Degassing in Inland Shipping

An Exploratory Study on Identifying Degassing Patterns and Hotspots on Dutch Inland Waterways Based on AIS and IVS data

Master Thesis
MSc. Geographical Information Management and Applications (GIMA)

Bram van der Pas
b.t.m.vanderpas@students.uu.nl

Supervisor: Dr.ir. B.M. Meijers
Responsible professor: Prof.dr.ir. P.J.M. van Oosterom
February 2025

Degassing in Inland Shipping

An Exploratory Study on Identifying Degassing Patterns and Hotspots on Dutch Inland Waterways Based on AIS and IVS data

Bram van der Pas

21-02-2025

9778128

Master thesis
MSc. Geographical Information Management and Application (GIMA)

Delft University of Technology (TU Delft)

With support of:
Dr.ir. B.M. Meijers - TU Delft
Ir. B. Turpijn- Rijkswaterstaat
Drs. M.C.M Schuur - ILT

Preface

This work is a thesis submitted in fulfilment of the requirements of the MSc. Programme Geographical Information Management & Applications (GIMA).

The findings of this research are not as detailed or conclusive as I had initially hoped. However, they provide valuable insights and lay the groundwork for further exploration of this topic. When the initial aim of the research was determined, I had assumed that the requested data could be provided and the process of data delivery was already well established. But, due to the first time of delivering such a dataset and the circumstances at the time of the research, only part of the requested data could be delivered and the delivery was only two weeks before the deadline. Nevertheless, a number of valuable insights emerged from the research.

I want to sincerely thank my supervisor, Martijn Meijers, for his advice and guidance through this dynamic process. Further, I would also like to thank Peter van Oosterom from the TU Delft for taking the role of responsible professor and providing feedback on the research. From the Inspectie Leefomgeving en Transport (ILT) I would like to thank Margje Schuur and Rony Nedkov for helping to supervise the research and providing connections and information from the ILT. From Rijkswaterstaat I would like to thank Bas Turpijn and Fedor Baart for helping to supervise the research, providing information and helping with the data provision process at Rijkswaterstaat.

At last, I want to thank my friends and family, especially my brother Mark, aunt and uncle Geertje and Joost and my friend Luuk for providing feedback on my research report and lending me some pieces of literature. I would also like to thank you, my reader: I hope you enjoy your reading.

Bram van der Pas
Heeswijk-Dinther, February 21, 2025

Abstract

Vessels transporting liquid gasses still contain some vapour after discharging their cargo. These vessels need to degas their cargo tanks of this vapour in order to not contaminate their next cargo. Most of the vessels release these vapours unprocessed into the atmosphere. This uncontrolled degassing of vapours, especially volatile organic compounds (VOCs), from inland tanker vessels can pose a serious risk for the environment and human health. To combat the uncontrolled degassing of vessels, the Convention on the collection, Deposit, and reception of waste generated during Navigation on the Rhine and other Inland waterways (CDNI), which is also ratified by the Netherlands, introduced a phased ban on the uncontrolled degassing of multiple substances on 1 October 2024. However, the enforcement of this ban proved to be a challenge due to the lack of effective detection methods. This study explores the potential of using Automatic Identification System (AIS) and Informatie en volgsysteem Scheepvaart (IVS) data to identify degassing patterns and hotspots on Dutch inland waterways.

A methodology is developed to detect deviations in tanker vessel movement by comparing the actual taken routes with the optimal route between the start and destination point. After filtering out explainable deviating behaviour, a density estimation technique is applied on the data to identify potential degassing hotspots and an emission volume analysis is conducted to estimate the emission created by degassing. The results indicate that locations with certain characteristics are preferred for degassing activities. However, this methodology relies on multiple assumptions, and the current validation using a confusion matrix indicate a precision of 63.8% (based on 160 predicted classifications) and a recall of only 42.1% (based on 242 actual cases). The accuracy of the methodology is indicated on 98.2%, but this is heavily skewed by the large amount of true-negative results (10.583) .

