision are to be made in due time, in a large enough manner, and course manoeuvres are preferred over speed manoeuvres, and further reduction of speed is the appropriate speed change manoeuvre.

All three types of situations are present in the data, as can be seen from panels 1–3 in Figure 2. The data also includes less visually discernible situations, such as the situation visualized in panel 4 in Figure 2. Thus, it is necessary to set out a clearly defined statistic for measuring the angle of approach between two vessels, in order to classify the situation correctly.

In order to quantify the angle of approach, COG for both vessels in the 20 seconds preceding the starting point t_1 is used to calculate the mean angle of approach. This is calculated as the difference between the mean COG for the two vessels, $\alpha_{app} = \overline{\alpha_1} - \overline{\alpha_2}$.

Categorization by vessel types involved in encounter

Table 1: Vessel types included in the Norwegian AIS data set.

T	Chemical Tankers	**C**	General cargo ships
	Gas tankers		Container ships
	Bulk Carriers		Ro-Ro Cargo Ships
	Oil product tankers		Refrigerated Cargo Ships
	Crude oil Tankers	**F**	Fishing vessels
O	Offshore supply ships	**P**	Passenger ships
	Other service offshore vessels	**U**	Cruise ships

The sample AIS data set provides a categorical data field indicating the type of vessel implied by the type of cargo carried. In other data sets, more detailed information on vessel type is included. Mean estimates for a subset of

the various statistics are calculated for each type of encounter, categorized by the type of vessel involved.

There are 14 types of vessels, the vessel types are presented in Table 1. These can be further grouped, as shown in the table. Mean estimates are made for the following statistics, and used in comparison with estimates for the whole data set.

- Yield time
- Yield distance
- Passing distance
- Approach speed

Estimation of the various statistics

Below, the main approaches to producing estimates for the listed statistics are presented. The statistics are presented in Chapter 4 by distributions for the data set as a whole and stratified for each of the three situations defined by COLREGs.

Yield time

The time spent on the evasive manoeuvre is calculated using the starting time t_1 and the exit time t_F, and is simply the difference between these two time points:

$$\Delta t = t_F - t_1$$

This statistic is estimated in order to ascertain how much time is used by the vessel to execute the manoeuvre and exit from the situation.

Yield distance

The distance to CPA ($d_{\rightarrow CPA}$) at point of yield is estimated for both vessels, and is calculated as the distance from the current position of the vessel to the closest point of approach if the vessel keeps on current course with the current speed. The statistic is calculated at t_1, when vessels start exiting the situation.

$$d_{\rightarrow CPA} = |SOG_i| \cdot t_{CPA}$$

The statistic indicates the distance from the near collision at time of action for the vessel, and thus gives an indication of how early action was taken.

Passing Distance

The passing distance is calculated as the CPA distance at the time of exit from the evasive manoeuvre, t_F, that is when t_{CPA} turns from being positive to negative. The statistic is calculated as:

$$d_{pass} = |P(t_F) - Q(t_F)|$$

The parametrized equations for P and Q are given in equation (4). This statistic gives an indication of how close the two vessels were from each other at the time when the vessels successfully evacuated the situation, and is a good indication of how unsafe the situation was allowed to become.

Approach speed

The approach speed is calculated using the mean speed and course over ground for the two vessels in the first half of the evasive manoeuvre, from t_1 to t_M. The statistic is calculated as:

$$|\mathbf{v}_{app}| = \sqrt{\overline{v}_{LAT}^2 + \overline{v}_{LON}^2}$$

where v_{LAT} and v_{LON} are given by:

$$\overline{v}_{LON} = \overline{|\mathbf{v}_1|} \cdot \sin(\overline{|\alpha_1|}) - \overline{|\mathbf{v}_2|} \cdot \sin(\overline{|\alpha_2|})$$
$$\overline{v}_{LAT} = \overline{|\mathbf{v}_1|} \cdot \cos(\overline{|\alpha_1|}) - \overline{|\mathbf{v}_2|} \cdot \cos(\overline{|\alpha_2|})$$

Here, $\overline{|\mathbf{v}_i|}$ indicates mean speed over ground between t_1 and t_M, and likewise, $\overline{|\alpha_i|}$ indicates mean course over ground in the same time period. The statistic gives the speed with which the two vessels are approaching each other.

Max course change

The maximum course change is calculated as the difference between the highest and lowest value for the course over ground implemented by the vessel during the evasive manoeuvre (between t_1 and t_F). This is calculated as:

$$\Delta\alpha_M = \max_{t\in[t_1,t_F]}\{|\alpha_i|(t) - \min_{t\in[t_1,t_F]}(|\alpha_i|(t))\}$$

The maximum course change indicates whether the vessel utilized a course change manoeuvre that is large enough to be readily apparent to another vessel observing visually or by radar.

Max speed change

The maximum speed change is calculated as the difference between the highest and lowest speed over ground attained by the vessel during the evasive manoeuvre. This is calculated as:

$$\Delta|\mathbf{v}_M| = \max_{t\in[t_1,t_F]}\{|\mathbf{v}_i|(t) - \min_{t\in[t_1,t_F]}(|\mathbf{v}_i|(t))\}$$

The maximum speed change indicates whether the vessel utilized a speed change manoeuvre that is large enough to be readily apparent to another vessel observing visually or by radar.

Manoeuvre interactions

The types of manoeuvres by the pair of vessels may be categorized as in Table 2.

This produces 16 different manoeuvre interaction types, which can be narrowed down to 9 categories of manoeuvre interactions. For future use

Table 2: Types of manoeuvres for ships.

	Speed change	Course change
Vessel 1	(y/n)	(y/n)
Vessel 2	(y/n)	(y/n)

(see Chapter 4) situations are labelled from A to D, where A indicates no action taken, B indicates one type of action taken (course or speed) by one or both vessels, C indicate mixed strategies (one vessel takes speed and the other course action, one vessel takes both types of action, and further mixed combinations) and D indicates both ships use both types of manoeuvres. The different narrowed-down categories are described in Table 3.

Table 3: Types of manoeuvre interactions.

A	No recognizable yield action taken
B1	Speed action taken only by one vessel
B2	Course action taken only by one vessel
B3	Speed action taken by both vessels (but no course action)
B4	Course action taken by both vessels (but no speed action)
C1	One vessel takes both speed and course action, other no action
C2	One vessel takes course action, the other speed action
C3	Mixed: One vessel takes both actions, the other takes one action
D	Both vessel takes both speed and course action

The type of manoeuvre interaction is defined by calculating the time derivative of the COG and SOG, thus obtaining an estimate for the acceleration in speed and change of course over time for the two vessels, $\frac{d}{dt}|\mathbf{v}_i|$, $\frac{d}{dt}|\alpha_i|$. Using t_1 to define the starting point of the evasive manoeuvre, two thresholds T_j are defined as the maximum "normal" acceleration and maximum "normal" course change. These are found as:

$$T_{dCOG} = \max_{t<t_1}\left(\frac{d}{dt}|\alpha_i|(t)\right)$$
$$T_{dSOG} = \max_{t<t_1}\left(\frac{d}{dt}|\mathbf{v}_i|(t)\right)$$

Using these thresholds, if $\frac{d}{dt}|\mathbf{v}_i|$ passes above T_{dSOG} during the evasive manoeuvre (between t_1 and t_F) for ship i, a speed change manoeuvre is registered for ship i. Similarly, if $\frac{d}{dt}|\alpha_i|$ passes above T_{dCOG} during the evasive manoeuvre for ship i, a course change manoeuvre is registered for ship i. This statistic indicates what types of manoeuvres have been applied by the vessels in each situation, and allows us to investigate what types of manoeuvre interactions are more frequent.

