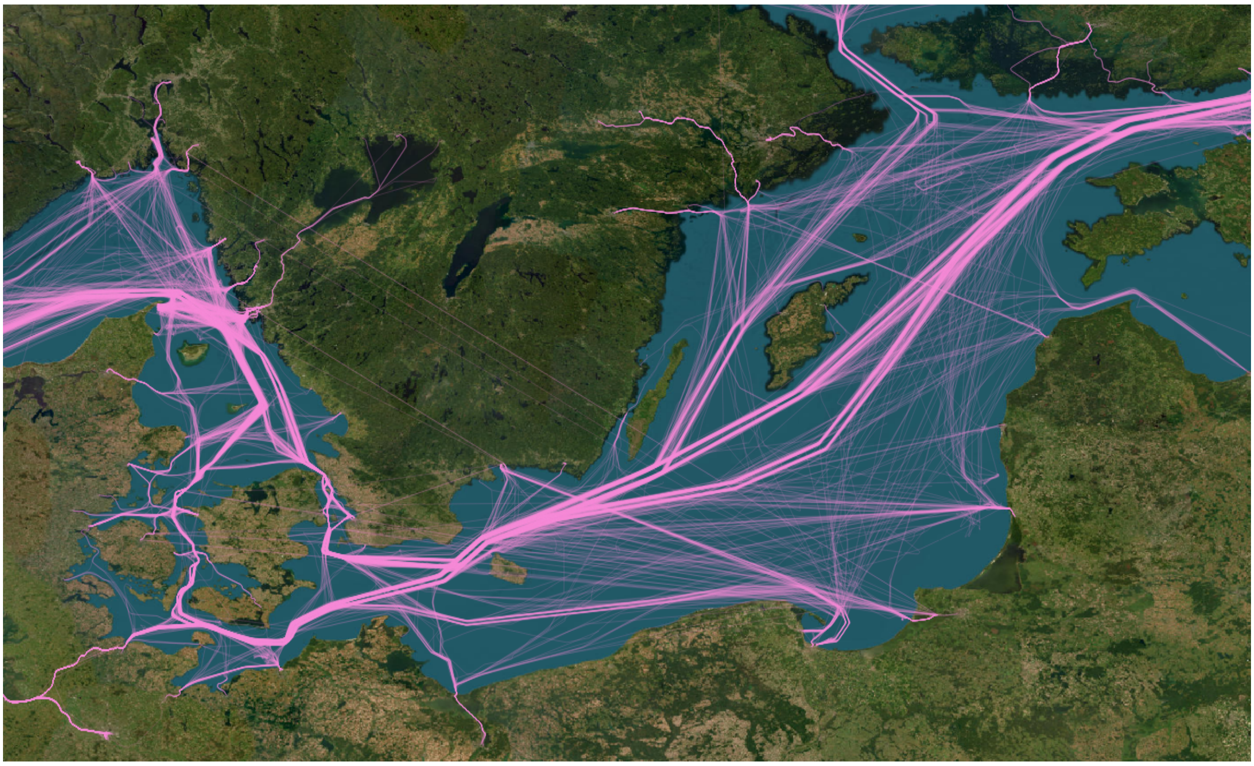
CHALMERS
UNIVERSITY OF TECHNOLOGY



AI/ML Applications to Identify Tank Cleaning Operations & Quantify Slop Discharge

Master's thesis in Data science and AI Programme

Omar Alabdalla
Alamin Alreda

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's Thesis in Data science and AI Programme

AI/ML Applications to Identify Tank Cleaning Operations & Quantify Slop Discharge

Alamin Alreda
Omar Alabdalla

Department of Mechanics and Maritime Sciences
Division of Division Name
Chalmers University of Technology
Gothenburg, Sweden 2026

Supervisor: André Liljegren, Scanjet AB
Examiner: Wengang Mao, Department of Mechanics and Maritime Sciences

Master's Thesis 2026
Department of Mechanics and Maritime Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2026

Abstract

Chemical tankers discharge contaminated washwater from tank-cleaning operations under conditions regulated by MARPOL Annex II and the IBC Code, yet compliance cannot generally be verified externally at the time of discharge. Vessels are not required to report when, where or in what volume tank cleaning occurs and current monitoring relies on aerial surveillance that is constrained by weather, daylight and geographic coverage. This leaves a gap in the environmental oversight of chemical tanker traffic in sensitive regions such as the Baltic Sea. Crucially, no reliable labelled record of confirmed tank-cleaning events exists, which precludes supervised machine learning approaches.

This thesis develops and evaluates a reproducible, label-free pipeline for identifying vessel trajectories whose navigational behaviour is potentially consistent with tank-cleaning activity, using historical AIS data. The work is conducted in collaboration with Scanjet AB and is based on AIS records for the Baltic Sea from January to March 2021. Segmented vessel trajectories are enriched with engineered behavioural features; speed variation, turning-angle variation and drift, and compressed before being rendered as multi-channel image tensors. A self-supervised representation learning framework (BYOL) with a custom convolutional encoder learns compact behavioural embeddings from these tensors without labels, and HDBSCAN clustering is applied to isolate behaviourally uncommon trajectories for expert-informed interpretation. Two trajectory compression methods, DP and TDKC, are systematically compared and the contribution of the engineered behavioural features is assessed through ablation.

Applied to 91,431 learned trajectory embeddings, the framework organised vessel movement into a small number of coherent behavioural clusters while isolating a minority of sparse, behaviourally distinct trajectories consistent with irregular manoeuvring, looping and offshore-deviation patterns. Behavioural feature engineering was found to be decisive: removing the drift channel fragmented the embedding space and increased the proportion of unassignable trajectories from 36 % to 69 %. The compression comparison showed that, in the configurations tested, DP supported more coherent representation learning than TDKC; this difference is attributable primarily to the amount of trajectory structure retained during compression rather than to the compression criterion itself. Detected anomalies were combined with operational washing-system metadata from Scanjet AB to produce scenario-based estimates of potential washwater discharge volumes.

The pipeline is presented as a detection-support tool rather than an autonomous classifier: its outputs are candidate trajectories requiring expert review, not confirmed evidence of discharge. The thesis demonstrates that self-supervised representation learning provides a viable foundation for large-scale behavioural anomaly analysis in label-scarce maritime monitoring contexts and offers a transferable basis for future environmental-monitoring systems.

Keywords: AIS, chemical tankers, tank cleaning, maritime anomaly detection, self-supervised learning, BYOL, trajectory compression, HDBSCAN, representation learning, Baltic Sea.

Preface

This thesis has been carried out at the Division of Marine Technology, Department of Mechanics and Maritime Sciences, Chalmers University of Technology, as part of the Master's Programme in Data Science and AI. Tank cleaning remains one of the least documented sources of environmental impact from shipping. There are no comprehensive records of when and where cleaning operations are performed or in what volumes, and the space between two port calls is from a regulatory standpoint largely unobservable. At the same time, tank cleaning may constitute a considerable burden on sensitive marine environments such as the Baltic Sea. The background to this work was therefore the need for methods capable of identifying such operations at scale, using data that is already collected continuously. The thesis builds upon the pre-study on AIS-based identification of tank-cleaning operations conducted at the Division of Marine Technology, and was carried out in collaboration with Scanjet AB between January and June 2026.

Acknowledgements

Behind a thesis of this kind there are many people whose contributions do not appear in the text itself. We would like to thank our supervisor, André Liljegren at Scanjet AB, for his guidance and for providing the operational knowledge of tank-cleaning systems without which the environmental estimation component of this work would not have been possible. To our examiner, Professor Wengang Mao, whose earlier work laid the foundation for this thesis and whose critical questions shaped its direction at several points. And finally, to the Division of Marine Technology and the Swedish Institute for the Marine Environment, for providing access to the AIS database that formed the empirical basis of the study. Thank you!

Alamin Alreda, Omar Alabdalla, Gothenburg, June 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AIS	Automatic Identification System
BYOL	Bootstrap Your Own Latent
CBT	Compression Binary Tree
CNN	Convolutional Neural Network
COG	Course over ground
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DP	Douglas-Peucker
DWT	Deadweight tonnage
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HELCOM	Helsinki Commission, governing body of the Convention on the Protection of the Marine Environment of the Baltic Sea Area
IBC	International Bulk Chemical
IMO	International Maritime Organization
MARPOL	The International Convention for the Prevention of Pollution from Ships
MMSI	Maritime Mobile Service Identity
PED	Perpendicular Euclidean Distance
SAR	Synthetic Aperture Radar
SED	Synchronous Euclidean Distance
SOG	Speed over ground
SOLAS	International Convention for the Safety of Life at Sea
SVD	Synchronous Velocity Difference
TDKC	Top-Down Kinematic Compression

Contents

List of Figures

List of Tables

1
Introduction

1.1 Background and Motivation

Maritime transportation plays a central role in global trade and industrial supply chains, with liquid bulk chemicals constituting an increasingly important segment of international shipping [14, 15]. Chemical tankers are responsible for transporting a wide range of liquid cargoes that require specialised operational procedures and strict environmental handling standards throughout loading, unloading and tank cleaning operations [7, 9, 15].

Many of these substances require strict operational and environmental handling procedures during loading, unloading and tank cleaning operations [7, 15]. As maritime transportation volumes continue to increase, concerns regarding the environmental impact of operational discharges from chemical tankers have gained growing attention among regulatory authorities and environmental organizations [9, 15].

Tank cleaning operations are necessary to remove cargo residues and prepare cargo tanks for subsequent voyages [15]. These operations may involve the discharge of contaminated washwater under regulated conditions governed by international frameworks such as MARPOL Annex II and the International Bulk Chemical (IBC) Code [7, 15]. Despite these regulations, monitoring and verification of tank-cleaning activities remain challenging in practice due to the limited availability of real-time operational reporting and the absence of transparent electronic monitoring mechanisms [9, 10].

The Baltic Sea and Swedish coastal waters are particularly sensitive to marine pollution due to their semi-enclosed geography, limited water exchange and ecologically vulnerable marine habitats [9]. Reports from Swedish maritime authorities and environmental agencies have documented recurring pollution events suspected to be associated with tank-cleaning discharges in Swedish waters [10]. However, these observations represent only the subset of events detectable under favorable operational conditions such as suitable weather, daylight and surveillance aircraft availability. Consequently, the true frequency and spatial distribution of tank-cleaning-related discharges remain uncertain [9, 10].

Current monitoring approaches are largely reactive and rely on aerial surveillance combined with retrospective trajectory backtracking analysis [10]. Although these methods provide important environmental monitoring capabilities, they remain limited in geographical coverage, continuous monitoring capability and temporal continuity. Furthermore, detailed operational information regarding when and where tank-cleaning activities occur is generally unavailable to authorities in real time.

This creates a significant monitoring gap in the environmental oversight of chemical tanker operations [9, 10].

To improve monitoring capabilities, researchers have increasingly investigated the use of Automatic Identification System (AIS) data for maritime traffic analysis and anomaly detection[4, 10]. AIS is a mandatory communication system for most commercial vessels engaged in international voyages under the International Convention for the Safety of Life at Sea (SOLAS) [6]. AIS continuously transmits navigational and operational information such as vessel position, speed, course, heading and identity[6]. The widespread availability of AIS data has enabled extensive research in maritime surveillance, traffic modeling, anomaly detection and environmental monitoring[4]. In the Baltic Sea region, historical AIS archives maintained by HELCOM and national maritime agencies provide extensive datasets suitable for long-term vessel behavior analysis [5, 10].

Despite these opportunities, AIS-based trajectory analysis introduces significant computational and methodological challenges. AIS datasets are characterized by very large volumes of high-frequency positional information, much of which is redundant during normal vessel navigation. Efficient trajectory preprocessing and compression methods are therefore essential for large-scale maritime analytics. Methods such as the DP algorithm and the TDKC algorithm have been proposed to reduce trajectory complexity while preserving navigational characteristics relevant to vessel behavior analysis, including speed variation, heading changes and maneuvering patterns [4].

One exploratory AIS pre-study examined chemical-tanker movements near two Swedish Coast Guard chemical-pollution observations that were hypothesised, rather than confirmed, to originate from tank cleaning. In these cases, the authors described sharp turns, sailing to open-sea areas outside the Swedish Economic Zone, differing speed conditions and ballast loading; the report separately examined extended AIS-message gaps as a possible screening characteristic. However, the report does not describe validation against onboard operational records, Cargo Record Books or confirmed tank-cleaning events. These observations are therefore treated as exploratory hypotheses for anomaly screening rather than diagnostic signatures of tank cleaning [10], providing a cautious basis for the application of machine-learning methods to larger AIS datasets.

At the same time, recent developments in machine learning have created new opportunities for analyzing large-scale maritime trajectory data. In particular, unsupervised clustering methods and self-supervised representation learning frameworks enable the extraction of latent behavioral representations that may support anomaly detection without requiring manually labeled datasets[4]. This is especially relevant in the present context, where confirmed records of tank-cleaning events are unavailable and reliable ground-truth labels do not exist. Under such conditions, supervised classification approaches become difficult to apply reliably.

This thesis builds upon previous maritime AIS analytics research by investigating how trajectory preprocessing, feature-preserving compression methods and self-supervised machine learning techniques can be combined to identify suspicious chemical tanker behavior in Swedish waters. The work is conducted in collaboration with Scanjet AB. The objective is to develop a reproducible machine learning pipeline capable of processing historical AIS data, extracting behavioral representations from vessel trajectories and identifying anomalous navigational patterns that may support maritime environmental monitoring and future regulatory analysis.

However, AIS-based anomaly detection identifies deviations from common movement patterns rather than the operational causes of those deviations. Whether a vessel follows an established route forms part of the learned traffic context, while unusual course or speed behaviour may reflect collision-avoidance manoeuvres or an intention to delay transit until a pre-arranged arrival time [12]. Wind and current can also influence vessel speed and the difference between course over ground and heading, while route geometry and ordinary navigational manoeuvring further shape the observed behaviour [16]. Consequently, the pipeline outputs are treated as suspicious trajectories requiring expert review rather than confirmed tank-cleaning operations or washwater discharges.

1.2 Problem Statement

Despite the regulatory framework established under MARPOL Annex II and the IBC Code, the enforcement of tank cleaning discharge requirements at sea faces a fundamental structural limitation: compliance cannot generally be verified externally at the time of discharge. The conditions governing lawful discharge, including minimum distance from the nearest land, minimum vessel speed and maximum effluent concentration, are defined in regulation but are not systematically monitored by any real-time vessel-independent system. Chemical tankers are not required to transmit electronic notifications when tank cleaning operations commence or conclude, nor are they required to report the location, duration, or volume of washwater discharged. The only mandatory record is an entry in the vessel's Cargo Record Book, an onboard document accessible to port state control officers only during port inspections, which are periodic and selective [7, 9]. This arrangement means that the space between two port calls, during which tank cleaning most commonly occurs, is from a regulatory standpoint largely unobservable.

This constitutes what may be described as a regulatory gap: the rules governing tank cleaning at sea are codified but the mechanisms for verifying adherence to those rules in real time do not exist. The gap is not merely procedural. Its practical consequence is that neither maritime authorities nor environmental agencies can determine from available data when and where tank cleaning discharges are made, by which vessels, involving which substances and in what volumes. Without this information, quantitative environmental risk assessment is uncertain, targeted enforcement is reactive rather than preventive and the full extent of chemical contamination from tank cleaning in sensitive marine areas such as the Baltic Sea remains unquantified [9].

The primary mechanism currently available for detecting suspected tank cleaning events is aerial surveillance conducted by national coast guard authorities. In Sweden, the Swedish Coast Guard operates systematic aerial monitoring flights over national waters during which pollution observations identified visually or through remote sensing are logged with their spatiotemporal coordinates. Where sufficient information is available, oil spill trajectory models are applied retrospectively to backtrack the likely origin of an observed slick to a vessel located in the vicinity at the relevant time [10]. According to Mao et al. [10], this approach has produced a record of 375 suspected tank cleaning events in Swedish waters over the period 2014 to 2022 and represents a meaningful contribution to environmental monitoring. However, it is subject to substantial limitations that collectively constitute a detection gap.

Aerial surveillance is inherently constrained by flight availability, weather conditions, daylight and geographic coverage [2]. Observations are possible only when a surveillance aircraft is present over a given area at the time a discharge is made or shortly thereafter, a condition satisfied for a small and non-representative fraction of actual events. Discharges conducted at night, in adverse weather, beyond the operational range of surveillance aircraft, or in areas outside routine patrol routes are effectively invisible to this system. Furthermore, trajectory backtracking models provide probabilistic attribution rather than confirmed identification: a vessel identified as a suspect through spatial and temporal proximity to a pollution observation has not been proven to have made a discharge and the evidentiary standard required for enforcement action typically cannot be met on the basis of trajectory matching alone [13]. The result is a detection system that is reactive, spatially sparse and unable to scale to the volume of chemical tanker traffic operating in Swedish and Baltic waters.

MARPOL Annex II governs certain discharges into the sea of cargo residues, tank washings and other mixtures containing noxious liquid substances. Where such discharge is permitted, relevant conditions include proceeding en route at a minimum speed of 7 knots for self-propelled ships or 4 knots for non-self-propelled ships, remaining at least 12 nautical miles from the nearest land, operating in water at least 25 metres deep and discharging below the waterline through the required underwater outlet within its approved rate [7]. These conditions regulate discharge rather than establish sharp turns, low-speed movement, AIS gaps, looping, loitering, drift or other trajectory anomalies as tank-cleaning indicators; moreover, tank cleaning and the subsequent retention, delivery ashore or discharge of washwater are distinct operations [9]. Mao et al. reported sharp manoeuvring, movement into open-sea areas outside the Swedish Economic Zone, ballast conditions and differing speed conditions around two suspected pollution cases, but these were exploratory observations rather than validated causal indicators [10]. Turning angle, speed change and the difference between course over ground and heading are instead analytical AIS features selected by this thesis to represent vessel movement. They are not regulation-derived or validated tank-cleaning indicators, and all resulting candidate behaviours are therefore treated as hypotheses for screening and expert review rather than confirmation of tank cleaning or washwater discharge.

A further constraint, perhaps the most consequential from a methodological standpoint, is the absence of labeled training data. Confirmed timestamped records of individual tank cleaning events specifying the vessel, the location, the time and the substances discharged do not exist in any publicly accessible database. The Swedish Coast Guard's pollution observation records provide spatiotemporal coordinates of suspected discharge events, but these records represent a small and uncertain sample of true events and matching a pollution observation to a specific vessel and cleaning operation introduces inference error at multiple steps. This absence of reliable ground truth prevents reliable conventional supervised training and evaluation, which require labeled examples of both positive tank cleaning and negative normal navigation instances to train a classifier. Standard evaluation metrics such as precision, recall and F1 score, which presuppose the availability of ground truth, cannot be applied in the conventional sense.

This constraint motivates a methodological choice central to the design of this thesis: the use of unsupervised and self-supervised machine learning methods capable of identifying statistically anomalous patterns in AIS trajectory data without requiring labeled training examples. Density-based clustering algorithms, specifically DBSCAN [1] and HDBSCAN [11], are capable of identifying compact regions of high behavioral similarity and isolating trajectories that do not conform to any dominant pattern. Such trajectories may, subject to expert review, correspond to unusual navigational behavior. Self-supervised representation learning, specifically the BYOL framework, offers the complementary capability of learning compact informative representations of trajectory behavior from unlabeled data, potentially capturing behavioral structures within the multi-channel trajectory representations that are not explicitly defined by the hand-engineered features [3]. Together, these methods provide a principled and scalable approach to the detection problem as it is actually constituted: a large unlabeled dataset, a behaviorally complex target signal and no available ground truth against which outputs can be calibrated.

In summary, the problem addressed by this thesis is defined by four interconnected gaps: a regulatory gap that leaves tank cleaning discharges unmonitored in real time, a detection gap that renders aerial surveillance insufficient in scale and coverage, a signal ambiguity challenge that prevents any single AIS feature from serving as reliable evidence of tank cleaning and a label scarcity problem that precludes supervised learning approaches. Addressing this problem requires a pipeline that is simultaneously data-driven and epistemically cautious, one capable of surfacing suspicious trajectory patterns at scale while explicitly acknowledging the inferential distance between an anomalous navigational signature and a confirmed discharge event. The design, implementation and evaluation of such a pipeline is the subject of this thesis.