Despite the many limitations that emerged during the research, the process provides valuable insights on the strengths and constraints of AIS and IVS data for detecting degassing activity. The findings can support the Inspectie Leefomgeving en Transport (ILT) and Rijkswaterstaat in further refining detection methods and enforcement strategies and showed the potential of the current proposed method. Future work should focus on applying the method on bigger datasets, increasing the accuracy of the method and a better validation of the results.

Table of Contents

Glossary

Term	Definition
Berth location	A place for a vessel to moor
Bunker station	A fuel station for inland shipping vessels
CDNI	Convention on the collection, Deposit, and reception of waste generated during Navigation on the Rhine and other Inland waterways
Dedicated transport	Transport in which a barge is loaded with the same kind of cargo, eliminating the need for degassing
Degassing	The removal of residual vapours from the tanks of a vessel
Deviating behaviour	Behaviour that deviates from the standard or expected (optimal) behaviour
Hotspots	A place of significant activity
ILT	Supervisor of the Ministry of Infrastructure and Water Management
Rijkswaterstaat	Executive agency of the Ministry of Infrastructure and Water Management
Tanker	Vessel designed to transport or store liquids or gases in bulk
Trip	Movement of a vessel between loading, discharging, berth or anchor location
Unexplainable deviating behaviour	Deviating behaviour that does not include explainable behaviour like bunkering or mooring
UN location code	The United Nations Code for Trade and Transport Locations
UN number	A four-digit number used to identify dangerous goods for transport
Vessel	Any vehicle designed for travel across or through water bodies

1 Introduction

In 1996, the Netherlands, together with Germany, Belgium, France, Luxembourg, and Switzerland, signed the Convention on the collection, Deposit, and reception of waste generated during Navigation on the Rhine and other Inland waterways (CDNI) (CDNI, 2019, 2023). Among the various restrictions on environmental pollution from inland shipping, the CDNI has agreed on a phased ban on the floating (uncontrolled) degassing of ships (CDNI, 2023). By signing the convention, the countries agreed to enforce the established rules. The convention is transposed into the national law of the contracting countries through their legal instruments (CDNI, n.d.). Degassing is needed when a tanker carrying liquid Volatile Organic Compounds (VOCs) has discharged its cargo and wants to load a new liquid due to safety, quality, and/or permit conditions (Buck et al., 2013; Erol & Arcuri, 2023; VNCI et al., 2024). Part of the old liquid remains in the tanks as vapours and can contaminate the subsequent cargo (Erol & Arcuri, 2023; Koop, 2016). Besides, these vapours are highly active and inflammable, so a sailing ship still carrying these vapours can constitute a serious threat to other users of the waterway and the environment (Erol & Arcuri, 2023). Erol & Arcuri (2023) even state that such a ship can be considered a 'floating bomb'. When a vessel is degassed, it is 'cleaned' and can load a wide variety of subsequent cargo (Buck et al., 2013).

Degassing can be divided into controlled and uncontrolled, where controlled degassing means that the vapours are being sucked from the ship and get treated at a facility. Uncontrolled degassing results in the vapours being ventilated into the atmosphere without treatment (Koop, 2016). The uncontrolled degassing of vessels results in the emission of the VOC and a small loss of the volume of the cargo (Sigrid & Leisner, 2021), but it can be done whilst the ship is sailing. The main reason for uncontrolled degassing is cost, not only by saving the expenses related to degassing at a degassing station but also the time saved by not having to wait the time it takes to degas, a possible queue at the station, and the time and fuel needed to sail to the degassing station (Geerlings & Kuipers, 2019).

Figure 1: Inland tanker ship. Source: vlootschouw.nl

In the CDNI, the six countries ratified the (phased) ban on the uncontrolled degassing of 18 dangerous substances (VOCs) and 2 miscellaneous dangerous substances classes in 2017 (CDNI, 2023; United Nations, 2011). The ban has gone into effect on the 1st of October 2024, but the Netherlands has anticipated this deadline by already banning the most harmful vapours (phases 1 and 2 of CDNI) on the 1st of July 2024 (CDNI, 2024; Erol & Arcuri, 2023; Ministry of Infrastructure and Water Management, 2023; VNCI et al., 2024). The degassing problem is especially relevant for the Netherlands as it has the highest concentration of tanker barges on inland waterways in all of Western Europe (Erol & Arcuri, 2023). The amount of tanker barges on inland waterways in the Netherlands is more than half of all tanker barges in the whole CDNI geographical scope (Erol & Arcuri, 2023; Koop, 2016). The Netherlands has banned the uncontrolled degassing of petrol since 2006 and the uncontrolled degassing of benzene is already banned in several provinces (Erol & Arcuri, 2023). Besides, Belgium and Germany have already introduced a ban on the uncontrolled degassing of several dangerous substances (Erol & Arcuri, 2023). However, the CDNI (2024) raises the following point on its website: “Local bans on venting are not sufficiently effective and create a risk of ‘waste tourism’”.