It should be noted that when a vessel implements a course change, a reduction in speed will often follow. Thus, by these measures, an action instance which

is in fact a clean course change may register as a double (C1) or mixed action (C3).

2.6 Comments on programming implementation

In the following, some short comments on the technological aids and programming languages utilized to implement the above algorithm.

Choice of programming language: Python

The main part of the code for this project is written in `Python`, using the `Pandas` module in order to handle tabular data, as well as `Numpy` to do vectorized calculations on arrays and `Numba` to be able to speed up computations by "just-in-time"-compiling the most computationally heavy parts of the operations to the faster programming language `C++`. The `Numba` module also allows for running certain elements of the code in parallel, further increasing computational speed somewhat, but more importantly allowing for possible scale-up of computations if desirable. This computational boost is most relevant for those computations applied to the whole data set, in the process of finding candidates for "near collision" situations, and less necessary for the process of estimating statistics.

In order to test and debug the code, `iPython` and `Jupyter notebooks` have been used, in addition to writing code to script. The `Python` code applied in this project can be found in Appendix A

Exploratory analysis and illustration: R

In order to quickly do exploratory analysis of the data set, the statistical programming language `R` has been used, in particular to produce data quality summaries of the data (see Chapter 3), as well as making the figures found later in this paper, and inspecting situations visually by plotting coordinates to map (an example of this can be seen in Figure 2). The libraries `data.table` and `dplyr`, allowing for fast handling of large data sets, have been useful. These allow for commands that are written in the faster programming languages `C/C++`. The `data.table` library allowed for the fastest exploration of the full 18.5 million data set.

Furthermore, libraries such as `ggplot2` and `OpenStreetMap` has enabled visualization and plotting to map. The `R` code applied in this project can be found in Appendix B.

Storage of data: HDF5 vs. CSV

In storing large data sets, the format of intermediate storage becomes important, in order to speed up work flows, decrease total hard disk and memory usage. The data supplied was provided in `.csv` format, and intermediate storage was done in `.hdf` format. The latter is a serialized storage format that allows for fast loading, reducing read times substantially, and allowing for reading only part of the data set if the analyst is constrained by RAM. This was not necessary in this project, but would be a necessity for a scaled-up version of the project.

3 Description of the data

In this section, a brief description of the data is presented, as well as a summary of data quality in the present data set, compared with two other data sets at hand, a presentation of sampling times in the data, and choice of sampling interval length.

3.1 On AIS data

AIS data is centered around positional records of latitude and longitude, recorded over time. Additionally, the AIS transponder of any vessel may record and transmit any number of other fields of data. Typical fields for AIS data, most of which are included in the data set forming the grounds for the application of the methods in Chapter 2 is presented in Table 4.

Table 4: Fields included in typical AIS data sets, such as the one used for the application of algorithms in this paper.

Field name	
Transmitter ID (MMSI)	Vessel length
Date and time of transmission (Time)	Vessel width
Longitude (LON)	Vessel draught
Latitude (LAT)	Other dimensions
Speed over ground (SOG)	Vessel ID (IMO)
Course over ground (COG)	Vessel call sign
Boat alignment (Heading)	Vessel type (substituted)
Navigational status	Type of transmitter (Source type)
Name of vessel	

Transmission of AIS data has been mandated by the International Maritime Organization (IMO) since 2002 for vessels of a certain size and of certain activity.

3.2 Sample data

The present paper implements the presented algorithms on a high-resolution data-set provided by an external source (sample 3). The data is in the following presented and compared to two other AIS data sets.

Sample 1: United States Zone 20

Data set sample 1 consists of 2.7 million records, recorded throughout one whole month, in January 2017, in United States Zone 20. The data is retrieved from MarineCadastre (link).

Sample 2: Denmark

Data set sample 2 consists of 5.8 million records, recorded during one day, on November 19, 2018 in Danish waters. The data is retrieved from the Danish Maritime Authority (link).

Sample 3: Norway

Data set sample 3 consists of 18.5 million records, recorded during one day, on August 20, 2015, in Norwegian waters. Figure 3 provides a snapshot of the unique vessels present within a 20 second interval at 10:20 in the morning on the day in question. The data set is provided by the maritime classification society DNV-GL.

3.3 Data quality and comparison

Table 5: Data quality comparison of the three data set samples.

Sample	1		2		3	
Data type	**Miss**	**Err**	**Miss**	**Err**	**Miss**	**Err**
SOG	0%	0%	6.0%	0.1%	0%	0.1%
COG	0%	0%	9.6%	0.1%	0%	0%
Heading	0%	55%	22.3%	0%	0%	18.0%
Length	22.4%	-	10.1%	-	4.1%	-
LON/LAT	0%	-	0%	-	0%	6.0%
Status	2.7%	-	15.1%	-	7%	4.2%
MMSI (platforms)	-	0%	-	0%	-	0.05%

Table 5 presents a comparison of the data quality of the three data sets [2].

AIS data have, for reasons presented in Chapter 1, varying degrees of quality. In particular, certain fields may be missing for a proportion of the records in the data set, or the values recorded may be erroneous. Examples of this is positional data recorded outside of the area in question (resulting from measurement error or transmission error), SOG exceeding the maximum attainable speeds of sea vessels (the present paper has used a threshold of 50 knots as cut-off point) or COG or heading exceeding 360°. Additionally, some of the data may be

[2] A hyphen indicates the question was not explored for US and DK data. For MMSI the number of irregular occurrences is shown. For Navigational status, "missing" indicates data is excluded and "erroneous" indicates values outside of the $[0, 15]$ range of possible statuses in the data set. US data is to some extent pre-filtered, and trimmed to 20 time zones, so the table indicates a lower percentage of missing data than what is the case, and LON/LAT error values are excluded by construction.