1.3 Research Aim and Objectives

1.3.1 Research Aim

The overarching aim of this thesis is to develop and evaluate a data-driven pipeline for the automatic identification of suspicious vessel trajectories potentially associated with tank cleaning activity in the Baltic Sea, using large-scale AIS trajectory data in combination with unsupervised and self-supervised machine learning methods. The aim is deliberately bounded in two respects. First, the pipeline is designed as a detection support tool: its outputs are candidate trajectories for expert review, not autonomous determinations of regulatory non-compliance. Second, the work is empirical and evaluative in character, the thesis does not propose new machine learning algorithms, but rather applies, adapts and systematically evaluates existing methods within a novel domain and a previously unaddressed operational context.

1.3.2 Research Objectives

The aim is pursued through sex research objectives, which correspond to the sequential stages of the detection pipeline and to the chapter structure of the thesis.
O1. Construct a clean, structured AIS dataset of chemical tanker trajectories, including validated preprocessing, quality filtering and dataset documentation with descriptive statistics.
This objective provides the empirical foundation for all subsequent analysis. It is addressed in Chapter 3 and encompasses AIS data extraction, vessel-type filtering, metadata integration, removal of invalid navigation records and the segmentation of continuous AIS sequences into discrete trajectory units suitable for downstream processing.
O2. Engineer and evaluate a set of behavioural features derived from segmented AIS trajectories, including turning angle, speed change and drift, to determine their relevance for distinguishing anomalous from typical navigational behaviour. This objective treats turning angle, speed change and drift as thesis-selected analytical features. Their selection was partly motivated by the exploratory trajectory and speed observations reported by Mao et al. [10], and they are used to characterise vessel movement for hypothesis-based anomaly screening rather than to confirm tank-cleaning events. The objective is addressed in Chapter 4 and contributes directly to Research Question RQ1.
O3. Implement and compare two trajectory compression methods, DP and the TDKC algorithm, with respect to compression ratio, kinematic fidelity and their impact on downstream trajectory representation and clustering behaviour.
Trajectory compression is a prerequisite for computationally tractable analysis of large AIS datasets and for the construction of fixed-size image-based representations required by the self-supervised learning pipeline. This objective is addressed in Chapters 3 and 4 and is the subject of Research Question RQ2.
O4. Develop a self-supervised representation learning pipeline using the BYOL framework to learn compact trajectory embeddings from image-based representations of compressed AIS trajectories and evaluate the quality of the learned representations with respect to their ability to capture latent motion patterns potentially associated with anomalous vessel behaviour. This objective constitutes the primary methodological contribution of the thesis. By operating without labels, BYOL addresses the data constraint established in Section 1.2. It is addressed in Chapter 4 and evaluated in Chapter 5 and forms the basis of Research Question RQ3.
O5. Apply HDBSCAN clustering to BYOL-learned trajectory embeddings to identify low-density and behaviourally distinct trajectory patterns and interpret flagged trajectories through systematic visual inspection in relation to thesis-defined screening hypotheses.
This objective produces the operational outputs of the anomaly-screening pipeline, namely the set of trajectories identified as behaviourally anomalous, and provides a structured, replicable protocol for their expert-informed interpretation in the absence of ground truth. It is addressed in Chapters 4 and 5 and contributes to

Research Question RQ3.
O6. Estimate the volume of washwater discharge potentially associated with detected anomalous trajectories.
This objective extends the anomaly detection framework toward an environmental impact assessment by translating detected behavioural patterns into approximate discharge estimates under a set of operational assumptions. Since direct measurements of discharge are unavailable, the analysis is formulated as a scenario-based estimation procedure rather than a deterministic calculation. The objective is addressed in Chapter 5 or Chapter 6 and contributes to the secondary Research Question RQ4.

1.4 Research Questions

The research objectives defined in Section 1.3 are operationalised through four research questions. Each question is formulated to be answerable within the scope of the available data and methods and each corresponds to a distinct stage of the detection pipeline or a distinct dimension of its evaluation. The questions are revisited explicitly in Chapter 7, where the findings of the thesis are assessed against them.
RQ1. How does the inclusion of drift alongside turning angle and speed change influence anomaly-clustering results for chemical-tanker AIS trajectories?
This question motivates the feature engineering stage of the pipeline (Objective O2, Chapter 4) and grounds the anomaly detection approach in analytical movement features selected by this thesis, partly motivated by the exploratory observations of Mao et al. [10]. Its importance lies in assessing how the inclusion of drift influences the HDBSCAN clustering results when turning angle and speed change remain fixed. An answer to RQ1 also provides interpretable evidence that can be communicated to domain experts and regulatory authorities, who require transparency regarding the basis on which trajectories are flagged as anomalous.
RQ2. Does trajectory compression using the TDKC algorithm preserve behavioural and kinematic trajectory characteristics more effectively than DP compression and how does this influence downstream trajectory representation and clustering behaviour?
This question addresses Objective O3 and is answered in Chapters 4 and 5. It is motivated by a practical tension inherent in the pipeline design: DP compression preserves geometric shape but discards kinematic information such as speed variation and rate of turn, while TDKC is designed to retain such information at the cost of a less aggressive compression ratio. Whether this improved preservation of kinematic information leads to more coherent trajectory representations and more meaningful clustering behaviour is an empirical question that has not previously been examined in the context of chemical tanker behavioural analysis.
RQ3. Can BYOL-based self-supervised learning produce trajectory embeddings that enable HDBSCAN to identify behaviourally distinct trajectory patterns potentially consistent with tank cleaning activity, without reliance on labelled training examples?
This question is the methodological core of the thesis and addresses Objectives O4 and O5. It is answered through the embedding space analysis, clustering results

and visual inspection protocol reported in Chapter 5. The question is formulated cautiously: it asks whether identified trajectory patterns are potentially consistent with tank cleaning activity, not whether they constitute proof of it, reflecting the constraint established in Section 1.2. A positive finding would suggest that self-supervised representation learning is a viable approach to behavioural anomaly detection in label-scarce maritime monitoring contexts, with implications beyond the specific application addressed here.

RQ4 To what extent can anomalous trajectories identified by the pipeline, combined with operational washing performance data from Scanjet AB, support estimates of potential washwater discharge volumes ?

This question addresses the conditional Objective O6 and corresponds to the secondary environmental estimation strand of the thesis. The findings are reported as scenario analyses rather than empirical measurements and no causal claims regarding environmental outcomes are made beyond what the available data directly supports.

1.5 Scientific Contributions

This thesis makes three principal contributions to the research area of maritime trajectory analysis and environmental monitoring. These contributions are methodological, empirical, and operational in character. They build directly on the exploratory findings of Mao et al. (2023), which established the feasibility of AIS-based behavioural analysis for chemical tank cleaning detection but did not extend to machine learning methods or systematic pipeline evaluation. The contributions of the present work are distinguished from that predecessor not by the novelty of the individual algorithms employed but by their integration into a coherent and reproducible pipeline, their application to a previously underexplored operational problem and the empirical evidence produced through their systematic evaluation.

C1. An end-to-end detection pipeline operating without ground-truth labels.

The primary contribution of this thesis is the design, implementation and evaluation of a complete pipeline for suspicious trajectory detection in chemical tankers that operates without labelled training data. The pipeline integrates AIS data preprocessing, behavioural feature engineering, trajectory segmentation, two compression strategies, trajectory-to-image construction, BYOL self-supervised representation learning and HDBSCAN clustering into a single reproducible workflow. The contribution lies not in the invention of new algorithms but in the principled integration of existing methods to address a problem for which labelled datasets are unavailable and fully automated analytical approaches remain limited. The pipeline is designed to be transferable and could, with appropriate data preparation, be adapted to other geographic regions.

C2. A systematic comparison of DP compression and TDKC trajectory compression for maritime trajectory analysis.

Trajectory compression is a necessary preprocessing step in large-scale AIS analysis, but the implications of compression choices for downstream clustering behaviour and representation quality are not always explicitly examined.

This thesis provides a structured comparison of two compression strategies, DP, which preserves geometric shape and TDKC, which prioritises kinematic information, across multiple compression parameter settings. The comparison evaluates compression ratio, point retention, behavioural fidelity and the effect of compression on the clustering outputs produced by HDBSCAN. This contributes to the broader field of trajectory data mining by examining how preprocessing decisions influence downstream analytical behaviour in maritime trajectory analysis.

C3. An adaptation of BYOL self-supervised learning to maritime trajectory-image data.

BYOL is an established self-supervised learning framework originally developed in the computer vision literature [3]. Its application to maritime trajectory data represented as fixed-size images encoding spatial paths and kinematic attributes has received limited investigation in the context of vessel behavioural analysis. This contribution involves both the technical adaptation of the framework and the design of domain-appropriate augmentation strategies for trajectory-image data, which differ from the natural image augmentations used in conventional computer vision tasks. The thesis thereby extends the applicability of self-supervised representation learning to structured geospatial trajectory data characterised by limited label availability, a problem setting common in maritime and environmental monitoring applications.

1.6 Scope and Delimitations

The scope of this thesis is defined by the data available, the methods selected and the operational context established in collaboration with the supervising institution and industry partner. Explicit delimitation of scope is necessary in a study of this kind, where the problem domain is broad, the data is large-scale and imperfectly labelled, and the inferential distance between observable signals and the phenomenon of interest is substantial. The boundaries described in this section are not presented as deficiencies of the work, but as deliberate and reasoned choices intended to ensure that the thesis remains methodologically tractable and scientifically rigorous.

1.6.1 Geographic and Data Limitations

The analysis conducted in this thesis is restricted to vessel trajectories within the Baltic Sea and is based exclusively on AIS data collected during January, February, and March of 2021. This temporal and geographic scope was determined primarily by data availability, computional and time constraints rather than by statistical sampling considerations. Consequently, the findings of this thesis, including identified behavioural patterns and anomaly characteristics, should not be generalised beyond the analysed period or geographic region without further investigation.

The AIS dataset used in this thesis is sourced from archives maintained by the Swedish Institute for the Marine Environment and provided through the research collaboration established by Mao et al. (2023). The dataset therefore reflects the coverage characteristics and any systematic gaps present in those archives, and no independent verification of dataset completeness is performed. AIS data is known to contain recording flaws, including signal gaps attributable to equipment malfunction and receiver coverage limitations, which cannot always be distinguished from intentional signal suppression [8]. These flaws are acknowledged as a source of uncertainty throughout the analysis.

1.6.2 Why Supervised Learning is Outside Scope

Supervised classification methods, which learn to distinguish between positive and negative instances from labelled training examples, are not applicable in this study because no reliable labelled dataset of confirmed tank cleaning events exists. The Swedish Coast Guard's pollution observation records provide spatiotemporal coordinates of suspected discharge events, but the process of matching these observations to specific vessels and cleaning operations introduces inference error at multiple stages, and the resulting pseudo-labels would not meet the quality standard required for supervised learning.

1.6.3 Why Real-Time Deployment is Outside Scope

The pipeline developed in this thesis is designed for offline analysis of historical AIS data. Extending the system to support real-time operational monitoring would require substantial additional engineering work, including stream-processing infrastructure, latency-optimised model inference and integration with live AIS data feeds. These engineering requirements are beyond the scope of a research thesis and are therefore designated as future work.

The present thesis instead provides the analytical foundation of a systematically evaluated detection pipeline on which future operational systems may subsequently be built.

1.6.4 Why SAR and Satellite Imagery Integration is Outside Scope

Synthetic aperture radar (SAR) and optical satellite imagery have been proposed as complementary data sources for maritime pollution detection, capable of identifying surface slicks that may corroborate AIS-based suspicious trajectory findings. The integration of such data with AIS trajectory analysis is technically feasible in principle but would require access to satellite imagery archives, image-processing expertise and a dedicated data-fusion methodology that would itself constitute a substantial research contribution.

Such integration is therefore outside the scope of the present thesis and is identified as a valuable direction for future work, particularly in the context of validating pipeline outputs against independent evidence of surface contamination.

1.6.5 Epistemic Delimitation

The most important boundary of this thesis is epistemic rather than technical. The pipeline produces a set of trajectories identified as anomalous with respect to the

behavioural norms of the chemical tanker population represented in the dataset. These trajectories are described throughout the thesis as suspicious, meaning that their navigational characteristics are potentially consistent with behavioural patterns associated with tank cleaning activity as documented in prior literature.
They are not described as confirmed tank cleaning events and no claim of regulatory non-compliance is made on the basis of trajectory analysis alone. The inferential step from an anomalous navigational signature to an actual discharge event requires domain expert review, corroborating evidence and a standard of proof that trajectory data cannot independently satisfy. This distinction is maintained consistently throughout the thesis.

2
Theory

2.1 Chemical Tanker Operations

Chemical tankers are specialized vessels designed for the transportation of liquid bulk chemicals, petroleum products, vegetable oils, biofuels and other hazardous or sensitive cargoes, as mentioned in chapter 1.1. Unlike conventional oil tankers, chemical tankers are equipped with multiple segregated cargo tanks and advanced cargo handling systems that enable simultaneous transportation of different chemical products under controlled operational conditions. Due to the wide variety of transported substances and their differing chemical properties, chemical tanker operations involve strict safety cleaning and environmental protection procedures [9].
Cargo operations on chemical tankers typically include loading, transportation, unloading, tank cleaning, gas freeing and cargo preparation for subsequent voyages. After cargo discharge, residual substances often remain attached to tank surfaces, pipelines and pumping systems. To avoid contamination between cargoes and maintain operational safety standards, tanks must be cleaned before loading new cargoes [9]. The required cleaning procedures depend on factors such as cargo type, chemical compatibility, tank coating materials and subsequent cargo requirements.
Tank cleaning operations are therefore an essential part of chemical tanker logistics and operational planning. Cleaning processes may involve:

- water washing,
- chemical cleaning agents,
- heating systems,
- high-pressure washing machines,
- ventilation and drying procedures,
- slop and residue handling systems.

Modern chemical tankers are commonly equipped with fixed or portable tank cleaning systems that spray cleaning fluids throughout cargo tanks to remove remaining cargo residues. The resulting wash water and slop mixtures may contain oils, chemicals, detergents and hazardous substances depending on the previous cargo composition. Consequently, tank cleaning operations are closely linked to both operational safety and environmental protection concerns [9].
The discharge of tank-cleaning residues is regulated internationally through MARPOL Annex II, which governs the handling and discharge of noxious liquid substances carried in bulk. MARPOL classifies chemicals according to their environmental hazard levels and specifies operational requirements for cargo residue discharge, tank prewashing, stripping efficiency and discharge limitations. Depending

on cargo category and operational conditions, some tank-cleaning discharges may legally occur at sea under controlled circumstances, while others require mandatory discharge to port reception facilities [7, 9].
Despite existing regulations, monitoring tank-cleaning activities remains challenging in practice. Current reporting systems generally do not provide real-time public information regarding when and where tank-cleaning discharges occur. Operational details are typically documented onboard vessels but are not continuously transmitted to maritime authorities. This lack of transparency complicates environmental monitoring and makes it difficult to assess the frequency, location, and environmental impact of tank-cleaning discharges [9, 10].
Some studies have indicated that even legally permitted discharges may pose environmental risks depending on the discharged substances, local environmental conditions, and cumulative pollution effects.
Understanding the operational characteristics of chemical tankers is therefore essential for interpreting AIS-based vessel behavior and developing anomaly detection methods related to potential tank-cleaning activities [10]. In this thesis, knowledge of chemical tanker operations provides domain context for interpreting movement patterns selected for anomaly screening, without treating them as evidence of cleaning or discharge operations.

2.2 AIS Data

AIS is a maritime communication and tracking system developed to improve navigational safety and situational awareness between ships and shore-based monitoring stations [6]. Following amendments to SOLAS, AIS transponders became mandatory for most commercial vessels, including cargo ships above 300 gross tonnage and all passenger ships engaged in international voyages [6]. AIS enables continuous transmission of vessel-related information, providing an important foundation for modern maritime traffic monitoring and analysis [10].
AIS messages contain both dynamic and static vessel information. Dynamic information includes navigational parameters such as vessel position, speed over ground (SOG), course over ground (COG), heading, and timestamp, while static information includes vessel identity, ship type, dimensions, and draught [10]. The transmission frequency of AIS messages varies depending on vessel movement conditions, ranging from a few seconds during active navigation to several minutes when vessels are anchored or moving slowly [10]. This high-frequency data generation enables detailed reconstruction of vessel trajectories and operational behavior over time.
AIS data has been widely used in studies related to maritime traffic modeling, route prediction, collision risk assessment, anomaly detection, port activity monitoring, environmental impact assessment, and vessel behavior analysis [4, 13]. Advances in big data analytics and machine learning have further expanded the ability to extract operational insights from large-scale maritime datasets [3].
One of the major strengths of AIS data is its ability to capture vessel movement patterns continuously over large geographical areas [5]. By analyzing vessel trajectories, researchers can identify recurring navigational patterns, operational routines and potentially anomalous behaviors [4]. In maritime surveillance applications, AIS-

based analytics have been applied to detect suspicious vessel activities, abnormal route deviations and unusual navigation behavior [10]. In environmental monitoring studies, AIS trajectories have also been used to support analyses of vessel emissions, fuel consumption and pollution-related events [13].
AIS trajectory analysis commonly involves the extraction of movement-related features from vessel navigation data. Typical trajectory features include vessel speed profiles, turning behavior, stop durations, route deviations, spatial density distributions and temporal movement characteristics [4]. Such features describe vessel movement patterns but do not, by themselves, establish operational states or vessel intent. In the context of chemical tanker operations, trajectory analysis may reveal behaviors potentially associated with tank-cleaning activities, including prolonged AIS signal gaps, unusual maneuvering patterns, abrupt course changes, or movement patterns occurring in offshore areas where discharges may take place [10].
Despite its extensive usefulness, AIS data presents several technical and analytical challenges. Maritime AIS datasets are characterized by extremely large data volumes due to high-frequency message transmissions from thousands of vessels operating continuously. As a result, AIS data processing often requires computational resources for storage, preprocessing, filtering and analysis [4]. Additionally, AIS datasets may contain missing messages, duplicated records, transmission noise, inconsistent timestamps, and positional inaccuracies caused by communication limitations or equipment-related issues [8, 10].

2.3 Trajectory Compression Methods for AIS Data

To address the problem of extremely large AIS datasets that are computationally expensive to process and analyze, trajectory compression methods at the point level were applied in this thesis [4]. Trajectory compression methods aim to simplify vessel trajectories by reducing the number of trajectory points while maintaining the overall shape and important behavioral features of vessel movement. In maritime applications, compression techniques are particularly valuable for improving computational efficiency in machine learning, trajectory mining, and anomaly detection tasks [4]. Compression methods can generally be categorized into lossless and lossy compression approaches. In AIS trajectory analysis, lossy compression methods are more commonly applied due to their ability to achieve significantly higher compression rates while maintaining acceptable trajectory fidelity [4]. In this thesis, two trajectory compression approaches were applied during AIS preprocessing: the Douglas-Peucker (DP) algorithm and the Top-Down Kinematic Compression (TDKC) algorithm. These methods were selected due to their suitability for maritime trajectory simplification and their ability to preserve operationally relevant vessel behavior [4].

2.3.1 DP Compression Algorithm

The DP algorithm is one of the most widely used trajectory simplification methods and has become a foundational approach in maritime AIS trajectory compression research [4]. Originally developed for line simplification in cartography, the method has been extensively applied in vessel trajectory analysis due to its simplicity, computational efficiency, and ability to preserve the overall geometric shape of trajectories.

The DP algorithm operates using a recursive top-down approach. Initially, the first and last points of a trajectory segment are connected to form a baseline. The spatial distance between trajectory points can be calculated using the Haversine formula shown in Equation (2.1), which computes the great-circle distance between two geographic coordinates:

$$d_{i,j} = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_j - \phi_i}{2}\right) + \cos(\phi_i)\cos(\phi_j)\sin^2\left(\frac{\lambda_j - \lambda_i}{2}\right)}\right) \tag{2.1}$$

where r denotes the Earth's radius, ϕ_i and ϕ_j represent the latitudes, and λ_i and λ_j represent the longitudes of two AIS points.