Since the start of 2025, investments have been made by the Dutch government to establish an effective enforcement system and create an adequate infrastructure for degassing (VNCI et al., 2024). The Inspectie Leefomgeving en Transport (ILT) and Rijkswaterstaat (RWS), both part of the Dutch government and clients of the research, want to gain insight into the most common degassing locations. This research tries to get more insight into degassing behaviours to enforce and prevent the problem. Besides, the ILT wants an indication of the amount of volume that is degassed into the atmosphere, to get a feel for the significance of this problem in the Netherlands. So, the identified degassing behaviour over a yearly period will be used to estimate the amount of volume of degassed cargo. The ILT is interested in the enforcement of the ban and wants to use the results of this report to increase the effectiveness of its monitoring. Rijkswaterstaat can use the results of this research to define new places for degassing stations in the most effective locations. Besides, other government or private organisations connected to this subject can use the outcomes of this research to get more insight or to check if the current known information is still up to date. This includes organizations like the provinces, environmental services, port authorities, the Ministry of Infrastructure and Water Management (IenW), Koninklijke Binnenvaart Nederland (KBN) and tank- and ship organizations (VOTOB, CTGG, Vemobin and VNCI) (Stuurgroep Varend Ontgassen, 2023).

1.1 Problem Statement

Currently, there is no straightforward method to detect degassing. The only method is to check sailing ships manually by boarding them or using a drone to identify if a vessel is ventilating. There are e-noses installed at some waterways which can alert the ILT, but correct signalling is difficult and degassing can only be confirmed by an inspection on board of the suspected vessel. The reason there is no effective method to detect degassing is mostly due to the recent introduction of the degassing ban, which results in the monitoring method still being in its early stages, but also due to the legal restrictions around privacy. As some skippers live on their vessel, the obtained AIS data is considered personal information and thus needs to follow the Dutch AVG (Algemene Verordening Gegevensbescherming) law and the General Data Protection Regulation (GDPR) of the EU (Autoriteit Persoonsgegevens, n.d.; European Commission, n.d.). The main priority of this research is to explore if IVS and AIS data can be used to detect degassing patterns within these legal restrictions.

Ships sailing on Dutch inland waterways occasionally deviate from normal behaviour between the start location and destination, like taking a detour or sailing between berths before loading at the next harbour. Whether or not behaviour is normal is very difficult to identify, especially after a vessel discharged its cargo. This is due to the different activities a vessel and its crew can perform after discharging the cargo, like resting or waiting on the next trip at a berth location, refuelling at a bunker station or mooring the vessel, e.g. to do some shopping.

Since degassing is a time-consuming process, longer deviating activities are very likely to be linked to the uncontrolled degassing of the vessel. As AIS data is a tool for maritime authorities to track and monitor vessel movement, it should be possible to detect these deviating activities of vessels with the use of historic AIS data. The IVS data can be used to validate the AIS data, as it contains the start and end locations and the timestamps of when a vessel reaches an object (locks, bridges or traffic posts) of all the trips made in or through the Netherlands. Besides, the IVS data has valuable information for the second part of the research about an estimation of the type and volume of vapour that is degassed, as it contains information on cargo type and amount.

1.2 Research Objectives/Questions

The problem statement results in the following research questions:

> *What sections of the Dutch inland waterway can be identified as 'hotspots' for illegal degassing activities based on historic IVS and AIS data, and what are the estimated volumes of gas emitted with these activities?*

This will lead to multiple sub-questions, where each sub-question will address a different aspect of the main research question. The first sub-question is part of the preprocessing of the data. The AIS data consists of a long continuous list of data points for every vessel and is not split into separate trips. Currently, there is no standardized way to split the AIS data into separate trips. The first sub-question will explore the different methodologies that are possible to split the AIS data.