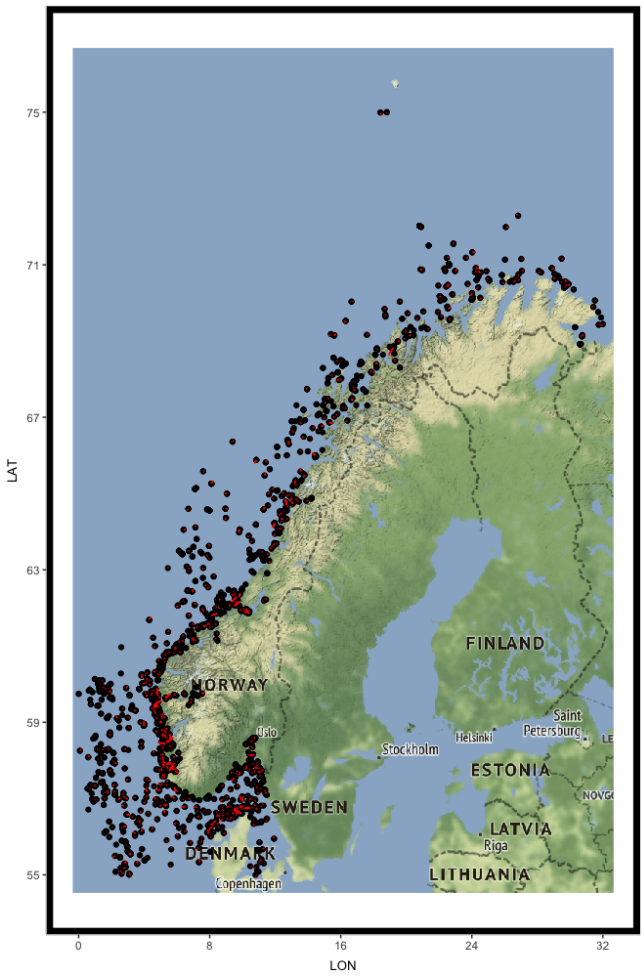

Figure 3: Map of unique seafaring vessels within a 20 second interval at 10:20, August 20, 2015. Red line segments annotate direction and speed of travel.

transmitted from platforms or other stationary AIS transmitting entities. These are generally identified by having an MMSI number of more or less than 8 digits.

For the core fields (LAT, LON, Time, SOG and COG), incidence of missing records is less frequent, even though data is still noisy. Missing data is more frequent for the fields *heading* and *length*.

Vessel type or navigational status is used in order to filter out pilots and tugs (as discussed in Chapter 2). Navigational status is often erroneous, as it is set manually. For this reason, most navigational status is recorded as "Under way using engine", while this might in fact not be the activity at hand. Vessel type is less often erroneous. The data set at hand did not provide this. This was substituted with a similar data field, "Vessel risk category", relating to the cargo carried by the ships, enabling exclusion of tugs and pilot vessels.

3.4 Sampling time interval distribution

Table 6: Sampling time statistics for three data set samples.

Sample	1	2	3
Mean	187.1 s	17.3 s	77.5 s
Median	69 s	10 s	9 s
90th percentile	180 s	20 s	21 s
95th percentile	211 s	41 s	49 s
99th percentile	560 s	181 s	361 s

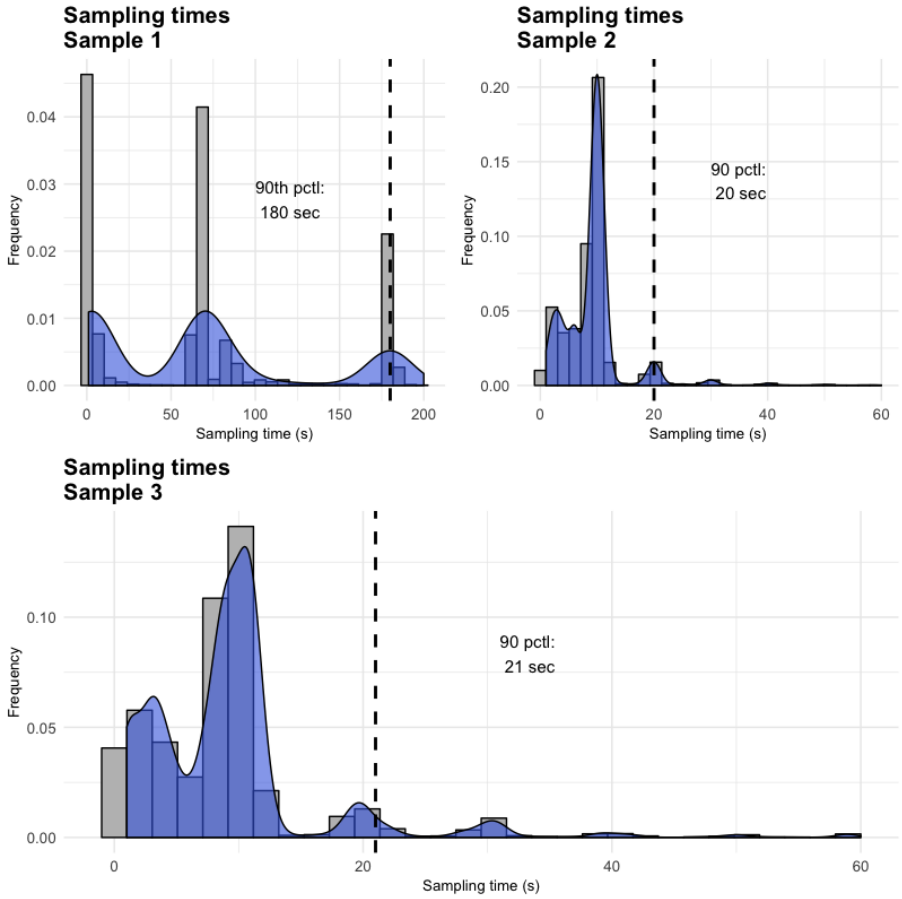

Figure 4: Sampling time distributions for the three sample data sets. We see that a 20 second interval will include approximately 90 percent of the records in each interval for the data sample 3.

Central to the algorithm described in Chapter 2 is the concept of the assumed "sampling interval length" of the data set. In order to define this interval length,

the distribution of the sampling time for the individual records in the data sets needs to be ascertained. Statistics for the sampling times for the three data sets are presented in Table 6 and the distributions in Figure 4. With the 90th percentile as an acceptable threshold, the sampling time distribution for the Norwegian data support choosing an interval size of 21 seconds. The size was approximated by choosing a 20 seconds interval length for application in the algorithm.

3.5 Summary of sample data

In summary the Norwegian AIS data is of adequate quality to work with, and it is not necessary to implement substitution or interpolation of data in order to have sufficient quality for analysis. The sampling time distribution justifies a 20 second time interval for use with the presented algorithm.

4 Results

The following presents results from the analysis presented in Chapter 2. The results are presented first for all situations (found in Norway, August 20, 2015 using data sample 3), then stratified by situation types and lastly by category of vessels involved in the situation. For the first part, distributions for the different statistics are presented. For the latter, mean estimates for a subset of the statistics for each category is provided.