To determine whether an intermediate trajectory point should be preserved, the DP algorithm computes the perpendicular distance between the point and the baseline segment. This perpendicular Euclidean distance (PED) is calculated using Equation (2.2):

$$d_p = \frac{|(y_2 - y_1)x_0 - (x_2 - x_1)y_0 + x_2y_1 - y_2x_1|}{\sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}} \tag{2.2}$$

where (x_0, y_0) denotes the intermediate trajectory point, while (x_1, y_1) and (x_2, y_2) define the segment endpoints.

The point with the maximum deviation from the baseline is identified, and if this deviation exceeds a predefined threshold, the point is preserved as a key trajectory point. The trajectory is subsequently divided into sub-segments and the same process is recursively repeated until no remaining points exceed the distance threshold [1, 4].

The recursive trajectory simplification procedure of the DP algorithm is illustrated in Figure 2.1. As shown in the figure, trajectory points with small perpendicular deviations are removed while the major geometric characteristics of the vessel trajectory are preserved.

The primary advantage of the DP algorithm is its ability to significantly reduce the number of trajectory points while maintaining the overall geometric structure of vessel movement [4]. This makes the method computationally efficient for large-scale AIS datasets and suitable for visualization, traffic analysis, and route simplification tasks. Additionally, the algorithm is relatively simple to implement and has been widely adopted as a compression method in maritime trajectory research.

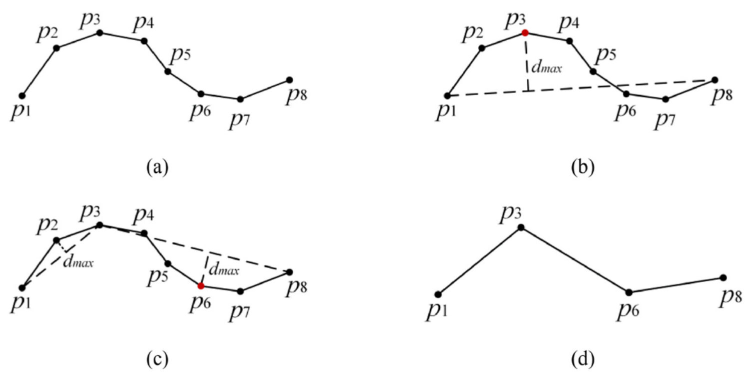


Figure 2.1: Illustration of the Douglas-Peucker trajectory compression algorithm.

2.3.2 TDKC algorithm

To address the limitations of purely geometry-based compression methods, more advanced trajectory simplification approaches have been developed that incorporate vessel kinematic behavior during compression [4]. One such method is the TDKC algorithm, proposed for adaptive trajectory compression and feature preservation in maritime traffic analysis [4].

Unlike the DP algorithm, TDKC incorporates multiple AIS attributes simultaneously, including timestamp, position, speed over ground, and course over ground [4]. The objective of TDKC is not only to reduce data size but also to preserve operationally significant vessel behavior during compression. This is particularly important in maritime anomaly detection applications, where navigational characteristics such as speed variations, maneuvering behavior, and abrupt course changes may contain critical behavioral information [4, 10].

TDKC is based on a top-down recursive compression framework similar to DP, but it introduces additional kinematic-aware measurement techniques [4]. The algorithm utilizes Synchronous Euclidean Distance (SED) to incorporate temporal-spatial trajectory deviations and Synchronous Velocity Difference (SVD) to preserve changes in vessel velocity behavior. By combining spatial and kinematic information, TDKC improves the preservation of important trajectory features that may otherwise be removed by purely geometric simplification methods [4].

The SED measures the spatial deviation between the original trajectory point and its synchronized point on the baseline, as defined in Equation (2.3):

$$SED_{i+1} = \sqrt{(x_{i+1} - x''_{i+1})^2 + (y_{i+1} - y''_{i+1})^2} \tag{2.3}$$

where (x_{i+1}, y_{i+1}) denotes the original trajectory point and (x''_{i+1}, y''_{i+1}) denotes the synchronized point on the baseline.

Similarly, the SVD quantifies the deviation between the original vessel velocity vector and the synchronized velocity vector, as shown in Equation (2.4):

$$SVD_{i+1} = \sqrt{(v_{x_{i+1}} - v'_{x_{i+1}})^2 + (v_{y_{i+1}} - v'_{y_{i+1}})^2} \tag{2.4}$$

where $(v_{x_{i+1}}, v_{y_{i+1}})$ represents the original vessel velocity vector and $(v'_{x_{i+1}}, v'_{y_{i+1}})$ represents the synchronized velocity vector.

The geometric interpretation of PED, SED, and SVD is illustrated in Figure 2.2.

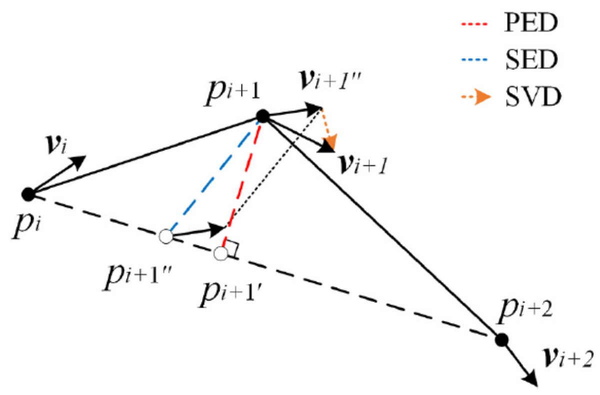


Figure 2.2: Illustration of PED, SED, and SVD measurements used in the TDKC algorithm.

Another important contribution of TDKC is its adaptive threshold determination mechanism [4]. Traditional compression algorithms often require manually predefined thresholds that may not generalize effectively across different vessel trajectories or operational environments. TDKC addresses this limitation by using a Compression Binary Tree (CBT) structure and adaptive threshold calculations based on trajectory characteristics [4]. This allows the algorithm to dynamically balance compression efficiency and feature preservation.

2.4 Machine Learning for Maritime Anomaly Detection

Machine learning refers to computational methods that enable systems to learn patterns and relationships from data without being explicitly programmed for every decision. In maritime applications, ML models can analyze historical AIS trajectories and operational behavior to identify recurring navigation patterns, classify vessel activities, and detect deviations from expected movement behavior [4, 10]. Such methods are particularly suitable for anomaly detection problems, where abnormal events may occur infrequently and exhibit complex behavioral characteristics [1, 11]. In its detailed examination of two suspected pollution cases, Mao et al. reported irregular manoeuvring and movement toward open-sea areas, but these observations were not validated as tank-cleaning indicators [10]. Although such movements are observable in AIS data, their relationship to tank cleaning cannot be established from individual features or simple thresholds alone. They therefore provide exploratory motivation for machine learning approaches that identify complex and statistically unusual trajectory patterns for expert review.

2.4.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a specialized type of neural network designed for image-based pattern recognition tasks. Unlike traditional fully connected neural networks, CNNs use convolution operations to capture spatial relationships and local patterns within images. The convolution process applies small filters, known as kernels, across an image to detect characteristic features such as shapes, edges, textures or movement patterns.
Multiple convolution layers can be stacked to progressively learn increasingly abstract representations of the input data. Early layers may detect simple geometric structures, while deeper layers learn more complex behavioral patterns and semantic representations. After convolution operations, the extracted features are transformed into feature vectors that can be used for classification or clustering tasks.
In this thesis, CNN-based methods are particularly relevant because AIS vessel trajectories can be transformed into trajectory images representing vessel movement behavior over time. This allows maritime trajectory analysis to be approached as an image recognition problem, where unusual sailing patterns can potentially be identified similarly to irregular visual patterns in computer vision applications [3].
The general CNN feature extraction process used for image-based representation learning is illustrated in Figure 2.3. As shown in the figure, the input trajectory image is processed through multiple convolution and pooling stages before being transformed into a compact latent feature vector representation [3].

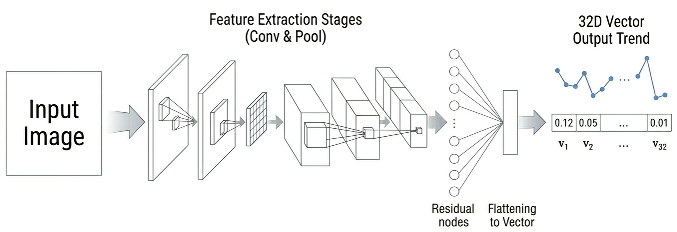


Figure 2.3: General architecture of a CNN used for feature extraction and latent vector generation.

2.4.2 Feature Embeddings and Latent Space Representation

An important concept in deep learning is the generation of feature embeddings, also referred to as latent vectors [3]. A feature embedding is a lower-dimensional numerical representation of input data learned by a neural network during training. Similar input samples tend to produce similar feature vectors, causing related patterns to cluster together within the latent feature space [3].
For trajectory analysis, embedding vectors provide a compact representation of vessel movement behavior while preserving important navigational characteristics. By

mapping trajectories into latent space, machine learning models can group similar vessel behaviors and separate anomalous patterns from normal navigation [3, 11].
The clustering behavior of trajectory embeddings within latent space is illustrated in Figure 2.4. As shown in the figure, trajectories with similar navigational characteristics are grouped together after dimensionality reduction, while anomalous movement patterns appear as isolated or distinct clusters within the feature space [11]. This approach is particularly useful for large-scale AIS datasets where manual trajectory inspection is impractical.

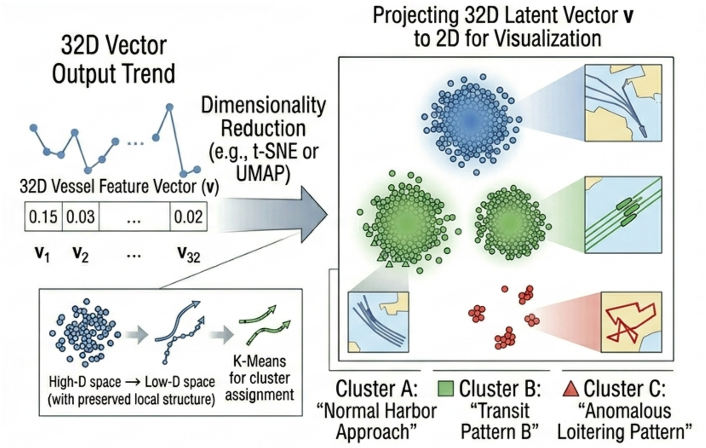


Figure 2.4: Illustration of latent space representation and clustering of vessel trajectory embeddings after dimensionality reduction. Similar vessel behaviors form dense clusters, while anomalous trajectories appear as isolated patterns in latent space.

2.4.3 Self-Supervised Learning

One of the major challenges in maritime anomaly detection is the limited availability of labeled AIS datasets [10]. Manual labeling of vessel trajectories is time-consuming and requires domain expertise, particularly when analyzing rare or uncertain operational events such as potential tank-cleaning activities [10]. To address this challenge, self-supervised learning approaches have gained increasing attention in maritime AI research [3].
Self-supervised learning enables models to learn useful feature representations from unlabeled data by generating supervisory signals directly from the data itself [3].
BYOL is a self-supervised learning framework in which multiple augmented versions of the same trajectory image are generated and trained to produce similar latent representations [3].
The objective of BYOL is to learn latent representations in which trajectories with similar structural and behavioural characteristics are mapped to nearby regions of the embedding space, while dissimilar trajectory patterns are represented farther apart [3].
The BYOL self-supervised learning framework applied to AIS trajectory representations is illustrated in Figure 2.5. As shown in the figure, multiple augmented versions of the same trajectory image are processed through online and target net-

works in order to learn invariant latent feature representations without requiring manual labeling [3].
This enables the neural network to learn meaningful trajectory representations directly from unlabeled AIS data. Such approaches are particularly valuable for maritime anomaly detection problems involving large-scale unlabeled AIS datasets [3].

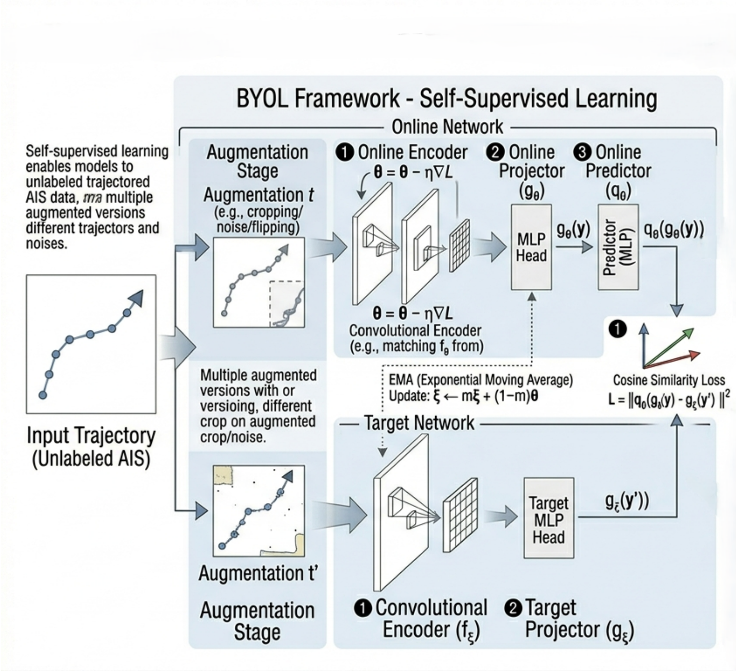


Figure 2.5: Illustration of the BYOL self-supervised learning framework.

2.5 HDBSCAN

In AIS-based maritime analytics, clustering is applied after feature extraction or embedding generation. When vessel trajectories are transformed into latent feature vectors using machine learning models such as CNNs, trajectories exhibiting similar movement behavior tend to be positioned close together within latent space [3]. Clustering algorithms can then be used to identify groups of normal navigation behavior and isolate outlier trajectories that may represent anomalous operations [1, 11].
In this thesis, the clustering method Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) was applied for trajectory behavior analysis and anomaly detection [11]. HDBSCAN is an extension of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm and is specifically designed to identify clusters of varying densities while handling noise and outliers effectively [1, 11]. Unlike traditional clustering methods such as k-means, HDBSCAN does not require the number of clusters to be predefined, making it particularly suitable for exploratory anomaly detection problems where the underlying structure of the data is unknown [11].
HDBSCAN operates by identifying dense regions within feature space and constructing a hierarchical clustering structure based on density connectivity between data

points [11]. The algorithm progressively evaluates clusters across multiple density levels and extracts the most stable cluster structures from the hierarchy. Points that do not belong to sufficiently dense regions are classified as noise or outliers [11]. This property is valuable in maritime anomaly detection, where unusual vessel behaviors may naturally appear as isolated or low-density observations within the dataset.
In the context of this thesis, HDBSCAN was applied to clustering embeddings generated from processed AIS trajectories. The objective was to identify dominant groups of regular vessel behavior while isolating trajectories exhibiting unusual movement characteristics potentially associated with tank-cleaning operations. Since anomalous maritime activities are expected to represent only a small fraction of all vessel trajectories, the ability of HDBSCAN to naturally classify low-density samples as outliers provides an effective framework for unsupervised anomaly detection [11].

3
Data and Preprocessing

3.1 AIS Dataset Description

This chapter describes the AIS dataset, preprocessing pipeline and trajectory preparation procedures used in this thesis. The objective of the preprocessing stage was to transform large-scale raw vessel trajectory data into structured representations suitable for subsequent trajectory compression, feature engineering and machine learning analysis. The chapter focuses on the practical implementation and preparation of the dataset, while the theoretical foundations and methodological frameworks of the employed approaches are presented in Chapter 2 and Chapter 4, respectively.

The primary data source used in this study was AIS trajectory data stored in a PostgreSQL/PostGIS database maintained within the Division of Marine Technology at Chalmers University of Technology. The database contained historical AIS trajectory records for vessel traffic in the Baltic Sea and served as the foundation for all preprocessing and trajectory analysis performed in this thesis.

The primary large-scale dataset used in this thesis consisted of candidate tanker vessel trajectories extracted from monthly trajectory tables covering January to March 2021. The preprocessing and extraction workflow was developed iteratively through several stages of experimentation, beginning with exploratory analyses around the Port of Stenungsund before later being extended and scaled to larger datasets and additional regions.

The dataset used for the primary preprocessing and machine learning pipeline was obtained from the following monthly trajectory tables:

- `trajectories_2021_01`
- `trajectories_2021_02`
- `trajectories_2021_03`

The trajectory tables stored AIS observations in trajectory-array structures rather than as independent point records. Each trajectory entry contained arrays of timestamps, geographic coordinates, speed-over-ground values, course-over-ground values, vessel headings, and navigational status information associated with a specific vessel movement sequence.

Additional vessel metadata was retrieved from a separate ship information table linked through the MMSI identifier. This metadata included vessel IMO number, ship name, vessel dimensions, draught information, and ship type classifications. Candidate tanker vessels were identified using the AIS vessel classification field `type_of_ship_and_cargo`, where values between 80 and 89 correspond to tanker-related vessel categories according to IMO AIS ship-type standards. Although these

AIS classifications reliably identify tanker traffic, the classification codes alone do not always provide sufficient granularity to independently distinguish all tanker subclasses and cargo specializations.
Trajectory extraction was performed using PostgreSQL/PostGIS spatial operations. The stored trajectory geometries were expanded into point-level AIS observations using the `ST_DumpPoints()` function, enabling each AIS message within a trajectory to be analysed individually during preprocessing. The extracted point-level dataset contained the following attributes:

- trajectory identifier
- MMSI
- IMO number
- vessel name
- point index within trajectory
- latitude
- longitude
- timestamp
- SOG
- navigational status
- COG
- vessel heading

The extraction process produced a total of 13,359,189 AIS observations across the three monthly datasets. Table 3.1 summarises the number of extracted AIS points per month.

Table 3.1: Extracted AIS observations from monthly trajectory tables.

Dataset	Number of AIS observations
January 2021	4,522,969
February 2021	3,955,066
March 2021	4,881,154
Total	13,359,189

In addition to AIS navigational information, vessel deadweight tonnage (DWT) data was integrated into the dataset using external vessel metadata retrieved from the S&P Maritime Portal (Sea-web/IHS database). The metadata enrichment process linked IMO numbers to vessel deadweight value. DWT information was included in the preprocessing stage to support filtering of small service vessels and to focus the analysis on seagoing tanker operations relevant to the objectives of this thesis. Vessels with DWT below 1000 tonnes were removed from the dataset
In addition to AIS and vessel metadata, supplementary operational washing-system information was provided by Scanjet AB to support the environmental estimation component of the thesis, as described in Objective O6 and Research Question RQ4. The provided data included vessel-level information related to tank-cleaning configurations, including vessel type, number of cargo tanks, nominal cleaning flow capacity, associated system identifiers (order and part numbers) and the number of installed cleaning machines.

These operational specifications were not used during trajectory preprocessing, anomaly detection, trajectory compression or representation learning. Instead, the Scanjet metadata was retained as auxiliary information for later scenario-based estimations of potential washwater discharge volumes associated with detected anomalous vessel behaviour.

The AIS dataset described in this section formed the empirical foundation for all subsequent preprocessing, feature engineering, compression, image generation and anomaly detection procedures presented in the remainder of this thesis.

3.2 Spatial Data Extraction

The AIS database used in this thesis contained vessel trajectories covering the Baltic Sea area and included multiple vessel categories beyond the scope of this study. The initial stage of the work focused on the Port of Stenungsund, which represents one of the major Swedish industrial and chemical shipping regions. Stenungsund was selected as the first prototype study area due to its high concentration of tanker-related industrial traffic associated with petrochemical operations. Exploratory analyses were initially performed using AIS trajectories from vessels operating within a radius of approximately 50 km from the port area. This prototype extraction stage was used to visually inspect vessel trajectories, evaluate preprocessing procedures, test trajectory compression methods and investigate the feasibility of AIS-based anomaly detection approaches prior to scaling the methodology to larger datasets.

Spatial filtering was implemented using PostgreSQL/PostGIS geographic operations. Port locations were represented as geographic coordinate reference points, and vessel trajectory points were extracted using the `ST_DumpPoints()` function. Spatial filtering was subsequently performed using the `ST_DWithin()` function, which enabled selection of AIS observations located within a specified geographic radius from the selected port coordinates. The extraction process retained only vessel trajectory points satisfying the specified distance threshold.