> *SQ1: Which processing steps can be applied to segment continuous AIS data into separate trips based on temporal, spatial, or voyage-related characteristics?*

By answering the first sub-question, the data can be split into separate trips for each vessel. These trips will be compared to the optimal routes to extract deviating behaviour. These deviating trips will be analysed to check if the behaviour is explainable, all the unexplainable behaviour could be related to degassing. This will result in the answer to the second sub-question:

> *SQ2: What deviating patterns can be related to uncontrolled degassing?*

The results of the second sub-question will be used in a density estimation to create a heat map of the deviating behaviour related to degassing. This will result in the identification of degassing hotspots which gives an insight into where degassing is happening in the Netherlands and answers the third sub-question:

> *SQ3: Where are these unexplainable deviating patterns located?*

The second part of the research question is about the estimated quantity of cargo emitted as vapour due to degassing. Besides the locations of the degassing, the ILT also requested an estimation of the amount of cargo that is released into the atmosphere, to indicate the significance of this problem. This question will be addressed in the fourth sub-question:

> *SQ4: What is the total volume of cargo released into the atmosphere in a year due to degassing vessels?*

The last sub-question is about the validity and accuracy of the results. This includes questions like: how logical are the results of the research? And can true-positive, true-negative, false-positive and false-negative results be identified?

> *SQ5: What is the quality and reliability of the proposed method to identified degassing behaviour of vessels?*

The results of these five sub-questions combined should be able to answer the main research question. This research, however, will not result in legal advice on degassing of ships. Neither will it present concrete locations for degassing stations. It will merely create a map containing the routes displaying deviating behaviour, which could be related to degassing activities, based on the degassing patterns that can be detected. The research is a first exploration into the possibility of using AIS and IVS data to detect degassing. The proposed process of this research can however be used as a proof of concept. Besides, the results of this research can be used as input to identify new monitoring locations or as one of the factors in the decision of new locations for degassing stations. The research is only aimed at the Dutch waterways and the movement of tankers starting or ending in inland harbours. Deviating behaviour outside of the Dutch inland waterways will not be included in this research as the AIS data only covers the inland waterways of the Netherlands. The proposed method could also be applied in other countries, if a dataset similar to IVS is available in that country.

1.3 Research Methodology

This study employs a quantitative research approach to systematically answer the proposed research questions. It adopts an exploratory, experiment-based approach to investigate the possibility to identify degassing patterns and hotspots on Dutch inland waterways based on AIS and IVS data. The methodology is designed to ensure objective, reproducible and statistical results. The research follows a structured process, including preprocessing, analysis and validation. The data used in this research is gathered from Rijkswaterstaat or is openly available. The provided data by Rijkswaterstaat is privacy sensitive, thus this data is pseudonymized. On this data, preprocessing techniques such as joining and filtering are applied to ensure data quality. On the pre-processed data the analysis is performed to extract results and insights. Finally, validation is performed using a confusion matrix to assess the performance and reliability of the results.

1.4 Report Structure

The following chapter, Chapter 2, gives some theoretical background to the concepts and theories that are used in this research in Chapter 2. The theoretical framework will discuss AIS and IVS data, the shortest path algorithm, density estimation, VOCs and their emission. Secondly, in Chapter 3, the used methods and the important choices that were made during the research are explained in more detail in the methodology. Subsequently, the results of the research are presented in Chapter 4. The results answer the proposed sub-questions, where each sub-question is addressed in a separate sub-chapter. The results are followed by a discussion on the meaning, importance and relevance of the results. This chapter also includes recommendations for further research. Lastly, Chapter 6 presents the conclusion of the research, where the findings from the sub-questions are combined to address the main research question.

2 Theoretical framework

The theoretical framework will elaborate on concepts and theories that are used in the research to help explain the problem mentioned in the research question. These concepts and theories are explained based on literature and provide background information on the problem that is researched and the methodology that is used. The chapter ends with a conceptual framework, which gives an indication of the different variables in this research and how they are connected.