4.1 Distribution of core statistics

Core statistics

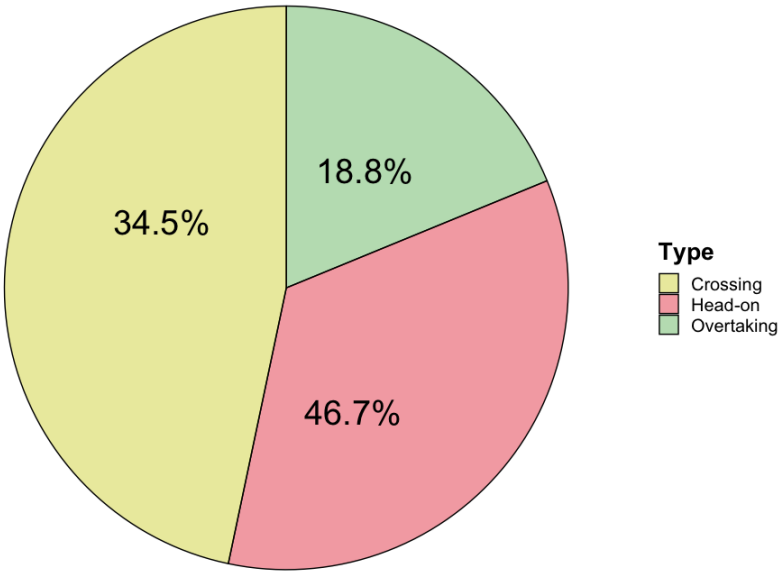

Figure 5: Proportion instances of near-collision situations by COLREGs situation type.

Panel 1 of Figures 7 to 12 show the distributions of the six core statistics, while Figure 6 shows the types of manoeuvre interactions taken, categorized as described in Chapter 2. Figure 5 summarizes the proportion of instances of each of the three COLREGs type situations.

Table 7: Legend for manoeuvre interaction types.

A	No recognizable yield action taken (not included in figure)
B1	Speed action taken only by one vessel
B2	Course action taken only by one vessel
B3	Speed action taken by both vessels (but no course action)
B4	Course action taken by both vessels (but no speed action)
C1	One vessel takes both speed and course action, other no action
C2	One vessel takes course action, the other speed action
C3	Mixed: One vessel takes both actions, the other takes one action
D	Both vessel takes both speed and course action

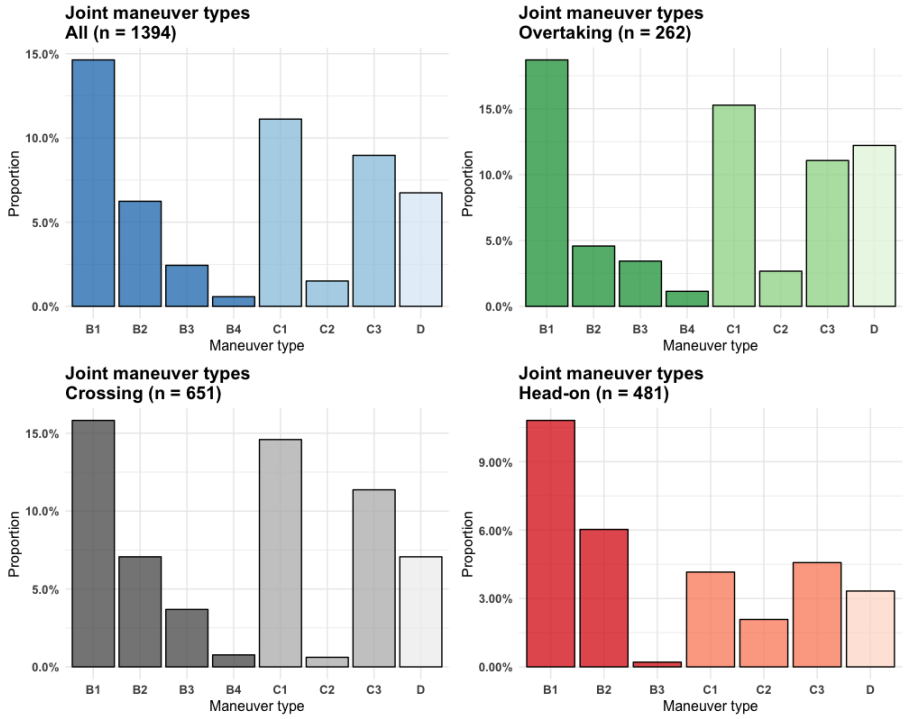

Figure 6: Types of manoeuvre interactions for the full data set and subsets of different situation types (overtaking, crossing, head to head). Records of "No action" (A) are not included.

The mean yield time is 719 seconds, while mean yield distance is 5506 meters. The distributions are heavily skewed, with long and heavy positive tails. Similarly, passing distance has a mean at 1113 meters, approach speed at 20.1 m/s (appx. 39.1 knots), maximum course change has estimated mean

at 21.8°, and max speed change at 2.89 m/s (5.61 knots). The distributions are generally skewed, with a long and heavy positive tail, except for approach speed, which is closer to having a bell shaped curve.

Manoeuvre types

In Figure 6, panel 1, the distribution of manoeuvre interactions are given, in those cases where manoeuvres are registered. The majority of situations where action is taken are situations where one vessel takes speed or course change action exclusively, or where one vessel takes both types of actions (B1, B2 og C1). Secondly, there is a certain proportion of situations where actions are mixed or involves speed and course actions from both vessels.

4.2 Distributions of core statistics - by situation type

Figures 7 to 12, panels 2-4, show the distribution of core statistics for the subset of situations in the different COLREGs classified situations: Overtaking, Crossing and Head-on. As can be seen from Figure 5, there is a plurality of situations in the "Crossing" category ($n = 651$). This is not surprising, as the crossing category covers the largest interval of degrees for crossing angles. Furthermore, there are $n = 481$ head-on situations and $n = 262$ overtaking situations.

Core statistics

Yield time

Figure 7 shows that vessels meeting in overtaking have a higher mean and a more positively skewed distribution than the average for the data set as a whole, at 1058 seconds. On the contrary, the distribution for vessels in a crossing situation is quite similar to the data set average. The largest discrepancy can be found for vessels partaking in head-on situations, where the mean yield time is 363 seconds, and thus distinctly lower than the data set average.

Yield distance

Figure 8 shows that vessels in a overtaking and crossing situation have higher estimated yield distances than the data set as a whole,at 8350 and 6758 meters. Overtaking situations have a similarly more skewed distribution as is the case with yield time. For vessels in a crossing situation, the estimate for mean yield distance is slightly higher than for the whole data set, but with a similarly shaped distribution. For the vessels in a head-on situation, the estimated yield distance is distinctly lower than the data set average, at 2504 meters.