The initial Stenungsund prototype was later extended to two additional Swedish port regions: Nynäshamn and Umeå. These locations were selected in order to evaluate whether the preprocessing workflow could be applied across different geographic and operational contexts.

The three-port extraction stage primarily served as a development and validation phase for the preprocessing workflow and did not define the complete dataset used in the final large-scale pipeline. Following validation of the spatial extraction procedures, the methodology was subsequently scaled to the full Jan-Mar 2021 candidate tanker AIS dataset described in Section 3.1. This progressive scaling strategy reduced implementation risk while supporting iterative validation of preprocessing and compression performance across different geographic and temporal contexts.

The coordinates used for spatial extraction are summarised in Table 3.2. A uniform extraction radius of 50 km was applied for all study regions.

Table 3.2: Port coordinates and extraction radius used during spatial filtering.

Port region	Latitude	Longitude	Radius (km)
Stenungsund	58.0775	11.81	50
Nynäshamn	58.55	17.58	50
Umeå	63.41	20.20	50

The prototype multi-port extraction stage produced approximately 1.25 million AIS observations across the three selected regions. Table 3.3 summarises the number of extracted AIS observations per port region.

Table 3.3: AIS observations extracted during the prototype multi-port stage.

Port region	Extracted AIS observations
Stenungsund	1,006,825
Nynäshamn	182,952
Umeå	55,244
Total	1,245,021

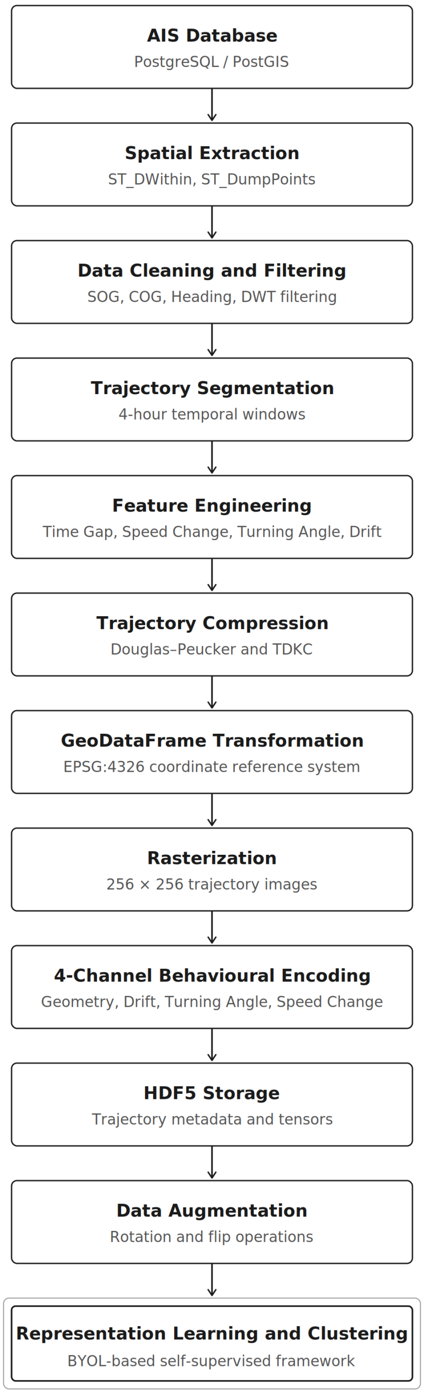


Figure 3.1: Overview of the AIS preprocessing and trajectory representation pipeline used in this thesis.

Figure 3.1 summarises the complete preprocessing and trajectory representation workflow developed in this thesis. The pipeline progressively transforms large-scale AIS trajectory data into multi-channel image representations suitable for representation learning, clustering, and anomaly detection.

3.3 Data Cleaning and Filtering

AIS trajectory data frequently contains inconsistencies, missing values, duplicated observations, and physically implausible navigational records arising from sensor inaccuracies, transmission interruptions, equipment malfunction, or limitations in AIS reporting systems. Consequently, preprocessing and quality-control procedures are necessary before AIS data can be reliably used for trajectory analysis and machine learning applications. The objective of the cleaning stage in this thesis was therefore to remove clearly invalid observations while preserving legitimate vessel behaviour relevant to the investigation of anomalous navigational patterns.

Filtering thresholds were selected conservatively in order to remove physically implausible AIS observations while minimising distortion of legitimate vessel movements. The preprocessing workflow was iteratively refined through exploratory trajectory visualisation and progressive scaling experiments performed during the prototype stages described in Section 3.2. Additional vessel metadata was subsequently integrated using IMO-linked ship information retrieved from the S&P Maritime Portal database. The integrated metadata included vessel DWT, which was used as an additional filtering criterion in order to remove small service vessels and non-representative tanker traffic from the analysis.

Only vessels with available DWT information were retained. Subsequently, vessels with deadweight tonnage below 1000 tonnes were removed from the dataset. This threshold was selected in order to focus the analysis on operational seagoing tanker traffic while excluding smaller harbour, support, and auxiliary vessels that were unlikely to be relevant to the operational patterns investigated in this thesis.

3.3.1 Removal of Invalid Navigational Values

Several filtering operations were applied to remove AIS observations containing invalid or physically implausible navigational values.

SOG observations exceeding 40 knots were removed from the dataset. Sustained vessel speeds above this threshold are operationally unrealistic for tanker vessels and were therefore interpreted as likely AIS transmission or recording artefacts.

Additional filtering was applied to COG and vessel heading observations. AIS records with COG or heading values greater than 360 degrees were removed, since such values fall outside the valid angular range defined by AIS navigational standards. These observations were interpreted as corrupted or improperly encoded AIS records.

The filtering operations were implemented sequentially using the following criteria:

- `SOG` $\leq$ `40`
- `COG` $\leq$ `360`
- `Heading` $\leq$ `360`
- non-missing DWT values
- `DWT` $>$ `1000`

3.3.2 Temporal Ordering and Consistency

Following navigational and metadata filtering, the AIS observations were temporally ordered according to vessel MMSI and timestamp. The AIS timestamps were converted into standard datetime format prior to ordering operations. Observations were then sorted chronologically for each vessel trajectory. Maintaining correct temporal order was particularly important for the computation of behavioural features dependent on consecutive vessel movements, including turning angle, speed variation, drift behaviour and temporal gaps between AIS transmissions.

3.3.3 Filtering Statistics

Table 3.4 summarises the effect of the preprocessing and filtering stages on the extracted AIS dataset.

Table 3.4: AIS preprocessing and filtering statistics.

Processing stage	Remaining AIS observations
Initial extracted dataset	13,359,189
After SOG filtering	13,339,024
After COG filtering	13,338,143
After heading filtering	13,082,711
After removal of missing DWT	13,032,430
After DWT filtering (>1000)	12,537,945
Final cleaned dataset	12,537,945

The cleaned and temporally ordered dataset obtained through these preprocessing operations formed the basis for the trajectory segmentation and feature engineering procedures described in the following sections.

3.4 Trajectory Segmentation

Raw AIS observations represent continuous streams of vessel movements transmitted over time. However, subsequent preprocessing stages in this thesis, including behavioural feature engineering, trajectory compression, image generation and machine learning analysis, required the AIS data to be organised into discrete and computationally manageable trajectory units. Consequently, a trajectory segmentation procedure was applied in order to partition the continuous AIS records into standardised analytical segments suitable for downstream processing.

In this thesis, a trajectory segment was defined as the collection of temporally ordered AIS observations associated with a single vessel within a fixed four-hour temporal window. The segmentation procedure was performed independently for each vessel using the MMSI identifier and chronologically ordered AIS timestamps.

The segmentation strategy adopted in this work was based on fixed temporal windows rather than complete voyage segmentation. Consequently, the segmented trajectories used throughout the thesis do not necessarily correspond to full port-to-port voyages or operational shipping routes. Instead, the segmentation approach was designed to produce standardised trajectory units appropriate for behavioural analysis, trajectory compression and machine learning representation learning.

The segmentation window length was selected iteratively during the exploratory stages of the preprocessing workflow described in Section 3.2. Initial experiments were conducted using shorter temporal windows in order to evaluate segmentation behaviour under different operational conditions. However, excessively short segmentation intervals frequently produced fragmented trajectories containing insufficient navigational variation for reliable behavioural analysis. Conversely, excessively long temporal windows increased the likelihood of combining unrelated vessel behaviours into a single analytical unit. A four-hour segmentation interval was ultimately selected as a practical compromise between behavioural continuity, computational tractability and trajectory interpretability during downstream processing.

Following segmentation, behavioural features including turning angle, speed variation, drift behaviour and temporal transmission gaps were computed using consecutive AIS observations within each trajectory segment. Since these calculations relied on temporal differencing operations between sequential observations, missing values were introduced at the beginning of some trajectories. AIS observations containing missing values generated during feature computation were therefore removed prior to construction of the final trajectory dataset.

Behavioural features were intentionally computed prior to trajectory compression in order to preserve local navigational characteristics that could otherwise be distorted or removed during compression operations. This was particularly important for preserving manoeuvring-related behavioural signatures including abrupt course changes, local speed variations and drift behaviour relevant to subsequent anomaly analysis.

Very short trajectories containing insufficient sequential information were also removed from the dataset. Single-point trajectories and trajectories containing fewer than two valid AIS observations were excluded since such trajectories could not support meaningful behavioural analysis, trajectory compression or image-based representation learning.

Table 3.5 summarises the trajectory segmentation process and the resulting dataset sizes following segmentation and trajectory filtering operations.

Table 3.5: Trajectory segmentation and filtering statistics.

Processing stage	Value
AIS observations after cleaning	12,537,945
Initial segmented trajectories	158,910
Trajectories after feature-based NaN removal	158,499
Removed single-point trajectories	344
Final valid trajectories	158,155

3.5 Feature Engineering

Raw AIS trajectory data primarily describes vessel movement through geographic coordinates and navigational observations such as speed, heading and course. However, spatial coordinates alone are often insufficient for identifying complex behavioural patterns associated with anomalous vessel operations.
Consequently, additional behavioural features were engineered from the AIS trajectories in order to provide richer representations of vessel movement dynamics for subsequent trajectory analysis and machine learning applications.
The feature engineering stage focused on extracting temporal and navigational characteristics from consecutive AIS observations within each segmented trajectory. All behavioural features were computed prior to trajectory compression in order to preserve local navigational signatures that could otherwise be distorted or removed during compression operations. This preprocessing strategy was particularly important for retaining manoeuvring-related behavioural patterns relevant to anomaly detection.
The engineered behavioural features used in this thesis included:

- speed variation
- turning angle
- drift behaviour

3.5.1 Speed Variation

The speed variation feature, referred to as `speed_change`, was computed as the difference in SOG between consecutive AIS observations within each trajectory segment.
This feature captures local acceleration and deceleration behaviour during vessel movement. Abrupt speed variations may indicate manoeuvring activity, operational adjustments, changes in navigational state or irregular vessel behaviour relevant to anomaly analysis.

3.5.2 Turning Angle

The turning angle feature was derived from differences between consecutive COG observations within each trajectory segment. Angular differences were normalised to the interval $[-180°, 180°]$ in order to preserve directional consistency and avoid discontinuities associated with circular angular measurements.
Turning angle represents short-term manoeuvring behaviour and directional change intensity during vessel movement. Large turning angles may correspond to abrupt manoeuvres, route deviations, unstable navigation patterns or local operational activities potentially relevant to the identification of anomalous vessel behaviour.

3.5.3 Drift Behaviour

The drift feature, referred to as `drift`, was computed as the difference between vessel COG and vessel heading measurements.

Drift behaviour describes deviation between the vessel heading direction and the actual movement direction over ground. Such deviation may arise due to environmental influences including wind, waves and currents or from operational manoeuvring behaviour and vessel control adjustments. Drift-related behaviour may therefore provide additional information regarding local navigational instability and vessel operational characteristics.
Angular drift values were normalised to the interval $[-180°, 180°]$ in order to ensure consistent representation of directional differences.

3.5.4 Feature Computation and Sequential Consistency

All behavioural features were computed using temporally ordered AIS observations within each segmented trajectory. Since several feature calculations relied on temporal differencing operations between consecutive AIS messages, missing values were naturally introduced at the beginning of some trajectory segments. Observations containing undefined feature values were therefore removed prior to subsequent trajectory compression and image generation procedures.
The feature engineering stage transformed the cleaned AIS trajectories into enriched behavioural representations containing both geometric and navigational characteristics.

3.5.5 Feature Statistics

Table 3.6 summarises the statistical characteristics of the engineered behavioural features used in this thesis.

Table 3.6: Statistical summary of engineered behavioural features.

Feature	Mean	Std	Min	Max
Time gap (min)	2.89	2.89	0.00	230.97
Speed change	-0.0003	0.3537	-32.60	32.10
Turning angle	-0.0378	34.9610	-180.00	180.00
Drift	-1.0596	56.4229	-180.00	179.90

3.6 Trajectory Compression

The cleaned and segmented AIS dataset used in this thesis contained more than 12.5 million AIS observations distributed across over 158,000 trajectory segments. Although AIS data provides detailed descriptions of vessel movement behaviour, the high temporal resolution and large number of recorded observations introduce significant computational challenges related to storage, image generation and machine learning scalability. In many cases, consecutive AIS observations contain highly redundant positional information due to stable vessel movements during normal navigation. Consequently, trajectory compression was applied in order to reduce data volume while preserving the navigational and behavioural characteristics relevant to subsequent anomaly analysis.

The primary objective of the compression stage was not only geometric simplification of vessel trajectories, but also preservation of behavioural information associated with vessel manoeuvring patterns, speed variation and navigational dynamics.
Behavioural features including turning angle, drift behaviour, speed variation and temporal transmission gaps were therefore computed prior to trajectory compression. This preprocessing strategy ensured that local navigational characteristics were preserved before trajectory simplification procedures removed intermediate AIS observations.

3.6.1 DP Compression

The DP algorithm was used as the primary geometric trajectory compression method in this thesis. The DP compression workflow was progressively evaluated under increasing data scales during the exploratory stages of the preprocessing pipeline. Initial experiments focused on AIS trajectories extracted around the Port of Stenungsund using single-month datasets. The methodology was subsequently extended to multi-month and multi-port datasets before finally being scaled to the complete Jan-Mar 2021 AIS dataset.
Selection of the DP compression threshold parameter was performed empirically through exploratory compression experiments. Multiple compression thresholds were evaluated in order to balance trajectory simplification against preservation of navigational behaviour and local trajectory structure. An elbow-method-inspired analysis was used to identify practical compression settings capable of substantially reducing data volume without excessive loss of behavioural information. Based on these experiments, a compression threshold of 50 meters was selected for the primary DP preprocessing pipeline.
The DP compression procedure was applied independently to each segmented trajectory. Following compression, trajectories containing insufficient sequential information for downstream processing were excluded from further analysis.

3.6.2 TDKC

In addition to geometric compression, this thesis also investigated TDKC as a behaviour-aware trajectory compression approach. The TDKC method was evaluated in order to investigate its capability to preserve local vessel manoeuvring behaviour during AIS trajectory simplification. Similar to the DP workflow, TDKC experiments were initially evaluated using smaller prototype datasets before later being extended to Jan-Mar 2021 AIS dataset.
In practice, the TDKC configuration adopted in this thesis reduced almost all trajectory segments to the minimum representation of three points, largely independent of the complexity of the underlying vessel movement, as documented in Section 3.6.4. This behaviour, and its consequences for downstream representation learning and clustering, is examined in detail in Section 6.2. On the basis of that analysis, the finalized large-scale image-generation and storage pipeline (Section 3.7) was constructed using DP-compressed trajectories, while TDKC was retained as part of the comparative evaluation.

3.6.3 Compression Workflow

Trajectory compression was implemented independently for each segmented trajectory after completion of the cleaning, segmentation, and behavioural feature engineering stages. The compression workflow therefore followed the sequential preprocessing order:

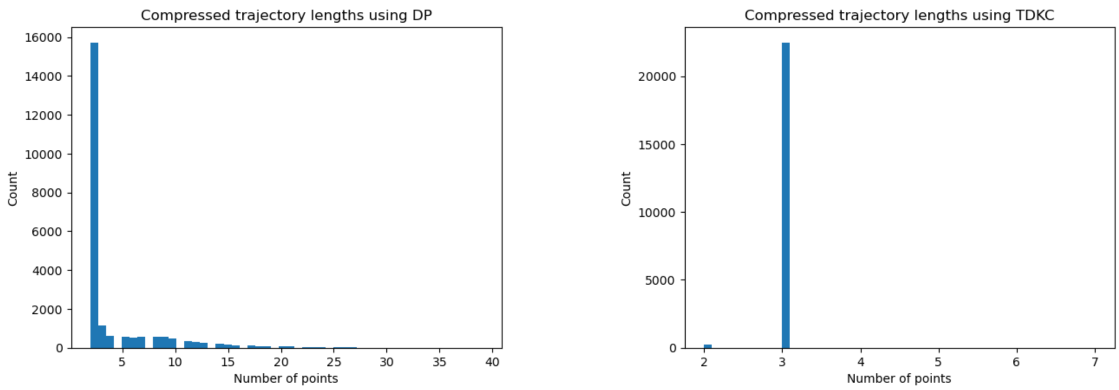
1. AIS data extraction
2. data cleaning and filtering
3. trajectory segmentation
4. behavioural feature engineering
5. trajectory compression
6. image generation and dataset storage

This preprocessing order was intentionally designed to preserve behavioural trajectory characteristics before simplification operations removed intermediate AIS observations.

Following compression, the resulting trajectories were transformed into GeoDataFrame representations and subsequently converted into image-based trajectory representations used during the machine learning stages described later in this thesis.

3.6.4 Compression Statistics

The two compression methods differed markedly in the distribution of compressed trajectory lengths they produced. Figure 3.2 presents these distributions for the three-port dataset. DP produced a broad, heavy-tailed distribution in which the number of retained points scaled with the geometric complexity of the trajectory, extending up to a maximum of 39 points. The TDKC configuration, by contrast, reduced almost every trajectory to exactly three points, with a maximum observed length of seven points across the dataset. The implications of this difference for representation learning and clustering are examined in Section 6.2.



(a) DP compression. **(b)** TDKC compression.

Figure 3.2: Distribution of compressed trajectory lengths for the two compression methods on the three-port dataset.

Table 3.7: Comparison of DP and TDKC trajectory compression on the three-port dataset. Both methods were applied to the same set of extracted AIS observations.

	DP	TDKC
AIS points after filtration	1,245,021	1,245,021
Compressed AIS points	91,826	67,917
Overall compression rate	0.9154	0.9374

Table 3.8 summarises the trajectory compression results obtained using the finalized DP preprocessing pipeline with four-hour trajectory segmentation and an epsilon value of 50.

Table 3.8: Trajectory compression statistics for the DP preprocessing pipeline.

Compression stage	Trajectories	AIS points
Before DP compression	158,155	12,378,691
After DP compression	158,155	1,287,987

The DP compression pipeline reduced the total number of AIS observations by approximately 89.6% while preserving the overall trajectory structure and vessel movement patterns required for downstream image generation, representation learning and anomaly analysis. The compression process reduced the number of AIS observations within each trajectory segment while preserving the original trajectory segmentation structure. Consequently, the compression procedure reduced computational complexity and storage requirements while maintaining the trajectory patterns necessary for subsequent machine learning analysis.

3.7 Image Generation and HDF5 Storage

Following trajectory compression, the AIS trajectories were transformed into image-based representations suitable for machine learning and self-supervised representation learning applications. Direct use of raw AIS trajectories in machine learning models is challenging due to varying trajectory lengths, irregular temporal sampling, and inconsistent spatial scales between vessel movements. Consequently, a standardised image-generation pipeline was developed in order to convert compressed trajectories into fixed-size tensor representations capable of preserving both spatial and behavioural vessel characteristics.

3.7.1 GeoDataFrame Conversion

Each compressed trajectory was first converted into a Pandas DataFrame representation containing the navigational and behavioural variables associated with the trajectory segment. Geographic point geometries were subsequently generated using vessel longitude and latitude coordinates.

The resulting trajectory representations were converted into GeoDataFrames using the EPSG:4326 geographic coordinate reference system. Each GeoDataFrame therefore represented a temporally ordered vessel trajectory containing both spatial geometry and associated behavioural features including drift behaviour, turning angle, and speed variation.