2.1 Volatile Organic Compounds

Volatile organic compounds (VOCs) is an umbrella term for organic chemicals that evaporate at room temperature and under normal pressure (David & Niculescu, 2021). Emissions of VOCs have a variety of direct and indirect impacts on human health and the environment. As David & Niculescu (2021) state: “Some of the volatile organic compounds are more volatile than others, those that evaporate faster are more dangerous and pose a higher risk to the environment and humans.” (p. 2). While most of the VOCs do not directly increase the concentration of greenhouse gases (as they have a short atmospheric lifetime and are decomposed), they all directly contribute to global warming by changing the concentration of ozone (David & Niculescu, 2021; Erol & Arcuri, 2023). Ozone is produced in a photochemistry process between VOCs, nitrogen oxides, and light, and is considered a strong greenhouse gas (Erol & Arcuri, 2023). Ozone, and other chemicals created by the photochemistry process, can also severely affect human health (David & Niculescu, 2021). Besides the indirect effects on human health, high concentrations of volatile organic compounds may also directly affect human health. Prolonged exposure to high concentrations of VOCs may damage the nervous system (Organic Psycho-syndrome) and some of the VOCs have carcinogenic or mutagenic properties (Rijkswaterstaat Environment, n.d.; RIVM, n.d.). Not all VOCs are equally dangerous, some VOCs are safe or hardly dangerous for human health, whilst a VOC like benzene is carcinogenic (RIVM, n.d.).

2.2 Automatic Identification System and Informatie en Volgsysteem Scheepvaart

Automatic Identification System (AIS) data makes it possible to track and monitor vessels but was originally introduced to identify vessels in maritime navigation (Emmens et al., 2021). AIS data contains various information mainly related to the current status of a vessel (Table 1). AIS data is sent every 10 seconds on the inland waterways in the Netherlands and, in the present day, AIS data is used for various purposes including the protection of the environment, management of vessels in waterways, and the overall surveillance of the vessels. There are however several limitations to AIS data, including the data quality (Emmens et al., 2021; Haskins et al., 2024). First of all, data quality issues include the noise present in the data. Noise can be defined as meaningless data related to uncertainty, precision, or corruption of the data. AIS data can contain noise in static, dynamic and voyage related information. Examples are wrongly communicated information (like speed over ground, timestamps and position), there may be duplicated data or some data could be missing (Emmens et al., 2021). Secondly, the paths of vessels can look illogical or go overland. This is mainly due to extended intervals between two data points, but can also be due to external conditions like traffic density, weather conditions, or high buildings. Lastly, data quality is influenced by human input of static or voyage-related data which can lead to intentional or unintentional inaccuracies in AIS data (Emmens et al., 2021). Emmens et al. (2021) therefore recommended to always use the AIS data in combination with other data sources. When combined, the AIS data can be used for environmental impact reduction and the protection of the environment (Emmens et al., 2021). In this research, the AIS data will be combined with the Informatie en Volgsysteem Scheepvaart (IVS) data.

IVS data is used to control the locks and bridges on the Dutch waterway but also contains the data of a vessel that is registered in the Binnenvaart Informatie en Communicatie Systeem (BICS) before it sails off (Rijkswaterstaat, n.d.). Other (standard) values that are included in the IVS data, like ship type and ship size, are stored in the Casco database, which makes this data standardized and reliable. Amongst others, the IVS data contains information like the start and end location, start time (the timestamp of when the ship arrived at an object), the type of ship, and the type and amount of cargo (Table 2) (Rijkswaterstaat, 2024a). As the IVS dataset also relies on human input, it will not be completely error-free. However, when dangerous substances are involved, the data is expected to be highly accurate. Therefore, it is likely that the most important data for this research is largely complete. These two datasets combined should give a clear picture of the vessel movements in Dutch inland waterways. However, there is some contradiction between the AIS and IVS data. This will be further explored in the following paragraph.