Passing distance

Figure 9 shows that for vessels in overtaking, the estimate for mean passing distance is somewhat higher than for the data set as a whole, at 1641 meters, while the estimate for vessels in a crossing situation is quite similar to the data set average, at 1279 meters. For the vessels in head-on situation, the passing

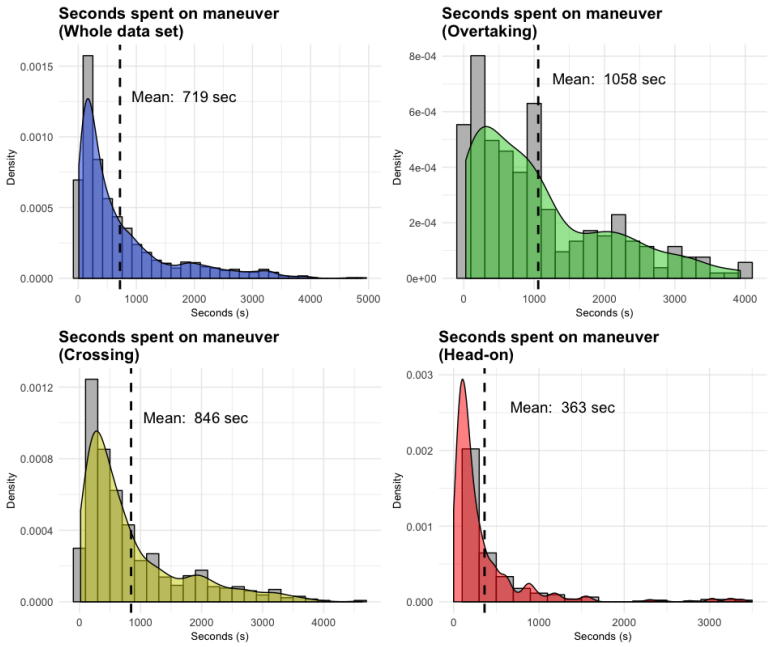

Figure 7: Distribution of yield times for whole data set (panel 1) and the three COLREGs situation types (panels 2-4).

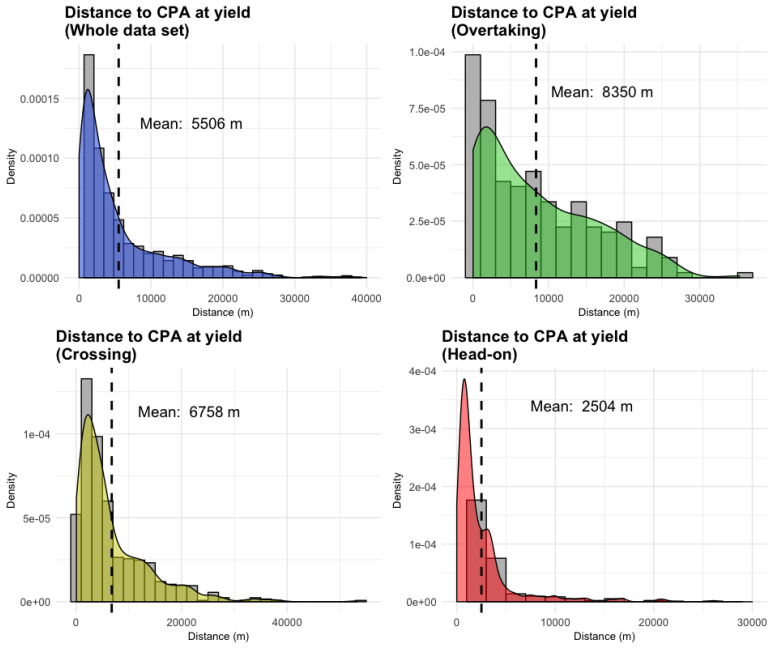

Figure 8: Distribution of yield distances for whole data set (panel 1) and the three COLREGs situation types (panels 2-4).

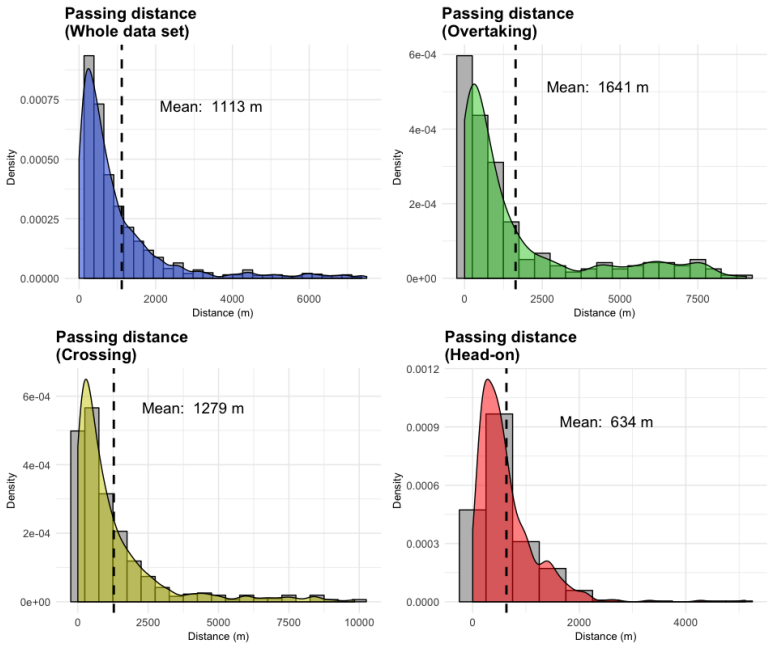

Figure 9: Distribution of passing distances for whole data set (panel 1) and the three COLREGs situation types (panels 2-4).

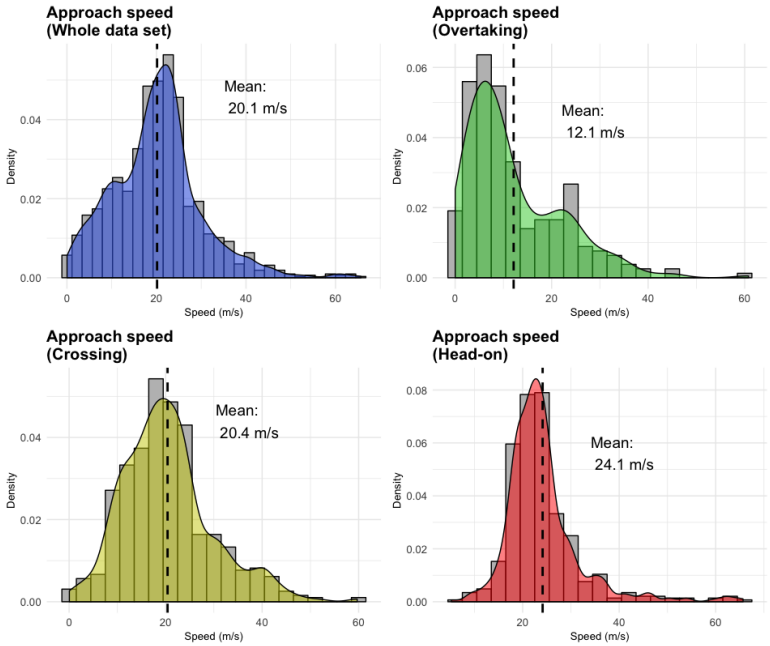

Figure 10: Distribution of approach speeds for whole data set (panel 1) and the three COLREGs situation types (panels 2-4).

distance is estimated distinctly lower, at 634 meters, with a somewhat less positively skewed distribution than the data set as a whole.