3.7.2 Rasterized Trajectory Representation

The GeoDataFrame trajectories were transformed into rasterized image representations using a fixed spatial resolution of 256×256 pixels. Rasterization enabled the conversion of variable-length AIS trajectories into standardised image tensors suitable for machine learning processing.
During rasterization, vessel trajectories were spatially normalised and projected onto fixed-size image grids. This preprocessing procedure preserved the geometric structure of vessel movements while enabling consistent trajectory representation across different spatial regions and trajectory lengths.
The rasterization process transformed each vessel trajectory into a structured spatial representation capable of encoding both geometric trajectory information and associated behavioural characteristics.

3.7.3 Multi-Channel Behavioural Encoding

The final image-generation pipeline used a four-channel trajectory representation in order to combine geometric vessel movement information with behavioural navigational characteristics.
The generated trajectory images consisted of the following channels:

- Channel 1: rasterized vessel trajectory geometry
- Channel 2: drift behaviour
- Channel 3: turning-angle behaviour
- Channel 4: speed-change behaviour

The first channel represented the geometric vessel trajectory obtained from the rasterized AIS path. The remaining channels encoded behavioural features computed during the feature-engineering stage described in Section 3.5.
This multi-channel representation enabled simultaneous encoding of spatial vessel movement patterns and local navigational dynamics within a unified machine learning representation. Consequently, the generated trajectory tensors preserved both geometric and behavioural information relevant to subsequent anomaly analysis and self-supervised learning.

3.7.4 Geometry Smoothing

Following rasterization, smoothing operations were applied to the geometric trajectory channel in order to improve trajectory continuity and reduce sparsity effects introduced during rasterization. Convolution-based smoothing using a kernel size of three pixels was applied exclusively to the geometric trajectory channel.
The smoothing procedure improved continuity of sparse trajectory paths and enhanced the stability of the resulting image representations during downstream machine learning processing.

3.7.5 HDF5 Storage Structure

The generated multi-channel trajectory images were stored using the Hierarchical Data Format version 5 (HDF5). HDF5 storage was selected due to its scalability, efficient data access capabilities and suitability for large-scale machine learning datasets.

Each stored trajectory dataset contained both the generated image representation and associated trajectory metadata. The stored metadata included:

- MMSI identifier
- IMO number
- vessel name
- trajectory start timestamp
- trajectory end timestamp
- trajectory coordinates
- trajectory length

The HDF5 structure enabled efficient storage and retrieval of large numbers of compressed trajectory representations while preserving the navigational metadata required for subsequent trajectory interpretation and analysis.

3.7.6 Tensor Formatting and Data Augmentation

The stored trajectory images were converted into tensor representations compatible with the PyTorch machine learning framework. The resulting tensors followed a channel-first representation format suitable for convolutional neural network processing.

In order to improve representation robustness during self-supervised learning, multiple augmented trajectory views were generated using rotational and mirror-based spatial transformations. These augmentations included:

- 180-degree rotational transformations
- horizontal and vertical reflection operations

The augmentation procedure generated multiple spatially transformed representations of the same trajectory while preserving the underlying navigational behaviour. This preprocessing strategy improved invariance and robustness during subsequent representation learning stages.

3.7.7 Dataset Statistics

Table 3.9 summarises the final image-generation and HDF5 dataset statistics obtained following the finalized Douglas–Peucker compression and rasterization workflow.

Table 3.9: Image-generation and HDF5 dataset statistics.

Dataset characteristic	Value
Compressed trajectories	158,155
Valid trajectories after rasterization filtering	91,431
Stored HDF5 datasets	91,431
Image resolution	256×256
Number of image channels	4

The finalized image-generation pipeline operated on the compressed four-hour trajectory dataset produced by the DP compression workflow described in Section 3.6. During rasterization and image generation, trajectories containing fewer than three AIS observations after compression were excluded since such trajectories could not generate meaningful spatial representations for machine learning processing. Consequently, 91,431 valid trajectory representations were retained and stored in the final HDF5 dataset.

The resulting image-generation and storage pipeline transformed large-scale AIS trajectory data into standardised machine-learning-ready representations suitable for self-supervised representation learning and subsequent anomaly detection procedures.

4

Methods

4.1 Overall Methodological Framework

This chapter describes the machine learning methodology and representation learning framework developed for unsupervised analysis of AIS vessel trajectories. Building upon the preprocessing pipeline presented in Chapter 3, the objective of the methodology stage was to transform compressed multi-channel vessel trajectory representations into latent behavioural embeddings suitable for clustering, anomaly analysis and subsequent operational interpretation.
The proposed framework focused on learning behavioural vessel representations directly from large-scale AIS trajectory data without relying on manually labelled training examples.

The complete methodological workflow consisted of the following principal stages:

1. generation of multi-channel trajectory tensor representations

2. self-supervised representation learning using BYOL

3. latent embedding extraction

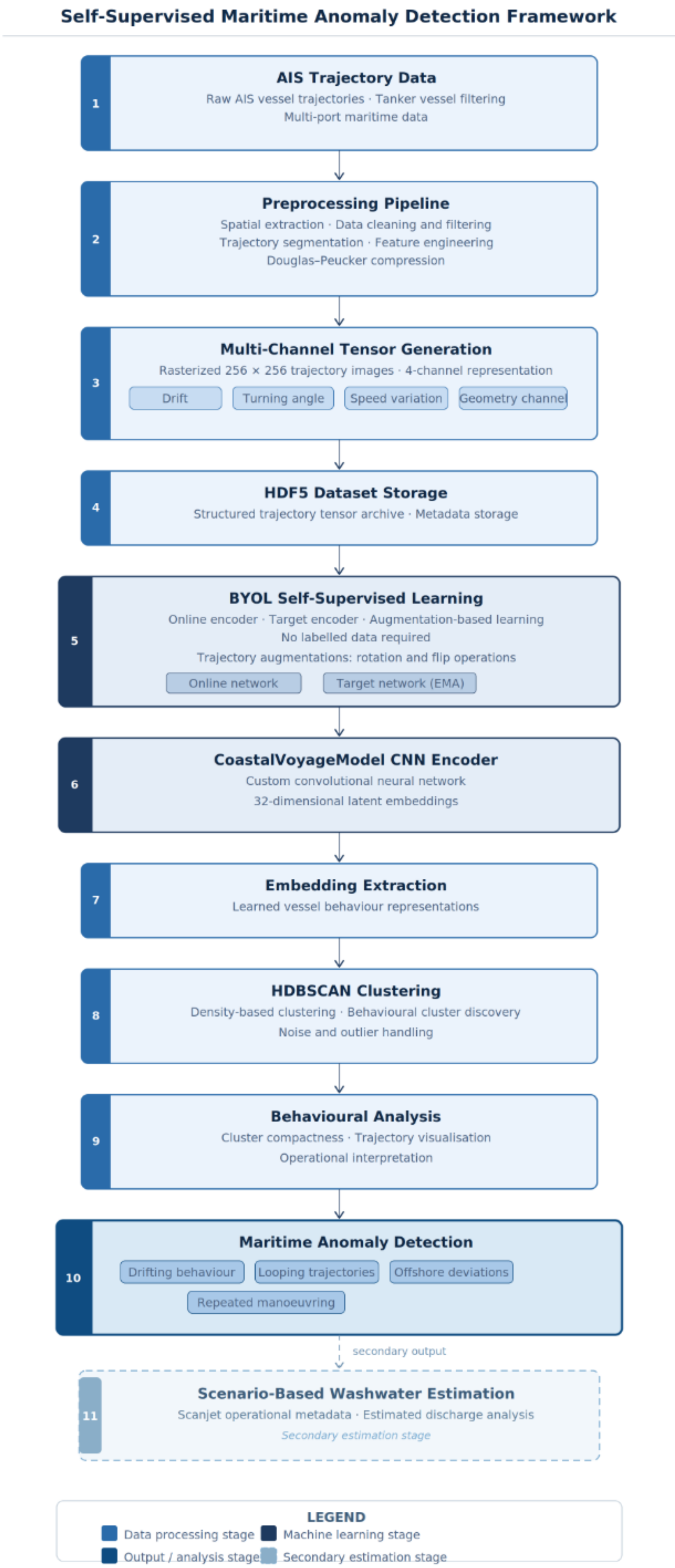
4. unsupervised clustering of learned embeddings

5. visual and operational interpretation of trajectory clusters

6. identification of anomalous vessel behaviour

7. scenario-based estimation of potential washwater discharge

Figure 4.1 summarises the complete machine learning and anomaly analysis framework developed in this thesis.


Self-Supervised Maritime Anomaly Detection Framework
1
AIS Trajectory Data
Raw AIS vessel trajectories · Tanker vessel filtering
Multi-port maritime data
2
Preprocessing Pipeline
Spatial extraction · Data cleaning and filtering
Trajectory segmentation · Feature engineering
Douglas–Peucker compression
3
Multi-Channel Tensor Generation
Rasterized 256 × 256 trajectory images · 4-channel representation
Drift
Turning angle
Speed variation
Geometry channel
4
HDF5 Dataset Storage
Structured trajectory tensor archive · Metadata storage
5
BYOL Self-Supervised Learning
Online encoder · Target encoder · Augmentation-based learning
No labelled data required
Trajectory augmentations: rotation and flip operations
Online network
Target network (EMA)
6
CoastalVoyageModel CNN Encoder
Custom convolutional neural network
32-dimensional latent embeddings
7
Embedding Extraction
Learned vessel behaviour representations
8
HDBSCAN Clustering
Density-based clustering · Behavioural cluster discovery
Noise and outlier handling
9
Behavioural Analysis
Cluster compactness · Trajectory visualisation
Operational interpretation
10
Maritime Anomaly Detection
Drifting behaviour
Looping trajectories
Offshore deviations
Repeated manoeuvring
secondary output
11
Scenario-Based Washwater Estimation
Scanjet operational metadata · Estimated discharge analysis
Secondary estimation stage
LEGEND
Data processing stage
Machine learning stage
Output / analysis stage
Secondary estimation stage

4.2 Self-Supervised Representation Learning

The representation learning framework implemented in this work was based on the BYOL self-supervised learning methodology introduced in Chapter 2. The primary objective of the BYOL framework was to learn compact latent representations capable of preserving meaningful vessel movement characteristics while remaining invariant to trajectory augmentations and geometric transformations.

The input to the self-supervised learning pipeline consisted of four-channel rasterised trajectory tensors generated during the preprocessing stage described in Chapter 3.

The BYOL framework implemented in this thesis consisted of two neural network branches referred to as the online network and the target network. Both branches shared the same underlying convolutional encoder architecture based on the custom `CoastalVoyageModel`. The online network was updated directly through gradient-based optimization, whereas the target network parameters were updated using an exponential moving average mechanism in order to provide stable representation targets during training.

For each augmented trajectory pair, the online network generated latent embeddings and projection vectors which were compared with the corresponding target-network representations. The training objective encouraged the latent representations of augmented views originating from the same vessel trajectory to remain similar in the embedding space while simultaneously preserving meaningful distinctions between different vessel movement patterns.

The self-supervised representation learning stage therefore enabled the extraction of compact latent vessel representations without requiring labelled anomaly data, thereby forming the foundation for the unsupervised clustering and anomaly analysis procedures described later in this chapter.

4.3 Neural Network Architecture

The self-supervised learning framework developed in this thesis utilised a custom convolutional neural network architecture referred to as the `CoastalVoyageModel`. The network was designed for the analysis of rasterised maritime trajectory tensors and implemented using the PyTorch deep learning framework.

The objective of the architecture was therefore not semantic object recognition, but rather the extraction of compact latent behavioural representations capable of preserving navigational movement characteristics from multi-channel AIS trajectory tensors.

The input to the network consisted of four-channel trajectory tensors with spatial dimensions of 256×256 pixels. The four channels represented the geometric vessel trajectory channel together with the engineered behavioural channels described in Chapter 3, including drift behaviour, turning angle and speed variation.

The `CoastalVoyageModel` architecture consisted of multiple convolutional processing stages combining convolutional layers, batch normalisation, rectified linear unit (ReLU) activation functions and spatial pooling operations. The initial convolutional layer employed a relatively large 7×7 kernel with stride 2 in order to capture broad geometric trajectory structures and large-scale vessel movement patterns present within the rasterised trajectory images.

Following the initial convolution stage, max-pooling operations were applied to reduce spatial dimensionality while preserving dominant trajectory structures and navigational features. Subsequent convolutional blocks progressively extracted increasingly abstract latent representations from the input trajectory tensors.

The intermediate convolutional stages utilised 3×3 convolutional kernels combined with batch normalisation and ReLU activation functions. Batch normalisation was incorporated in order to stabilise the learning process and improve convergence behaviour during self-supervised training. ReLU activation functions were used to introduce non-linear representational capacity while maintaining computational efficiency.

The convolutional architecture progressively increased the number of latent feature channels throughout the network while simultaneously reducing spatial resolution through strided convolution operations. This hierarchical feature extraction strategy enabled the model to capture both local navigational characteristics and larger-scale trajectory movement structures relevant to maritime behavioural analysis.

Following the convolutional stages, adaptive average pooling was applied to compress the final spatial feature maps into compact global feature representations independent of the exact spatial dimensions of the intermediate feature maps. The pooled feature representations were flattened and processed by a fully connected layer producing a final 32-dimensional latent embedding vector for each vessel trajectory tensor.

Compared with larger deep convolutional architectures commonly used for natural image analysis, the proposed network maintained a relatively lightweight structure in order to reduce computational complexity while remaining sufficiently expressive for maritime trajectory representation learning. This design choice was important given the large-scale AIS trajectory dataset and the computational demands associated with self-supervised representation learning on hundreds of thousands of vessel trajectories. The network was trained entirely from scratch without the use of pretrained image-classification weights. Figure 4.2 illustrates the overall structure of the `CoastalVoyageModel` architecture used throughout this thesis.

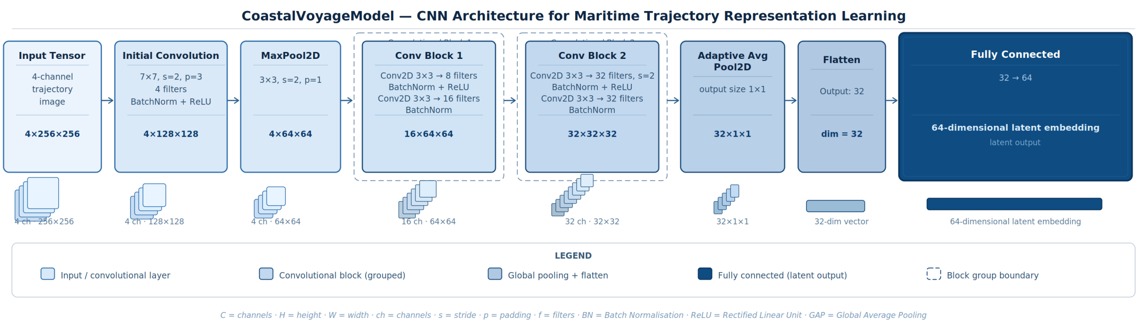


Figure 4.2: Architecture of the custom `CoastalVoyageModel` convolutional neural network used for self-supervised maritime trajectory representation learning.

4.4 Training Configuration and Augmentation Strategy

Prior to training, the dataset was randomly divided into training and testing subsets using an 80/20 split ratio. A fixed random seed was used during dataset partitioning in order to ensure reproducibility of the training and evaluation procedures. Data loading and batch preparation were implemented using the `DataLoader` framework with parallel worker processes in order to improve training efficiency during large-scale trajectory processing.
The final training configuration adopted for the large-scale experiments utilised a batch size of 64 trajectory tensors and a total training duration of 50 epochs. The learning process employed the Adam optimisation algorithm with an initial learning rate of 3×10^{-6}. In order to improve convergence stability during extended self-supervised training, an exponential learning-rate scheduler was applied with a decay factor of 0.98.
Table 4.1 summarises the principal training parameters used throughout the final representation learning experiments.

Table 4.1: Final training configuration used for self-supervised representation learning.

Parameter	Value
Learning framework	PyTorch
Training methodology	BYOL
Input tensor dimensions	$4 \times 256 \times 256$
Batch size	64
Training epochs	50
Learning rate	3×10^{-6}
Learning-rate scheduler	ExponentialLR
Scheduler decay factor	0.98
Latent embedding dimension	32
Dataset split ratio	80/20
Random seed	42
GPU acceleration	CUDA

Prior to representation learning, all trajectory tensors were normalised using min-max normalisation in order to stabilise numerical behaviour across samples exhibiting different navigational intensities and trajectory densities. This normalisation strategy improved convergence behaviour during self-supervised optimisation while preserving relative spatial and behavioural trajectory structures.

4.5 Latent Embedding Extraction

Following completion of the self-supervised training stage, latent vessel trajectory embeddings were extracted from the trained `CoastalVoyageModel` in order to con-

struct behavioural representations suitable for downstream clustering and anomaly analysis.
During inference, each trajectory tensor from the compressed HDF5 dataset was forwarded through the trained neural network without augmentation transformations. The resulting latent embeddings therefore represented the network's learned behavioural interpretation of each vessel trajectory under the final trained representation space.
Compared with the original high-dimensional rasterised trajectory tensors, the extracted latent vectors provided more compact representations suitable for efficient downstream clustering analysis. The embedding extraction stage therefore transformed complex multi-channel vessel trajectory tensors into low-dimensional behavioural representations wiht the intent to encode characteristics required for anomaly discovery and operational interpretation.
The resulting latent embedding vectors served as the primary input to the unsupervised clustering framework described in the following section. Clustering was performed directly within the learned embedding space in order to identify behavioural vessel groups, sparse trajectory regions, and potentially anomalous movement patterns.
In addition to clustering analysis, the learned embedding space was also explored through interactive visualisation procedures in order to investigate cluster compactness, latent representation structure and behavioural trajectory relationships. These analyses supported the qualitative interpretation of the learned vessel representations and assisted in the identification of unusual navigational behaviour within the large-scale AIS dataset.

4.6 Unsupervised Clustering and Behavioural Analysis

Following latent embedding extraction, unsupervised clustering techniques were applied in order to identify groups of vessel trajectories exhibiting similar navigational characteristics within the learned embedding space. The objective of the clustering stage was to discover recurring vessel movement behaviours, identify sparse trajectory regions and support the detection of potentially anomalous maritime activity without relying on predefined anomaly labels.
The clustering methodology adopted in this thesis was based primarily on the HDBSCAN algorithm introduced in Chapter 2. The input to the clustering stage consisted of the 32-dimensional latent embedding vectors extracted from the trained `CoastalVoyageModel`. The final clustering configuration utilised a minimum cluster size of 70 trajectory embeddings. This configuration provided a balance between cluster stability and sensitivity to less common vessel movement patterns during large-scale behavioural analysis. Smaller cluster thresholds produced excessively fragmented cluster structures during preliminary experimentation, whereas larger thresholds reduced sensitivity to operationally relevant trajectory variations. HDBSCAN assigned the trajectory embeddings to behavioural clusters or classified them as noise when they did not belong to a sufficiently dense cluster. The resulting

clusters were subsequently examined to identify recurring vessel movement patterns and potential anomalous behaviour.

The clustering results were assessed using both quantitative and qualitative criteria. The Silhouette coefficient and Davies-Bouldin index were used as internal clustering validation metrics. The Silhouette coefficient measures how similar trajectories are to other trajectories within their assigned cluster compared with trajectories in other clusters, where higher values indicate better cluster separation and compactness. The Davies-Bouldin index provides a complementary measure of cluster compactness and separation, where lower values indicate better clustering structure. The proportion of trajectories classified as HDBSCAN noise was also reported separately. In addition to these quantitative measures, visual inspection of the embedding projections and representative trajectory samples was performed to assess the resulting behavioural groupings.

HDBSCAN classified some trajectory representations as noise because they did not belong to sufficiently dense clusters. These noise trajectories were treated as candidates for further anomaly analysis rather than being automatically classified as anomalous. The identified clusters and noise trajectories were therefore examined together with the corresponding vessel trajectories to determine whether they exhibited unusual navigational behaviour.