Approach speed

Figure 10 shows that for approach speed, vessels in overtaking have a distinctly lower estimated mean at 12.1 m/s (23.5 knots), while for vessels in head-on it is distinctly higher at 24.1 m/s (46.8 knots). The estimated mean approach speed for crossing situations is close to the data set average at 20.4 m/s (39.7 knots). The approach speed retains its non-skewed form for all the three situation types, except for overtaking situations, which have two peaks.

Max course change

Figure 11 shows that for maximum course change, the estimated mean for vessels in overtaking is higher than for the data set as a whole, at 31.7°, and similarly for vessels in crossing at 28.0°. The distributions are more skewed with a heavier positive tail for these two situation categories. For vessels in head-on situations on the other hand, the estimate is lower, at 8.17°, and with a less heavy tail than the other two categories.

Max speed change

Figure 12 shows that for maximum speed change, the estimate for vessels in overtaking is higher than the data set as a whole, at 4.19 m/s (8.14 knots), and likewise with the estimate for vessels in crossing situation. Both the distributions have a heavier positive tail, and seem to have a substantial number of vessels with a speed change at close to 0 m/s. In the instance of crossing vessels, there is a two-peaked distribution, one with a high incidence of max speed change at close to 0 m/s, and one at approximately 6 m/s (appx. 11.5 knots). For the vessels in head-on situations, the estimate for mean maximum speed change is at 0.849 m/s (1.65 knots), distinctly lower than the two other situations, and than the data set as a whole.

Manoeuvre interactions

Figure 6 shows that for vessels in an overtaking situation, instances where one vessel does a speed manoeuvre (B1) or where one vessel does a speed and course manoeuvre C1) is somewhat higher than the data set as a whole.

For vessels in a crossing situation the distribution of manoeuvre types is more similar to the data set distribution, except for a higher incidence of situations where one vessel take both speed and course action (C1), as well as a somewhat higher incidence of mixed actions (C3).

For vessels in head-on situations, a smaller proportion of vessels seem to take action, and absolute numbers are lower relative to the total number of situations classified as head-on. No vessels take unilateral course action (B4). The proportion of joint and mixed action (C2, C3, D) are smaller than for the data set as a whole.

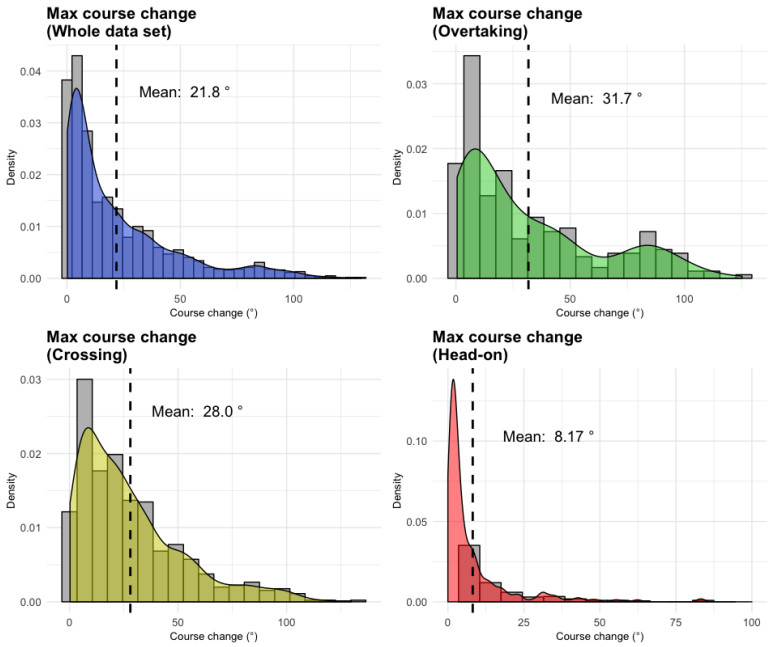

Figure 11: Distribution of maximum course change for whole data set (panel 1) and the three COLREGs situation types (panels 2-4).

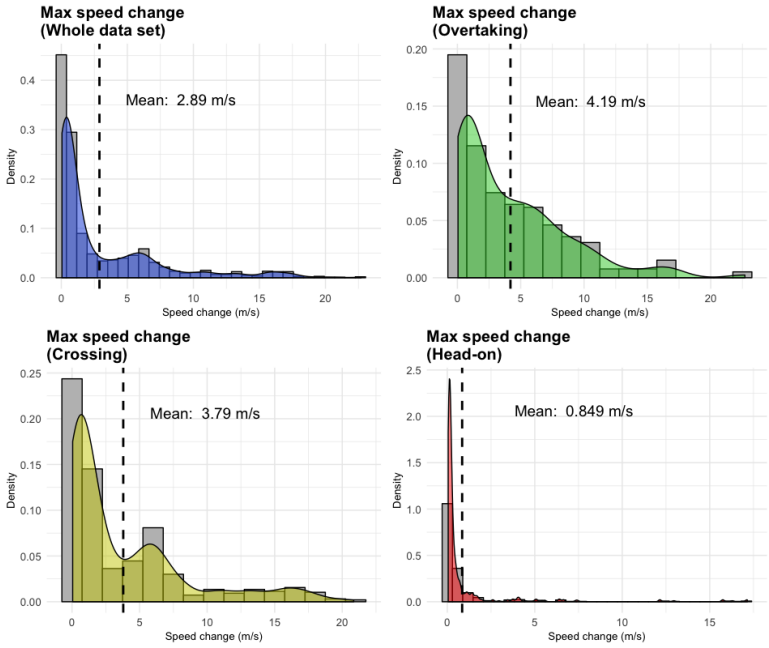

Figure 12: Distribution of maximum speed change for whole data set (panel 1) and the three COLREGs situation types (panels 2-4).

4.3 Statistics by vessel category

Table 9 present the count of situations stratified by vessel types involved in each situation. Tables 10 to 13 present the estimated means of a subset of the core statistics, stratified by the type of vessels involved in each situation. The tables are labelled according to the legend in Table 8, and vessels are grouped on the basis of the main categories of vessel types found in the provided data set. In particular, large natural resource carriers are grouped in T, offshore vessels are grouped in O and cargo vessels are grouped in C. Fishing vessels (F), passenger vessels (P) and cruise ships (U) are kept separate.

Table 8: Legend for ship categories.

T	Chemical Tankers	**C**	General cargo ships
	Gas tankers		Container ships
	Bulk Carriers		Ro-Ro Cargo Ships
	Oil product tankers		Refrigerated Cargo Ships
	Crude oil Tankers	**F**	Fishing vessels
O	Offshore supply ships	**P**	Passenger ships
	Other service offshore vessels	**U**	Cruise ships

These statistics provide somewhat less nuanced information than the distributions, but used in comparison with the plots in Figures 7 to 10 may provide insights into how each vessel category behaves in relation to the average. The estimates corresponding to encounter types with a low count convey less robust information than the ones with a high number of encounters.

Number of situations

Table 9: Count of situations by vessels involved in data set.

	T	**O**	**C**	**F**	**P**	**U**
T	27					
O	14	20				
C	99	53	**222**			
F	10	8	63	**55**		
P	48	34	**283**	51	**361**	
U	4	6	14	1	16	5

Table 9 presents the number of situations by which types of vessels are involved, grouped as laid out in Table 8. It is evident that the data set is rich in situations involving vessels of cargo ship type (C) and passenger ships (P). The highest number of near-collisions situations also occur within and between these two categories. In addition, for fishing vessels (F), the largest number of near-collision situations are with other fishing vessels. These four encounter types

are examples of situations where the data is richer, providing for more robust estimates. The following analysis will focus on these situations.

Yield time

Table 10 presents estimates for yield time stratified by vessel type. There is a significant spread among the different situations. We focus on the four situation types outlined above.

Within the category of cargo ships (C-C), the estimated yield time is somewhat lower than the data set as a whole, at 601 seconds. Within the category of passenger ships (P-P), the negative deviation is larger, at 541 seconds. For encounters between cargo and passenger vessels (C-P), the estimate on the other hand is distinctly higher, at 810 seconds. Within the fishing vessel category (F-F), the estimate is at the lowest of these four encounter types, at 528.

Table 10: Mean yield time categorized by vessels involved.

	T	**O**	**C**	**F**	**P**	**U**
T	1013 s					
O	1101 s	765 s				
C	1257 s	753 s	**601 s**			
F	653 s	302 s	764 s	**527 s**		
P	845 s	647 s	**810 s**	537 s	**541 s**	
U	1939 s	410 s	1144 s	624 s	1044 s	283 s

Yield distance

Table 11 presents estimates for the mean yield distance stratified by vessel type. Focus is directed towards the four relatively data-rich situations.

Within the category of cargo ships (C-C), the estimated yield distance is 4164 meters, a little lower than the data set average. Within the category of passenger ships (P-P) it is at 6476 meters a little higher, and between the two (C-P) at 9569 meters, almost twice the data set mean. Between fishing vessels (F-F) it is lower than the data set average at 4013.

Passing distance

Table 12 presents estimates for the passing distance stratified by vessel type. Focus is again directed towards the four relatively data-rich situations.

Within the category of cargo ships (C-C), the estimated passing distance is at 839, lower than for the data set as a whole. On the contrary, both within the passenger ship group (P-P) and between passenger and general cargo ships (C-P), the estimate for passing distance is at 2184 and 2084 almost twice the estimate for the data set as a whole. The estimate for passing distance between fishing vessels is somewhat lower than for the data set as a whole at 914.

Table 11: Mean yield distance categorized by vessels involved (irrespective of which vessel carries out the evasive manoeuvre).

	T	O	C	F	P	U
T	8000 m					
O	9154 m	5905 m				
C	10 386 m	6027 m	**4168 m**			
F	4910 m	3025 m	6970 m	**4013 m**		
P	21 461 m	11 573 m	**9569 m**	8790 m	**6476 m**	
U	17 253 m	2685 m	14 075 m	2852 m	14 583 m	5370 m

Table 12: Mean "Passing Distance" for vessels categorized by vessels involved.

	T	O	C	F	P	U
T	1170 m					
O	908 m	1169 m				
C	1451 m	1848 m	**839 m**			
F	489 m	482 m	1372 m	**914 m**		
P	3579 m	1872 m	**2086 m**	1955 m	**2184 m**	
U	2512 m	849 m	4355 m	1637 m	2350 m	662 m

Approach speed

Table 13 presents estimates for the approach speed stratified by vessel type. Focus is directed towards the four relatively data-rich situations.

Within the category of cargo ships (C-C), the estimated approach speed is at 20.0 m/s (39.0 knots), at the data set mean. Within the passenger ship category (P-P) it is somewhat higher at 21.6 m/s (42.0 knots), and between the two (C-P) at 20.1 m/s (39.0 knots). Between fishing vessels (F-F) it is distinctly lower at 16.4 m/s (31.9 knots).

Table 13: Mean Approach speed for vessels categorized by vessels involved.

	T	O	C	F	P	U
T	19.2 m/s					
O	19.3 m/s	15.7 m/s				
C	18.2 m/s	18.2 m/s	**20.0 m/s**			
F	15.6 m/s	18.3 m/s	17.3 m/s	**16.4 m/s**		
P	22.2 m/s	22.1 m/s	**20.1 m/s**	23.0 m/s	**21.6 m/s**	
U	19.2 m/s	17.4 m/s	16.6 m/s	15.0 m/s	26.2 m/s	26.8 m/s

4.4 Discussion

Core statistics, manoeuvres and situation types

There is a gap between yield time in particular between situations characterized as being head-on and the data set as a whole. In particular the estimate for overtaking situations is higher. Considering the nature of the situation, this is not strange, as in particular head-on situations involve two vessels headed towards each other, enabling a clearing of the situation fast.

Similarly, there is a gap between yield distance with the same nature, a low estimate for head-on and a higher than average for overtaking. This could also possibly relate to the situation at hand, where a ship approaching another from behind may more easily spot the situation early and act, whereas in the head-on situation, a situation may occur more suddenly, leading to a shorter distance to CPA. The same analysis would be applicable for the estimate of passing distance. Additionally, a vessel approaching another vessel in an overtaking situation will be closer to this vessel than to CPA. In a head-on situation, this will be the opposite. Thus, distances between vessels need not necessarily be as different.

In terms of approach speed, the direction of deviance from the data set average is the opposite, with a higher estimate for head-on and a lower for overtaking. This lends itself to explanation by the fact that vessels in a head-on situation have velocities with opposing directions, while the velocities for overtaking vessels are aligned or close to parallel, but where the situation arises because one vessel has a higher speed than the other.

In combination, these four statistics and their estimates imply that vessels in head-on situations act slower (in terms of being closer to a possible collision point at time of yield) in near-collision situations, leading vessels in such situations to have a shorter passing distance, all the while also exiting situations faster. This may be connected to the fact that the speed of approach between the two vessels is higher, leading to faster resolution of the situation.

For the maximum course change, both overtaking and crossing situations deviate positively (and both have positively skewed distributions), while head-on situations deviate negatively from the data set average. This implies that vessels in overtaking and crossing situations are more prone to make substantial course shifts in order to avoid collision situations, while this is less applied by vessels in head-on situations, or the amount of course change needed is less in head-on situations.

Lastly, maximum speed change exhibit the same split as for maximum course change, where overtaking and crossing have a substantially higher estimate than have head-on situations. In combination, and seen in light of the first four statistics, this may imply that in general, vessels in head-on situations, while reacting more slowly to near-collision situation, also take, or need to take, less action in order to exit these situations.

There is a similar discrepancy in the distribution of manoeuvre types, where the general level of manoeuvres is lower among vessels in head-on situations, and higher among crossing and overtaking. This is in line with the findings discussed above.

Encounters by vessel category

In terms of encounters between vessel categories, encounter types where data is sparse are disregarded, and focus is directed towards the four encounter types emphasized above: Within and between the categories of cargo ships and passenger ships (C-C, P-P, C-P) and within the category of fishing vessels (F-F).