The trajectory groups were visually inspected to identify movement patterns that could be relevant to the anomaly-analysis task. Particular attention was given to behaviours such as drifting, repeated manoeuvring, looping trajectories and other unusual navigational patterns. These observations were used to support the interpretation of the clustering results and to identify trajectories that warranted further investigation. The observed behaviours were not treated as direct evidence of tank-cleaning activity or discharge.

The anomaly-analysis methodology adopted in this thesis therefore combined unsupervised density-based clustering with quantitative clustering assessment and operationally informed trajectory interpretation rather than relying exclusively on fully automated anomaly classification. This approach was considered appropriate for maritime AIS analysis due to the absence of verified anomaly labels and the operational complexity associated with real-world vessel movement behaviour.

Figure 4.3 illustrates the clustering and behavioural analysis workflow applied to the learned vessel trajectory embeddings.

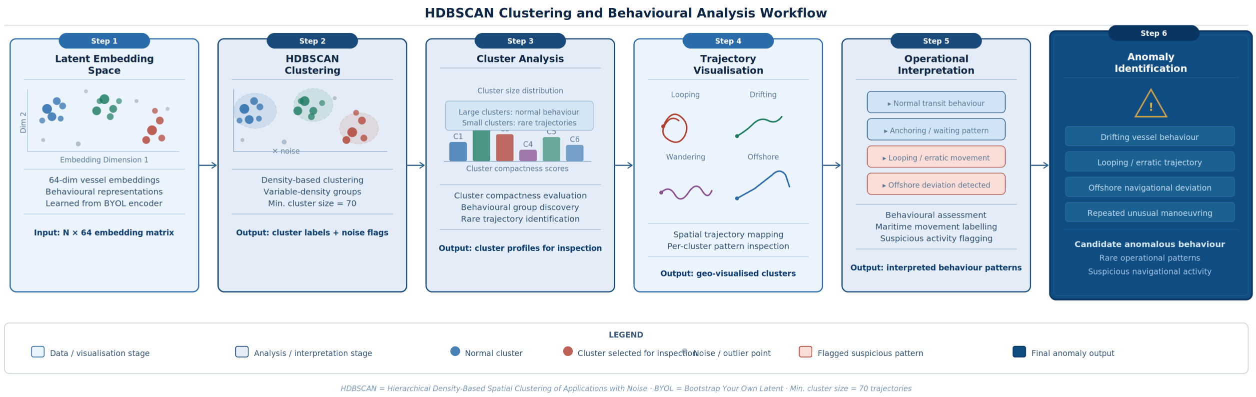


Figure 4.3: Overview of the unsupervised clustering and behavioural analysis workflow applied to the learned latent vessel embeddings.

4.7 Scenario-Based Washwater Estimation Methodology

In addition to unsupervised anomaly detection and behavioural trajectory analysis, this thesis also explored the potential use of detected anomalous vessel trajectories for scenario-based estimation of possible washwater discharge volumes associated with tank-cleaning operations. This component of the methodology addressed Objective O6 and Research Question RQ4 introduced in Chapter 1.

The washwater estimation methodology developed in this thesis did not attempt to directly measure or verify actual discharge events. Instead, the objective was to investigate whether anomalous vessel movement patterns identified through AIS trajectory analysis could be combined with operational vessel metadata in order to support exploratory estimation scenarios related to potential tank-cleaning discharge activity.

The estimation framework utilised supplementary operational metadata provided by `Scanjet AB`, including vessel type, deadweight tonnage, number of cargo tanks, cleaning-machine configurations, flow capacities and associated operational system specifications. These operational parameters were linked to selected anomalous vessel trajectories identified during the clustering and behavioural analysis stages described previously in this chapter.

Candidate anomalous trajectories considered during the estimation stage included vessel movement patterns exhibiting prolonged low-speed wandering behaviour, repeated manoeuvring activity, looping trajectories and offshore deviations potentially consistent with extended operational activity at sea. However, the presence of anomalous navigational behaviour alone was not interpreted as evidence of illegal discharge activity. Instead, these trajectory patterns were treated solely as candidate operational scenarios suitable for exploratory environmental estimation.

The estimation methodology was therefore formulated as a scenario-based analytical framework relying on operational assumptions rather than deterministic environmental measurement procedures. Estimated discharge volumes were derived by combining vessel operational metadata with assumed tank-cleaning activity durations and cleaning-system flow-rate specifications obtained from the supplementary Scanjet dataset.

Operational discharge scenarios were generated by estimating potential cleaning durations associated with selected anomalous vessel trajectories and subsequently combining these durations with the corresponding cleaning-system flow capacities and vessel tank configurations. Depending on the available vessel-specific operational metadata, the estimation procedure additionally considered the number of installed cleaning machines and the reported cleaning-system specifications associated with individual vessels.

The resulting estimates therefore represented hypothetical operational discharge ranges rather than empirically verified washwater quantities. The methodology was intended as an exploratory extension investigating how anomalous AIS movement patterns and operational assumptions could be combined in hypothetical tank-cleaning and discharge scenarios.

No causal environmental conclusions were inferred directly from the estimated discharge scenarios.

Figure 4.4 summarises the scenario-based washwater estimation framework developed in this thesis.

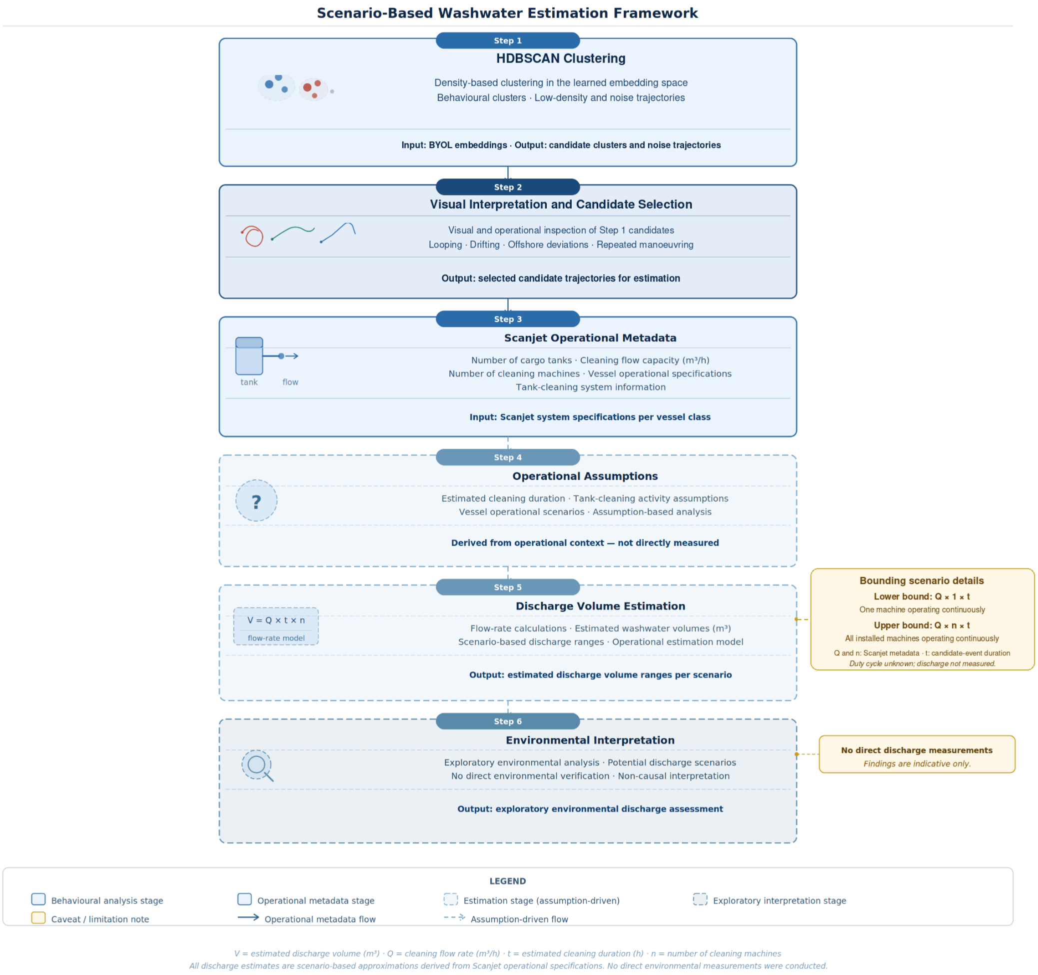


Figure 4.4: Overview of the scenario-based washwater estimation methodology combining anomalous trajectory analysis with operational vessel metadata provided by Scanjet AB.

4.8 Methodological Limitations and Experimental Considerations

Several methodological limitations and experimental considerations should be acknowledged when interpreting the results and conclusions presented in this thesis. The representation learning framework depended on several experimentally selected hyperparameters, including trajectory segmentation duration, compression thresholds, latent embedding dimensionality and clustering parameters. Different parameter configurations may potentially produce alternative embedding structures and clustering outcomes. While the final experimental configuration adopted in this thesis was selected based on cluster stability, representation quality and operational interpretability, the resulting behavioural group structures should still be interpreted within the context of the chosen experimental setup.

The self-supervised representation learning framework also introduced interpretability limitations commonly associated with deep neural network methodologies. Although the learned latent embedding space captured meaningful vessel movement characteristics suitable for clustering analysis, the precise internal feature representations learned by the neural network were not directly interpretable in a deterministic or physically explainable manner. Consequently, interpretation of the learned behavioural representations relied primarily on downstream trajectory visualisation and cluster-level behavioural analysis.

Additional limitations were associated with the scenario-based washwater estimation methodology developed as part of the exploratory environmental analysis framework. The operational discharge estimates generated in this thesis were derived entirely from assumed operational scenarios combined with supplementary Scanjet vessel metadata. No direct environmental measurements, discharge sensors, onboard operational logs or verified cleaning-event observations were available for validation purposes.

Accordingly, the estimated washwater discharge volumes presented later in this thesis should be interpreted exclusively as hypothetical operational estimation ranges rather than empirically validated environmental measurements. Furthermore, the methodology did not establish causal relationships between detected anomalous trajectories and actual discharge activity. The estimation framework instead served as an exploratory analytical extension intended to investigate how operational vessel metadata and anomalous AIS behaviour could potentially be combined within a broader maritime environmental analysis context.

Despite these limitations, the methodology developed in this thesis demonstrates the potential applicability of self-supervised representation learning and unsupervised behavioural clustering for large-scale maritime trajectory analysis. The framework additionally illustrates how AIS-based behavioural anomaly analysis may be combined with supplementary operational metadata in order to support exploratory maritime environmental assessment scenarios under explicitly stated methodological assumptions and limitations.

5

Results

5.1 Experimental Progression and Pipeline Refinement

The final anomaly-detection framework presented in this thesis was developed through a sequence of iterative experiments investigating the influence of dataset scale, trajectory compression methodology, behavioural feature selection and temporal segmentation configuration on the learned embedding space and the resulting clustering behaviour. To quantitatively assess the resulting clustering structures, the number of HDBSCAN clusters, the proportion of trajectories classified as noise, the Silhouette coefficient and the Davies-Bouldin index were considered. The number of clusters and noise proportion were used as descriptive measures of the clustering outcome, while the Silhouette coefficient and Davies-Bouldin index were used as internal clustering validation metrics. Higher Silhouette values indicate greater within-cluster compactness and separation, whereas lower Davies-Bouldin values indicate more compact and better-separated clusters. HDBSCAN noise trajectories were excluded from the calculation of the Silhouette coefficient and Davies-Bouldin index and were instead reported separately as the noise percentage. Consequently, a smaller number of clusters or a lower noise proportion was not considered sufficient evidence of improved clustering performance.

Table 5.1: Summary of the principal experimental configurations investigated during framework refinement.

ID	Dataset scope	Comp.	Feat.	Segmentation	Traj.
E1	1 port	DP	3	2-hour	5,237
E2	1 port	TDKC	3	2-hour	18,571
E3	1 port	TDKC	2	2-hour	18,571
E4	3 ports	DP	3	2-hour	6,976
E5	3 ports	TDKC	3	2-hour	22,494
E6	Full	DP	2	2-hour	154,944
E7	Full	DP	3	2-hour	154,944
E8	Full	DP	3	4-hour	91,431

Table 5.2: Summary of the clustering results and validation metrics.

ID	Clusters	Noise (%)	Silhouette	Davies-Bouldin
E1	6	9.2	0.789	0.331
E2	25	41.4	0.500	0.467
E3	28	36.8	0.297	0.413
E4	10	31.2	0.644	0.303
E5	56	38.2	0.400	0.447
E6	53	68.7	0.143	0.600
E7	31	36.3	0.126	1.185
E8	4	16.4	0.754	0.498

The initial experiments were conducted using AIS data extracted from a single operational port over a three-month period. These baseline experiments employed DP trajectory compression together with the three engineered behavioural channels described in Chapter 3. E1 produced six HDBSCAN clusters, with 9.2% of trajectories classified as noise. The resulting Silhouette coefficient was 0.789 and the Davies-Bouldin index was 0.331, indicating relatively strong internal clustering structure among the trajectories assigned to clusters.

Subsequent experiments investigated the use of the TDKC trajectory compression algorithm as an alternative to DP compression. Although TDKC compression visually preserved vessel movement geometry effectively during preliminary trajectory inspection, the resulting latent embedding distributions produced fragmented and unstable clustering structures during downstream HDBSCAN analysis. In particular, the TDKC-compressed trajectories frequently generated sparse behavioural groupings and reduced cluster compactness, thereby reducing operational interpretability during anomaly analysis.

Additional feature-ablation experiments were performed in order to evaluate the influence of the engineered behavioural channels on representation-learning quality. These experiments reduced the behavioural representation from three engineered channels to two channels while preserving the remaining preprocessing configuration. The reduced feature configuration consistently produced weaker behavioural separation within the latent embedding space and generated less coherent clustering structures during HDBSCAN analysis. The results therefore indicated that the combined behavioural representation consisting of drift behaviour, turning-angle variation and speed-change characteristics contributed positively to the learned vessel representations.

However, the single-port dataset represented a geographically and operationally restricted subset of the available AIS data. The subsequent experiments therefore investigated whether increasing the geographical scope of the AIS data altered the resulting clustering structure and the associated internal clustering metrics. The experimental framework was extended from a single-port environment to a larger multi-port AIS dataset covering three operational ports. Under the DP configuration, E4 produced ten clusters with 31.2% noise, a Silhouette coefficient of 0.644 and a Davies-Bouldin index of 0.303. The corresponding TDKC configuration, E5, produced 56 clusters with 38.2% noise, a Silhouette coefficient of 0.400 and a Davies-Bouldin index of 0.447. Thus, under the three-port experimental configuration, DP produced a higher Silhouette coefficient, a lower Davies-Bouldin index and a lower noise proportion than TDKC. These results provide quantitative evidence of stronger internal clustering structure for the DP configuration. However, as discussed in Section 6.2, the two compression methods did not retain the same number of trajectories for clustering, and this difference should be considered when interpreting the noise proportions.

Further large-scale experiments incorporated the full three-month AIS dataset containing all selected tanker trajectories across the analysed operational regions during 2021. E6 and E7 used the same full-dataset trajectories, DP compression and two-hour segmentation, while differing in the number of behavioural features. Although the configuration of E6 had better performance according to the metrics, the reduced feature configuration consistently produced weaker behavioural separation within the latent embedding space and generated less coherent clustering structures during HDBSCAN analysis and visual inspection of the samples.

The comparison between the full-dataset feature configurations therefore provided no single metric that could independently justify selecting the three-feature representation. Nevertheless, the three-feature configuration was retained for the subsequent experiment because the feature-ablation results demonstrated that changing the behavioural channels affected the resulting clustering structure, while the final configuration was evaluated further using the four-hour segmentation setting.

The final experimental configuration adopted in this thesis utilised the full three-month AIS dataset together with DP trajectory compression, four-hour trajectory segmentation windows and the three-feature behavioural representation. The adoption of four-hour trajectory segmentation windows produced more behaviourally coherent trajectory samples and improved latent embedding consistency during self-supervised learning compared with earlier experimental configurations. The resulting HDBSCAN clustering structures demonstrated improved behavioural compactness, increased operational interpretability and more stable separation between dominant navigational patterns and sparse behavioural trajectory groups.

The iterative experimental progression therefore played a central role in refining the final anomaly-detection framework presented in this thesis. Rather than selecting a configuration solely according to the number of clusters produced, the experiments considered the number of clusters, HDBSCAN noise proportion, Silhouette coefficient and Davies-Bouldin index together with qualitative inspection of the embedding projections and representative trajectories. The quantitative metrics provided an internal assessment of cluster compactness and separation, while the qualitative inspection supported the subsequent behavioural interpretation of the identified trajectory groups. The final configuration was selected based on the combined evidence from these assessments and the requirements of the downstream anomaly-analysis task.

5.2 Final Clustering Results

The final clustering analysis was performed on a total of 91,431 learned vessel trajectory embeddings generated by the trained `CoastalVoyageModel`. Using a minimum cluster size parameter of 70, the HDBSCAN framework identified four principal behavioural clusters together with a set of sparse embedding-space trajectories classified as noise points.
Within the HDBSCAN framework, noise points correspond to embeddings that could not be reliably associated with a stable high-density behavioural cluster within the learned latent representation space.
Table 5.3 summarises the final clustering distribution obtained from the learned embedding space.

Table 5.3: Final HDBSCAN clustering distribution for the learned vessel trajectory embeddings.

Cluster ID	Number of Trajectories	Percentage (%)	Interpretation
Cluster 2	63,809	69.79	Dominant transit behaviour
Cluster 3	11,230	12.28	More variable navigational behaviour
Cluster 1	1,223	1.34	Sparse low-motion behaviour
Cluster 0	199	0.22	Specialised directional trajectories
Noise (-1)	14,970	16.37	Behaviourally uncommon trajectories

The resulting clustering structure demonstrated a highly uneven behavioural distribution within the learned embedding space. A dominant behavioural cluster containing approximately 70% of all trajectory embeddings represented the majority of common vessel movement behaviour observed within the AIS dataset, while smaller clusters captured increasingly specialised navigational movement regimes exhibiting greater behavioural variability.
Figure 5.1 presents the distribution of trajectory counts across the final HDBSCAN behavioural clusters and trajectories classified as noise points.

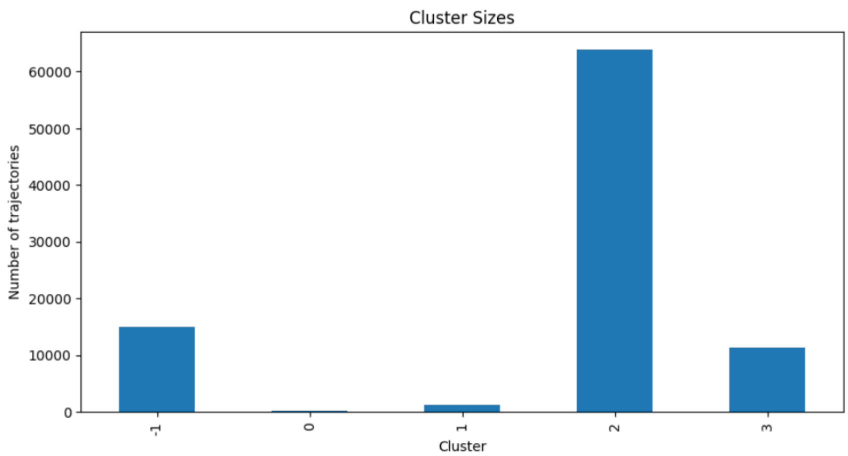


Figure 5.1: Distribution of vessel trajectory counts across the final HDBSCAN behavioural clusters and trajectories classified as noise points.

In order to further analyse the structure of the learned latent representation space, Uniform Manifold Approximation and Projection (UMAP) was applied to project the high-dimensional vessel trajectory embeddings into a two-dimensional representation space. The resulting projection provided a visual representation of the behavioural organisation learned by the self-supervised representation learning framework.
Figure 5.2 illustrates the UMAP projection of the learned vessel trajectory embeddings coloured according to the final HDBSCAN clustering assignments.