Within the category of cargo ships (C-C), both the time spent on evasive manoeuvres and the passing distance between vessels are shorter than the data set average. The approach speed at average, but yield distance is somewhat smaller than the data set average. This may imply that general cargo vessels are more prone to being in close encounters than the data set average. This may to some extent be explained by some proportion of the cargo ship encounters being in close quarters, e.g. in harbours, although the data is not consistent with the situations just being of this type.

Within the category of passenger ships (P-P), yield time is smaller than the average, but passing distance is larger, at twice the data set average. The approach speed and yield distance is higher than the data set average. This may imply that passenger ship captains put a higher price on exiting near-collision situations, exiting them faster than in the average encounter, and also passing at a greater distance than the average situation encounter, likewise reacting faster. The positive deviance in approach speed may result from a general higher speed for passenger ships, and may also result from the fact that many passenger ships are ferries operating routes where they are prone to encountering meeting ships on the same path, but in the opposite direction.

Between the category of general cargo ships and passenger ships (C-P), the yield time is higher, the passing distance higher and the approach speed lower than the average estimates. Yield distance is on the other hand substantially higher, at twice the data set average. It is easy to envision this as a consequence of passenger ship captains putting a higher price on close encounters with cargo ships, and there being less "natural points of encounter" between passenger and general cargo ships than within the category of passenger ships. On the other hand, a situation where a passenger ship is in close encounter with a general cargo ship may be less routine than an in-category encounter, and thus require more time for evasion.

Within the category of fishing vessels (F-F), yield time is lower, the passing distance slightly lower, the approach speed lower and the yield distance lower than the data set average. The lower estimate for yield time may indicate that fishing vessels generally are smaller and more nimble than the average vessel, thus allowing for faster exits from near-collision situations, and a passing distance which is only somewhat lower than average (the numbers in the present paper is not corrected for vessel length). The lower approach speed may indicate that fishing vessels in general are moving at a slower speed than the average vessel.

4.5 Comments on robustness of statistics

The statistics at hand are produced by first filtering and then calculating estimates for the given statistics. In the analysis of the data it is thus important to make a consideration of whether the modes of filtering in any way introduce

a bias in the data. The present paper will not make a comprehensive analysis of this question, but assume that the methods applied do not introduce skewness to the data. Is is though worth mentioning that the presence of heavy positive tails in the distributions in Figures 7 to 12 may indicate that there is some noise in the sample of near-collision situations that may originate from imperfect filtering.

By applying the methods laid out in Chapter 2 to a larger data set, it would be possible to analyse the predictive capabilities of the methods and estimates.

5 Conclusions

We have in this paper presented algorithms for identification, aggregation and analysis of near-collision situations in the maritime industry and applied these algorithms to a high-resolution temporal AIS data set. We have presented estimates for core statistics regarding evasive manoeuvres, in aggregate and stratified by situation type (as defined by angle of approach) and by the category of the vessels involved in the encounter.

The analyses in this paper are still imperfect, and further research may answer such questions as whether these estimates change if distances are corrected for ship length, and thus creating a more realistic measure of the closeness to collision. Furthermore, a larger dataset may provide dense enough data to stratify statistics further, such as looking at vessel categories by COLREGs situation type, or presenting distributions of core statistics conditional on the manoeuvre types applied in the situation. In terms of the narrowing of the data, better methods for creating sampling time intervals may enable sampling of a larger percentage of vessels within each time interval, while at the same time gathering records which are closer in time for each vessel.

Lastly, if data with sufficiently good speed and course over ground-measurements are provided, numerical differentiation of these may provide estimates for the functional form of the deceleration action applied and likewise the form of course change. This would also enable an extension of the present method for estimation of maximum speed and course change, measuring the gross changes in speed or course, not the net change from yield to exit.

References

[1] Beser, F. and Yildirim, T. “COLREGS Based Path Planning and Bearing Only Obstacle Avoidance for Autonomous Unmanned Surface Vehicles”. In: *Procedia Computer Science*. Recent Advancement in Information and Communication Technology: 131 (2018), pp. 633–640.

[2] Blaich, M. et al. “Extended Grid Based Collision Avoidance Considering COLREGs for Vessels”. In: *IFAC Proceedings Volumes*. 9th IFAC Conference on Manoeuvring and Control of Marine Craft 45.27 (2012), pp. 416–421.

[3] European Maritime Safety Agency. *Annual Overview of Marine Casualties and Incidents 2018*. 2018.

[4] He, Y. et al. “Quantitative analysis of COLREG rules and seamanship for autonomous collision avoidance at open sea”. In: *Ocean Engineering* 140 (2017), pp. 281–291.

[5] Hu, L. et al. “COLREGs-Compliant Path Planning for Autonomous Surface Vehicles: A Multiobjective Optimization Approach**The authors should like to thank Innovate UK, grant reference, TSB 102308, for the funding of this project”. In: *IFAC-PapersOnLine*. 20th IFAC World Congress 50.1 (2017), pp. 13662–13667.

[6] International Maritime Organization. *Guidelines for the onboard operational use of shipborne automatic identification system*. London, UK: IMO, 2002.

[7] Kujala, P. et al. “Analysis of the marine traffic safety in the Gulf of Finland”. In: *Reliability Engineering & System Safety* 94.8 (2009), pp. 1349–1357.

[8] Naeem, W. et al. “A Reactive COLREGs-Compliant Navigation Strategy for Autonomous Maritime Navigation”. In: *IFAC-PapersOnLine*. 10th IFAC Conference on Control Applications in Marine SystemsCAMS 2016 49.23 (2016), pp. 207–213.

[9] Ożoga, B. and Montewka, J. “Towards a decision support system for maritime navigation on heavily trafficked basins”. In: *Ocean Engineering* 159 (2018), pp. 88–97.

[10] Ozturk, U. and Cicek, K. “Individual collision risk assessment in ship navigation: A systematic literature review”. In: *Ocean Engineering* 180 (2019), pp. 130–143.

[11] Sunday, D. *Distance between Lines, Segments and their CPA (2D & 3D)*. Geometry Algorithms Home. 2012. URL: https://geomalgorithms.com/a07-_distance.html (visited on 04/22/2019).

[12] Ventura, M. "COLREGS - International Regulations for Preventing Collisions at Sea". In: *Lloyd's Register Rulefinder 2005* (2005), p. 74.

[13] Yim, J.-B. et al. "Modeling perceived collision risk in vessel encounter situations". In: *Ocean Engineering* 166 (2018), pp. 64–75.

[14] Yoo, S.-L. "Near-miss density map for safe navigation of ships". In: *Ocean Engineering* 163 (2018), pp. 15–21.

[15] Zhang, J. et al. "A distributed anti-collision decision support formulation in multi-ship encounter situations under COLREGs". In: *Ocean Engineering* 105 (2015), pp. 336–348.

A Appendix: Python Code

The full code implementing the project written in `Python` may be found at the project Github repository.

B Appendix: R Code

The full code for exploratory analysis in R may be found at the project Github repository.