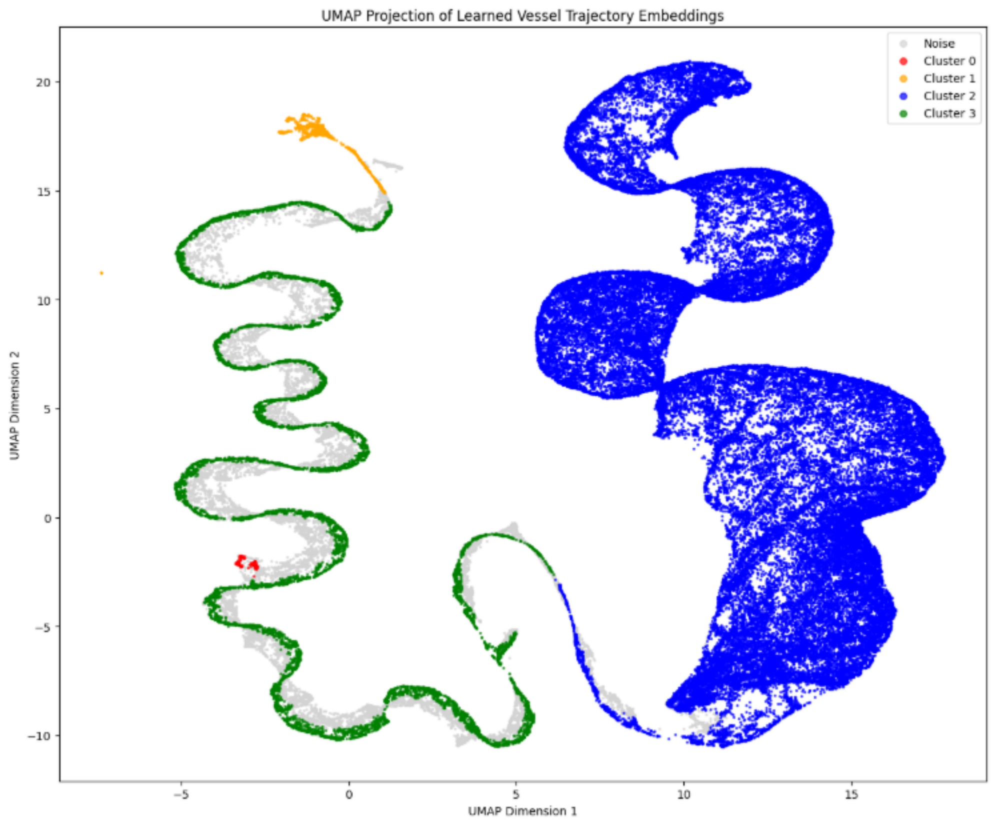


Figure 5.2: UMAP projection of the learned vessel trajectory embeddings coloured according to the final HDBSCAN clustering assignments.

The UMAP projection demonstrated that the learned embedding space exhibited behavioural organisation and density variation. The dominant behavioural clusters formed large coherent embedding manifolds corresponding to common navigational movement regimes, while smaller sparse clusters occupied isolated regions of the embedding space associated with behaviourally uncommon vessel trajectories.
In particular, Cluster 2 formed the largest and densest embedding structure within the latent space, representing the dominant vessel transit behaviour observed across the AIS dataset. Cluster 3 formed a secondary but clearly separated behavioural manifold characterised by more variable navigational movement patterns and increased directional variation. The smaller Clusters 0 and 1 occupied highly isolated embedding regions, indicating that the self-supervised learning framework separated sparse and specialised navigational movement patterns from the dominant vessel transit regimes.

The trajectories classified as noise by HDBSCAN were distributed primarily around sparse regions and transition boundaries between the dominant behavioural manifolds. This behaviour suggested that the noise trajectories represented vessel movement patterns exhibiting reduced behavioural similarity to the dominant navigational regimes identified within the learned latent embedding space.

Figure 5.3 presents representative vessel trajectories classified as noise points by the HDBSCAN clustering framework.

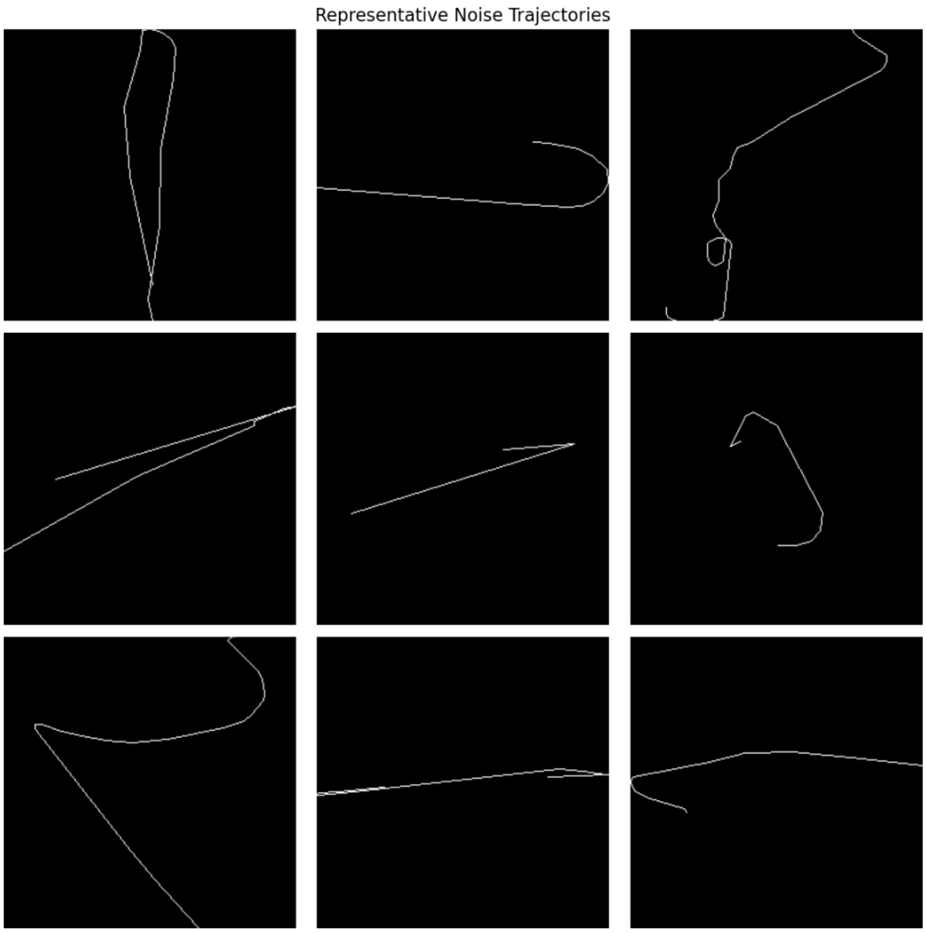


Figure 5.3: Representative vessel trajectories classified as noise points by the HDBSCAN clustering framework.

Visual inspection of representative trajectory samples further illustrated the behavioural separation observed within the embedding space. The dominant behavioural clusters primarily corresponded to stable transit navigation patterns characterised by continuous vessel movement and relatively smooth directional behaviour. In contrast, the smaller behavioural clusters frequently contained more specialised movement structures including compact trajectory fragments, local manoeuvring behaviour and sparse navigational movement patterns.

Figure 5.4 presents representative trajectory samples extracted from the principal HDBSCAN behavioural clusters.

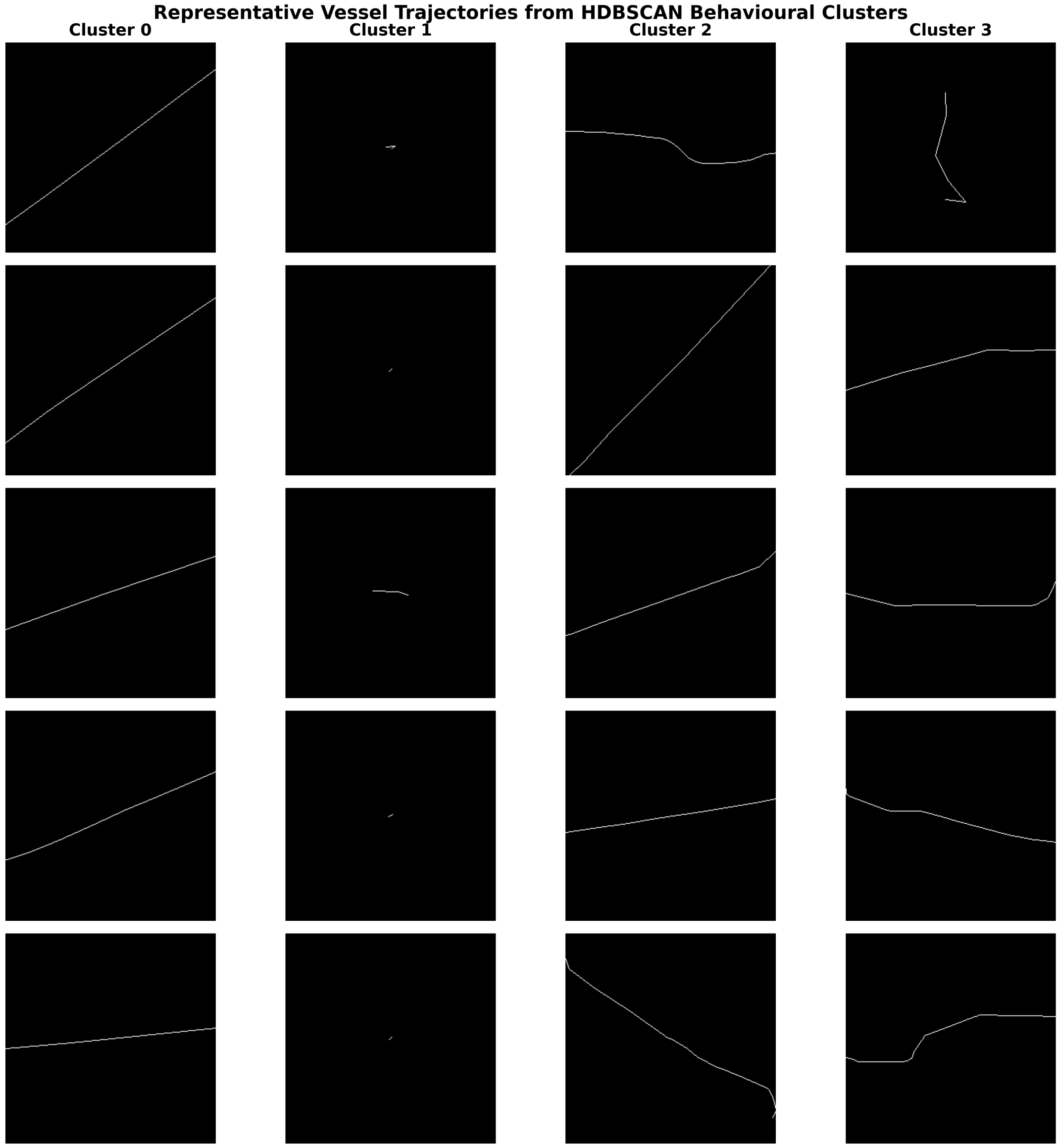


Figure 5.4: Representative trajectory samples extracted from the principal HDBSCAN behavioural clusters.

Additional inspection of the multi-channel trajectory tensors demonstrated that the learned behavioural representations were influenced not only by geometric trajectory structure but also by the engineered behavioural channels incorporated during preprocessing. While the primary trajectory channel encoded geometric vessel movement patterns, the additional drift, turning-angle and speed-variation channels introduced behavioural navigational characteristics enabling the representation learning framework to distinguish between operational movement regimes exhibiting otherwise visually similar geometric trajectories.

Figure 5.5 presents representative examples of the multi-channel vessel trajectory representations used during representation learning.

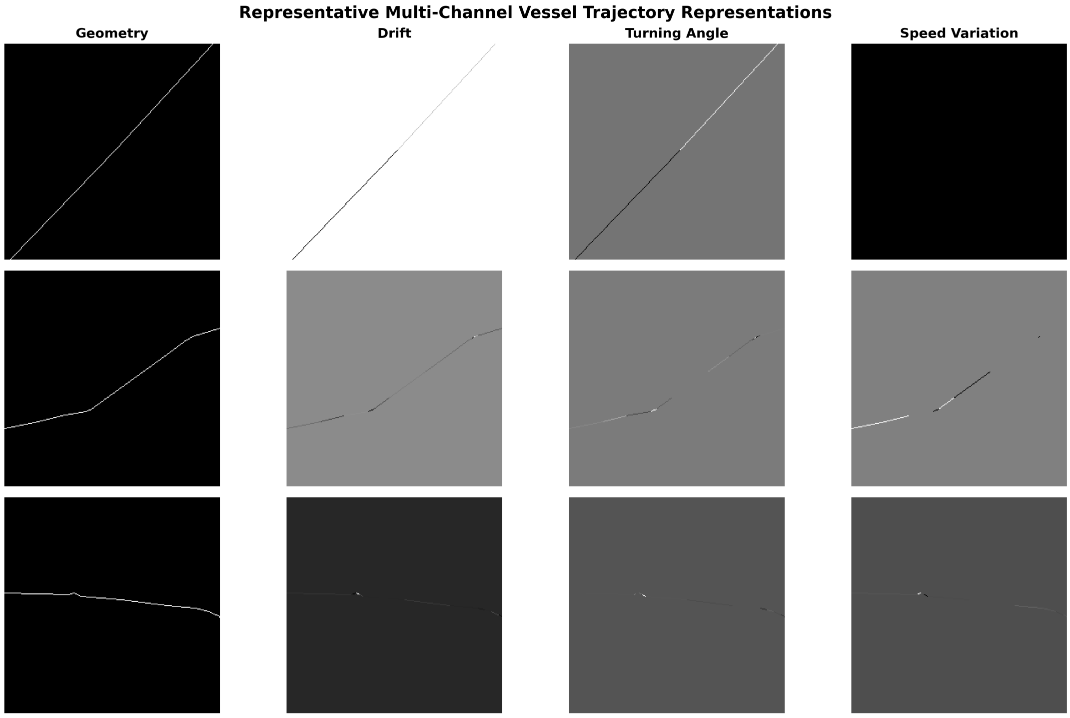


Figure 5.5: Representative examples of the four-channel vessel trajectory tensor representation used during self-supervised learning, including the geometric trajectory channel together with the engineered behavioural channels representing drift behaviour, turning-angle variation, and speed variation.

Overall, the final clustering results demonstrated that the proposed self-supervised representation learning framework was capable of organising large-scale AIS vessel trajectories into coherent behavioural movement regimes while simultaneously isolating sparse and behaviourally uncommon navigational patterns suitable for further operational interpretation and anomaly analysis.

5.3 Scenario-Based Washwater Estimation Results

This section presents the results of the scenario-based washwater estimation procedure described in Section 4.7. The analysis addresses Objective O6 and Research Question RQ4, and constitutes the secondary environmental strand of the thesis. The estimates reported below are derived from operational assumptions combined with vessel-specific metadata provided by Scanjet AB. They are not measurements and they are not presented as evidence that discharge occurred.

5.3.1 Selection of Candidate Trajectories

Candidate vessels were selected from the trajectories isolated during the clustering and behavioural analysis stage described in Section 5.2. Selection was restricted to vessels whose trajectories exhibited thesis-defined screening characteristics, namely prolonged low-speed movement, repeated manoeuvring, looping trajectory structures and deviations toward open sea. Vessels for which the corresponding operational washing-system metadata was available in the Scanjet dataset were retained, since vessel-specific flow capacity and machine configuration are required inputs to the estimation model.

Because the trajectory segmentation procedure partitions AIS records into fixed four-hour windows (Section 3.4), a single extended period of anomalous behaviour is distributed across several consecutive segments. For the purposes of this analysis, temporally adjacent anomalous segments belonging to the same vessel were therefore aggregated into a single candidate event, and the duration of that event was taken as the total elapsed time spanned by the aggregated segments. The resulting durations consequently exceed the four-hour segmentation window.

Four candidate events were carried forward to the estimation stage. Vessel identities are reported in partially redacted form, since the purpose of the analysis is methodological demonstration rather than the identification of individual operators.

5.3.2 Estimation Assumptions

The estimated discharge volume for each candidate event was obtained as

$$V = Q \times n \times t \tag{5.1}$$

where V denotes the estimated washwater volume (m^3), Q the flow capacity of a single cleaning machine (m^3/h), n the number of cleaning machines installed on the vessel and t the estimated duration of the candidate event (h).

Since neither the actual duty cycle nor the number of machines operating simultaneously can be determined from AIS data, each candidate event was evaluated under two bounding scenarios rather than a single point estimate. The lower bound corresponds to a single machine operating and the upper bound to continuous operation of all installed machines. The ratio between the upper and lower bounds is therefore determined by the vessel-specific number of installed cleaning machines used in the calculation rather than by properties of the observed trajectories.

5.3.3 Estimated Discharge Volumes

Table 5.4 presents the estimated washwater discharge volumes for the four candidate events.

Table 5.4: Scenario-based estimates of potential washwater discharge volumes for four candidate anomalous events. Vessel identifiers are partially redacted. Estimated volumes represent bounding operational scenarios and are not measured quantities.

Vessel (IMO)	Event duration	Estimated volume (m^3)
98xxxxx	10 h 08 min	117 – 2 797
97xxxxx	24 h 18 min	571 – 19 987
96xxxxx	52 h 58 min	847 – 45 763
94xxxxx	61 h 19 min	1 840 – 44 148

The estimated volumes vary considerably across the four candidate events, spanning approximately 117 m^3 at the lower bound of the shortest event to approximately 45 800 m^3 at the upper bound of the longest. Two factors drive this variation. The first is event duration, which ranges from roughly ten hours to over sixty hours and enters the estimate linearly. The second is vessel-specific washing capacity: the implied effective flow rates differ substantially between vessels, reflecting differences in the number of installed cleaning machines and their nominal capacities as recorded in the Scanjet dataset. Vessel 98xxxxx, for example, exhibits a markedly lower implied flow rate than vessels 96xxxxx and 97xxxxx despite a comparable estimation procedure, indicating that vessel configuration exerts an influence on the estimates of similar magnitude to event duration.
It is notable that the two longest events, 96xxxxx and 94xxxxx, yield a quite similar upper bounds in the vloume ranges despite a difference of more than eight hours in duration. This reflects the installed washing capacity of vessel 94xxxxx, which partially offsets its longer observed anomalous period. The estimates are therefore not reducible to duration alone, and the integration of vessel-specific operational metadata materially affects the outcome. This observation supports the premise underlying RQ4, namely that behavioural anomaly detection and operational metadata are more informative in combination than either is in isolation.

6

Discussion

6.1 Interpretation of the Learned Behavioural Representations

One of the most significant findings of this study is that the proposed self-supervised learning framework was capable of learning useful behavioural representations directly from unlabeled AIS trajectory data. The absence of labelled examples of tank-cleaning operations prevented the use of traditional supervised learning approaches, making representation learning a critical component of the overall methodology. The results indicate that the BYOL-based framework transformed complex vessel trajectories into compact latent embeddings that preserved behavioural information while enabling downstream clustering and anomaly analysis.

The quality of the learned representations can be assessed through the clustering behaviour observed throughout the experimental progression. Across multiple experiments, trajectories exhibiting similar movement characteristics formed groups within the latent embedding space, while unusual trajectories appeared in sparse regions or were classified as noise by HDBSCAN. The clustering results were assessed quantitatively using the Silhouette coefficient and Davies-Bouldin index, together with the proportion of trajectories classified as noise. The emergence of distinct behavioural groupings suggests that the model learned underlying vessel movement patterns rather than simply memorising trajectory geometries.

An important observation is that the learned representations appeared to capture operational behaviour in addition to spatial movement. The trajectory tensors provided to the network contained not only geometric trajectory information but also engineered behavioural channels describing drift behaviour, turning-angle variation and speed variation. Consequently, the latent embeddings reflected both where vessels moved and how they moved. This distinction is particularly important in maritime anomaly detection, where vessels operating in the same geographical region may exhibit different operational behaviours. The experimental comparisons showed that changing the behavioural feature configuration affected the resulting clustering structure, although the internal clustering metrics did not consistently favour the same feature configuration across all experiments.

The learned embeddings were also affected by changes in dataset scale and experimental configuration. Experiments conducted on different dataset scopes produced different clustering structures and internal validation results. In particular, the three-port DP configuration (E4) produced a Silhouette coefficient of 0.644 and a Davies–Bouldin index of 0.303, while the final full-dataset configuration (E8) produced a Silhouette coefficient of 0.754 and a Davies–Bouldin index of 0.498. These results demonstrate that the clustering structure changed as the dataset scope and other preprocessing configurations were modified. Visual inspection of the embedding projections and representative trajectories was therefore considered alongside the quantitative metrics when interpreting the learned representations.

At the same time, the interpretation of the learned representations remains subject to an important limitation inherent to deep learning methodologies. Although the clustering results and visual inspection indicate that behavioural information was captured within the latent space, the internal decision-making process of the neural network is not directly interpretable. The learned embeddings cannot be associated with specific physical vessel actions in a deterministic manner, and therefore behavioural interpretation relies primarily on cluster-level analysis and trajectory visualisation. Consequently, the results should be understood as evidence that the model learned useful behavioural abstractions from AIS data rather than explicit representations of individual operational activities.

Overall, the findings demonstrate that self-supervised representation learning provides a viable approach for extracting behavioural information from large-scale unlabeled AIS datasets. The ability of the framework to generate behavioural groupings without requiring manually labelled examples represents an important advantage for maritime anomaly detection problems, where labelled operational data are often unavailable. These results suggest that self-supervised learning can serve as a useful foundation for exploratory maritime behavioural analysis and may support future environmental monitoring applications involving rare or difficult-to-observe vessel activities.

6.2 Influence of Trajectory Compression

Trajectory compression constituted a critical preprocessing stage within the proposed framework because the compressed trajectories formed the basis for all subsequent rasterization, representation-learning and clustering procedures. The experimental results demonstrated that the choice of compression algorithm had a crucial influence on the resulting latent representations and clustering structures. Although both DP compression and TDKC successfully reduced trajectory complexity, the two methods produced markedly different outcomes during downstream representation learning and clustering.

The clearest difference between the two methods lies not only in the geometry of the retained trajectories but also in the distribution of compressed trajectory lengths that each method produces. Figures 3.2a and 3.2b present these distributions for the three-port dataset.

The DP algorithm produced a broad, heavy-tailed distribution. A large proportion of trajectory segments were reduced to the minimum representation of two points, while the remainder extended across a continuous range of lengths up to a maximum of 39 points. The retained trajectory length therefore varied with the geometric complexity of the underlying vessel movement: segments corresponding to straight-line transit were simplified aggressively, whereas segments containing curvature, manoeuvring or directional change retained proportionally more points.

The TDKC configuration employed in this thesis behaved very differently. Rather than producing a distribution of trajectory lengths, TDKC compressed almost every trajectory segment to exactly three points, with a maximum observed length of only seven points across the entire dataset. Given that the segmented trajectories entering the compression stage contained on the order of tens of AIS observations each, this indicates that the adaptive threshold mechanism terminated the recursive subdivision procedure after a single level of decomposition for the overwhelming majority of segments. The compressed output consequently consisted of a start point, a single interior point and an end point for most trajectories, irrespective of the complexity of the original vessel movement.

This behaviour is significant because it means that the TDKC configuration used here did not, in practice, function as an adaptive compression method. A two-hour transit across open water and a two-hour segment containing repeated manoeuvring were reduced to representations of identical size and comparable structural poverty. When such three-point polylines are rasterized into 256 × 256 image tensors, the resulting representations approximate a straight line with at most a single inflection, and the engineered behavioural channels are expressed across only a handful of pixels. The information available to the self-supervised encoder is therefore severely restricted, independently of any property of the kinematic criteria that TDKC was designed to exploit.

6.2.1 Consequences for representation learning and clustering

The downstream consequences of the different compression outputs were observed across the experiments in which both methods were evaluated. At the single-port stage, DP compression (E1) produced six behavioural clusters with 9.2% of trajectories classified as noise, a Silhouette coefficient of 0.789 and a Davies-Bouldin index of 0.331. TDKC compression (E2) produced 25 clusters with 41.4% noise, a Silhouette coefficient of 0.500 and a Davies–Bouldin index of 0.467. At the three-port stage, DP compression (E4) produced ten clusters with 31.2% noise, a Silhouette coefficient of 0.644 and a Davies-Bouldin index of 0.303, while TDKC compression (E5) produced 56 clusters with 38.2% noise, a Silhouette coefficient of 0.400 and a Davies–Bouldin index of 0.447. Thus, in both comparisons, the DP configuration produced higher Silhouette coefficients and lower Davies–Bouldin indices than the corresponding TDKC configuration. The DP configurations also produced fewer clusters and lower noise proportions, although these two measures were treated as descriptive characteristics rather than direct measures of clustering quality.

Inspection of the embedding projections and representative trajectory samples provided additional qualitative evidence for the differences between the compression methods. The DP-derived embeddings showed larger behavioural groupings, whereas the TDKC-derived embeddings showed a greater degree of clustering fragmentation. This visual assessment was consistent with the larger number of clusters observed under TDKC. The quantitative metrics further indicate that the DP configurations produced stronger internal clustering structure under the tested configurations, as reflected by their higher Silhouette coefficients and lower Davies-Bouldin indices.

Taken together, these observations suggest that the downstream differences between the DP and TDKC pipelines are associated with the amount and distribution of trajectory information retained during compression. This does not imply that TDKC is inherently unsuitable for trajectory compression. Rather, the specific TDKC configuration used in this study produced highly simplified representations that resulted in different latent representations and clustering structures after rasterization and self-supervised learning. A compression method that preserves kinematically relevant points may therefore still be unsuitable for the subsequent representation-learning pipeline if the resulting trajectory contains insufficient structural information for the encoder.

One qualification must be attached to the comparison presented above: the two pipelines did not cluster the same population of trajectories. Because DP reduces a large fraction of segments to two points and because trajectories containing fewer than three points after compression were excluded prior to rasterization (Section 3.7), the DP pipeline discarded a proportion of the segmented trajectories before clustering. At the three-port stage, 6,976 trajectories entered the DP clustering analysis (E4) as against 22,494 under TDKC (E5), despite both methods operating on an identical set of segmented inputs. The trajectories removed by this filter are those that DP reduced to fewer than three points and which were therefore excluded before clustering. The noise proportions reported for the two methods are therefore computed over different populations and are not directly comparable in a strict sense. The comparison of the Silhouette and Davies-Bouldin values should also be interpreted in the context of this difference in analysed populations. Nevertheless, the difference in clustering structure and the consistent advantage of DP in both internal validation metrics provide evidence that the tested DP configuration produced stronger clustering structure than the tested TDKC configuration.

6.2.2 Basis for the final pipeline configuration

On the basis of the evidence available, DP compression was adopted for all subsequent large-scale experiments and for the final anomaly-detection framework. In both the single-port and three-port comparisons, the DP configuration produced higher Silhouette coefficients and lower Davies-Bouldin indices than the corresponding TDKC configuration. The DP configurations also resulted in lower noise proportions, although the noise proportions should be interpreted with caution because the two compression methods did not retain the same trajectory population for clustering. Visual inspection of the embedding projections and representative trajectory samples further supported the selection of DP for the subsequent experiments.

TDKC was not carried forward to the full-dataset experiments, both because of its observed clustering behaviour at the smaller scales and because of the computational cost of the full-scale pipeline. The answer to RQ2 offered by this thesis is accordingly a qualified one: at the scales tested and under the configurations implemented, DP compression produced stronger internal clustering results and retained more variable trajectory representations than TDKC for the purposes of downstream representation learning and clustering. This conclusion refers specifically to the compression configurations implemented in this study and does not imply that DP is generally superior to TDKC for all trajectory-compression applications.

6.3 Influence of Behavioural Feature Engineering

The feature-ablation experiments demonstrated that the choice of behavioural features influenced the learned vessel representations and the resulting clustering structure. The relevant comparison is between experiments E6 and E7, which were performed on the identical set of 154,944 full-dataset trajectories using identical DP compression and identical segmentation, and which differed only in the behavioural channels supplied to the encoder.

When the full three-channel behavioural representation was employed, comprising speed variation, turning-angle variation and drift behaviour (E7), HDBSCAN identified 31 behavioural clusters with 36.3 % of trajectories classified as noise. The corresponding two-channel representation, in which drift behaviour was removed and only speed variation and turning-angle variation were retained (E6), produced 53 clusters with 68.7 % of trajectories classified as noise. The internal clustering metrics provided a more mixed comparison. E6 obtained a Silhouette coefficient of 0.143 and a Davies-Bouldin index of 0.600, while E7 obtained a Silhouette coefficient of 0.126 and a Davies-Bouldin index of 1.185. Thus, the two-channel representation achieved slightly better values for both internal validation metrics, whereas the three-channel representation resulted in fewer trajectories classified as noise. Visual inspection of the embedding projections and representative trajectories, however, favoured E7, which provided more interpretable behavioural groupings. The feature configuration was therefore assessed using both quantitative clustering metrics and qualitative trajectory inspection rather than relying on a single measure.

The three engineered channels describe complementary aspects of vessel movement. Speed variation captures changes in propulsion state and characterises acceleration and deceleration behaviour. Turning-angle variation captures directional manoeuvring and changes in course. Drift, computed from the relationship between vessel heading and course over ground, captures the extent to which a vessel's actual track diverges from the direction in which it is pointed, and therefore provides information about heading and navigational consistency. Vessels exhibiting similar speed and turning characteristics may nevertheless differ in their drift behaviour, particularly during slow movement or manoeuvring. The inclusion of drift therefore provided an additional behavioural dimension that may help distinguish movement patterns that are not fully described by speed and turning-angle variation alone. The visual inspection of representative trajectories was consistent with this interpretation.

6.3.1 The interaction between feature engineering and compression

An apparently contradictory result was obtained at the single-port stage under TDKC compression and it merits explicit discussion rather than omission. Comparing experiments E2 and E3, both performed on the identical set of 18,571 TDKC-compressed trajectories, the three-channel behavioural representation (E2) produced

25 clusters with 41.4 % noise, whereas the two-channel representation (E3) produced 28 clusters with a lower noise proportion of 36.8 %. The internal validation metrics again showed a mixed result. E2 obtained a Silhouette coefficient of 0.500 and a Davies-Bouldin index of 0.467, while E3 obtained a Silhouette coefficient of 0.297 and a Davies-Bouldin index of 0.413. Consequently, removing drift reduced the Silhouette score but slightly improved the Davies–Bouldin index and reduced the proportion of trajectories classified as noise. This differs from the effect observed under DP compression and indicates that the contribution of the additional feature was dependent on the overall preprocessing configuration.

This result is not, however, inconsistent with the interpretation advanced above. It can be considered in relation to the compression behaviour documented in Section 6.2. The engineered behavioural features are supplied to the encoder as additional raster channels, in which feature values are represented along the compressed trajectory path. Their representation therefore depends on the number of trajectory points that survive compression. Because the TDKC configuration reduced almost every trajectory to three points, the behavioural channels were expressed across only a small number of spatial locations. Under these conditions, the additional drift channel may have contributed limited additional information to the rasterized representation. Under DP compression, by contrast, trajectories retained a more variable number of points and the behavioural channels could be represented along a longer spatial path. This provides a possible explanation for why the effect of feature removal differed between the two compression methods.

The two principal findings of this thesis regarding preprocessing are therefore not independent. Behavioural feature engineering and trajectory compression interact: the benefit conferred by additional behavioural channels is conditional upon the compression scheme retaining enough trajectory points for those channels to be meaningfully expressed in the rasterized representation. A compression configuration that reduces trajectories to a near-minimal point count effectively nullifies the contribution of behavioural feature engineering, irrespective of how informative those features may be in principle. This observation has practical implications for the design of image-based maritime trajectory pipelines: compression thresholds and behavioural feature sets should be selected jointly rather than tuned in isolation, since the informational value of the latter is bounded by the former.

6.4 Limitations

Although the proposed framework demonstrated promising results for AIS-based behavioural analysis and anomaly detection, several limitations must be considered when interpreting the findings. These limitations relate primarily to the absence of verified tank-cleaning observations, the inherent constraints of AIS data, the interpretability of the learned representations, and the exploratory nature of the washwater estimation framework.

The most significant limitation of this study is the absence of ground-truth data describing confirmed tank-cleaning operations or verified washwater discharge events. While the proposed methodology successfully identified anomalous vessel behaviours within large-scale AIS datasets, it was not possible to establish a direct relationship between the detected anomalies and actual tank-cleaning activities. Consequently, the identified trajectories should not be interpreted as confirmed cleaning events but rather as behavioural outliers that may warrant further investigation. The lack of labelled operational data also prevented quantitative validation of the anomaly-detection performance and limited the ability to evaluate detection accuracy using conventional classification metrics.

A second limitation arises from the use of AIS data as the primary information source. Although AIS provides continuous information regarding vessel position, speed, heading and navigational behaviour, it does not contain direct information concerning onboard operational activities. Important factors such as cargo type, tank-cleaning procedures, discharge operations, weather conditions and crew decision-making processes are not represented within AIS transmissions. As a result, different operational activities may produce similar movement patterns, while unusual vessel behaviour may occur for reasons unrelated to tank cleaning. The behavioural anomalies identified by the framework should therefore be interpreted within the broader operational context of maritime navigation.

Additional limitations are associated with the self-supervised learning methodology employed in this thesis. While the BYOL framework successfully generated meaningful latent behavioural representations and enabled effective clustering of vessel trajectories, the learned embeddings remain difficult to interpret directly. Similar to many deep learning approaches, the internal representation space does not provide explicit explanations regarding which trajectory characteristics contributed most strongly to cluster formation or anomaly identification. Consequently, behavioural interpretation relied primarily on downstream trajectory visualisation and qualitative cluster analysis rather than direct examination of the learned feature space.

Finally, the washwater discharge estimation framework developed in this thesis should be regarded as exploratory rather than predictive. The estimated discharge volumes were derived from hypothetical operational scenarios combined with supplementary vessel metadata and were not validated using environmental measurements, onboard operational records, discharge sensors, or confirmed cleaning-event observations. The resulting estimates therefore represent plausible operational ranges rather than empirically verified discharge quantities. Despite these limitations, the framework demonstrates how AIS-based behavioural analysis can be combined with supplementary operational information to support future environmental monitoring applications.

7
Conclusion

This thesis investigated the potential of AIS-based machine learning methods for identifying vessel behaviours that may be associated with tank-cleaning activities in chemical tankers. Motivated by the limited visibility of tank-cleaning operations and the challenges associated with monitoring washwater discharge activities, a complete anomaly-detection framework was developed using AIS trajectory data, behavioural feature engineering, self-supervised learning and unsupervised clustering techniques.

To address RQ1, several behavioural trajectory features were extracted from AIS data, including turning-angle variation, speed variation and drift behaviour. The experimental results demonstrated that behavioural feature engineering played a critical role in the quality of the learned trajectory representations. In particular, the inclusion of drift behaviour significantly improved cluster coherence and reduced the proportion of trajectories classified as noise. The feature-ablation experiments showed that removing drift behaviour resulted in fragmentation of the latent space and a marked deterioration in clustering performance. These findings suggest that drift behaviour provides important information regarding navigational consistency and complements speed and turning-angle characteristics when describing vessel operations.

To address RQ2, the performance of DP and TDKC trajectory compression methods was evaluated. Although TDKC was designed to preserve kinematic trajectory characteristics, the results showed that DP compression consistently produced more coherent latent representations and more interpretable clustering structures. TDKC generated fragmented embedding distributions, a larger number of small clusters, and a higher proportion of noise trajectories during HDBSCAN analysis. The results therefore indicate that preserving trajectory geometry alone does not necessarily lead to improved behavioural representation learning and that DP compression was better suited to the proposed machine-learning framework.

To address RQ3, a BYOL-based self-supervised learning framework was developed to learn behavioural vessel representations without requiring labelled training data. The resulting latent embeddings enabled HDBSCAN to identify distinct behavioural trajectory groups and isolate anomalous vessel movements within large-scale AIS datasets. The results demonstrated that meaningful behavioural representations can be learned directly from unlabeled AIS data and that self-supervised learning provides an effective approach for maritime anomaly detection problems where verified operational labels are unavailable. Visual inspection of the identified anomalies revealed vessel movement patterns characterised by irregular manoeuvring, looping behaviour and atypical navigational structures that differed substantially from the dominant behavioural groups.

To address RQ4, the anomalous trajectories identified by the proposed framework were combined with supplementary operational washing-performance data provided by Scanjet AB in order to estimate potential washwater discharge volumes under a set of hypothetical operational scenarios. Although the resulting estimates could not be validated against direct environmental measurements or confirmed tank-cleaning events, the analysis demonstrated that AIS-derived behavioural anomalies can be integrated with vessel-specific operational metadata to support exploratory assessments of potential environmental impacts. The results indicate that anomaly detection alone provides limited environmental context, whereas the combination of behavioural analysis and operational washing information enables a more meaningful interpretation of anomalous vessel activity. Consequently, the proposed framework illustrates how AIS-based anomaly detection can serve as a foundation for future environmental monitoring systems capable of linking unusual vessel behaviour with operational estimates of washwater generation and discharge.
Overall, the findings of this thesis demonstrate that AIS-based self-supervised learning can provide a practical framework for large-scale behavioural analysis of chemical tanker operations. While the methodology does not directly identify confirmed tank-cleaning events, it successfully detects anomalous vessel behaviours that may warrant further investigation and therefore provides a promising foundation for future maritime environmental monitoring applications. By combining behavioural feature engineering, trajectory compression, representation learning, and density-based clustering within a unified analytical framework, this thesis contributes to the growing use of artificial intelligence and data-driven methods for maritime environmental assessment and anomaly detection.

Bibliography

[1] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231. AAAI Press, 1996.

[2] European Maritime Safety Agency. Cleanseanet service user manual, 2021.

[3] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Shaoqing Guo, Victor Bolbot, and Osiris Valdez Banda. An adaptive trajectory compression and feature preservation method for maritime traffic analysis. *Ocean Engineering*, 312:119189, 2024. doi: 10.1016/j.oceaneng.2024.119189.

[5] HELCOM. Ais and e-navigation, 2024. URL `https://helcom.fi/action-areas/shipping/ais-and-e-navigation/`. Accessed 2026-05-17.

[6] International Maritime Organization. Resolution a.1106(29): Revised guidelines for the onboard operational use of shipborne automatic identification systems (ais), 2015.

[7] International Maritime Organization. Marpol annex ii: Regulations for the control of pollution by noxious liquid substances in bulk, 2024. Consolidated version as of 1 May 2024.

[8] Jukka-Pekka Jalkanen et al. Ais coverage and data quality in the baltic sea region, 2021.

[9] Anna Lunde Hermansson and Ida-Maja Hassellöv. Tank cleaning and its impact on the marine environment, 2022.

[10] Wengang Mao, Shaobo Wang, and Ida-Maja Hassellöv. Pre-study on ais “big data” analyses for identification of tank cleaning operations at sea, 2023.

[11] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. doi: 10.21105/joss.00205.

[12] Giuliana Pallotta, Michele Vespe, and Karna Bryan. Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction. *Entropy*, 15(6):2218–2245, 2013. doi: 10.3390/e15062218. URL https://www.mdpi.com/1099-4300/15/6/2218.

[13] Antonio Augusto Sepp Neves, Nadia Pinardi, and Flavio Martins. It-osra: applying ensemble simulations to estimate the oil spill risk associated to operational and accidental oil spills. *Ocean Dynamics*, 66(8):939–954, 2016. doi: 10.1007/s10236-016-0960-0.

[14] United Nations Conference on Trade and Development. Review of maritime transport 2023, 2023.

[15] Wanqing Wu, Zhixing Cao, Bin Zhang, Yafei Guo, Min Du, Qinggong Zheng, Zhaoao Bai, and Zhongxin Shi. Safe and efficient tank cleaning of liquid cargo tankers: a review. *Ships and Offshore Structures*, 2023. doi: 10.1080/17445302.2022.2162752.

[16] Yang Zhou, Winnie Daamen, Tiedo Vellinga, and Serge P. Hoogendoorn. Impacts of wind and current on ship behavior in ports and waterways: A quantitative analysis based on AIS data. *Ocean Engineering*, 213:107774, 2020. doi: 10.1016/j.oceaneng.2020.107774. URL https://doi.org/10.1016/j.oceaneng.2020.107774.

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

AVANCEZ
1829

CHALMERS
UNIVERSITY OF TECHNOLOGY