Table 1: Most relevant variables of a pseudonymized AIS dataset

Variable	*Description*
Static data	
Track id	A unique pseudonymized identification number of a vessel
Id	A unique pseudonymized identification number of a send signal
Message type	Type of message that is sent
Object type	The type of object the vessel passes
Ship Type AIS	The type of vessel in categorical numbers
Type (geometry)	The type of geometry present in the data
Dynamic data	
Timestamp	Timestamp of the send signal
Latitude	Geographical position/coordinate of a vessel (north-south)
Longitude	Geographical position/coordinate of a vessel (east-west)
Speed over ground (SOG)	Speed of vessel relative to the ground
Course over ground (COG)	Position relative to North
Heading	Direction (between 0° and 360°)
Geometry	Geometry (point) of the signal
Voyage data	
Length	Length of the vessel
Beam	The width of a vessel at its widest point
Draught	Draft/draught of the vessel
Hazard	The type of hazard in numerical format
Status	Status of the vessel (e.g. 'at anchor')

Table 2: Variables of an anonymized IVS dataset

Variable	*Description*
Jaarmaand	4-digit code containing the year and month (yymm)
Jaar	The numeric year the data was recorded
Maand	The numeric month the data was recorded
Weeknr	The numeric week the data was recorded
Begindatum evenement iso	The date and time a vessel arrives at a lock or bridge with ISO-standard
Begindatum evenement	The date and time a vessel arrives at a lock or bridge
UN-locatie herkomst	UN-location code of the start harbour
UN-locatie bestemming	UN-location code of the destination harbour
Scheepstype RWS	The type of vessel in categorical numbers defined by RWS
SK_code	The size of a vessel in categorical codes defined by RWS
Laadvermogen	The amount of cargo a vessel can load
Beladingscode	The status of the loaded cargo in categorical numbers
Vervoerd gewicht	Amount of cargo that is loaded at that moment
Containers TEU	Amount of containers that are loaded at that moment
Nstr	The classification of goods based on the NST/R
Nst2007	The classification of goods based on the NST 2007

Even though the combined dataset gives a relatively complete overview of vessel movement in inland waterways in the Netherlands, the data quality issues persist. The ship type values between the two datasets can sometimes differ. This makes it difficult to only identify the tanker vessels in the combined dataset. Besides, the combination of AIS and IVS data for analysis is a relatively new method, so there is no standardized way to join the two datasets. In a previous project, the data was first split into separate trips that could be identified, which were then joined based on matching ship identification, date and timeframe (with a 4-hour deviation) (Rijkswaterstaat, 2024b). For the IVS data, the trip identification is relatively straightforward as it contains voyage-related information. However, for the AIS data this is challenging as it does not contain a field which can be used to identify separate trips, especially in the anonymised data. This is another limitation of the AIS data for this project. However, by combining the datasets, this data will be added to the AIS dataset. Therefore, in this project the data will be joined before different trips will be identified. The methodology will explain how the two datasets will be joined purely on timestamps and ship identification.

2.3 Datasets

To enhance the combined AIS and IVS dataset further, some extra data is needed. The extra datasets that are used in this project are the dataset containing all the berth locations on the Dutch waterways, all the harbour locations in the Netherlands, the navigability of the Dutch waterways, the bunker stations on the Dutch waterways and a dataset containing polygons of all the surface water bodies maintained by Rijkswaterstaat. Most of these datasets could easily be found on the sites of Rijkswaterstaat. Only the dataset containing the bunker stations was difficult to find. The used data now consists of multiple sources including OpenStreetMap, Rijkswaterstaat and inland shipping experience and knowledge sites which provided names and addresses of bunker station locations. Table 3 gives an overview of the different datasets that are used in the research.

Table 3: All the datasets used in the research other than the AIS and IVS data

DATASET	SOURCE	DESCRIPTION
Harbours	Rijkswaterstaat	Dataset containing all the harbours in the Netherlands. Location of the harbours is represented as a point or line.
Berth locations	Rijkswaterstaat	The berth locations on the inland waterways in the Netherlands.
Navigability	Rijkswaterstaat	Inland waterways of the Netherlands represented as a network. Includes information on the navigability of all the waterways.
Bunker stations	Rijkswaterstaat OpenStreetMap Binnenvaartkennis.nl	The bunker stations on the inland waterways in the Netherlands. A combination of three different datasets.
Surface water bodies managed by Rijkswaterstaat	Rijkswaterstaat	A polygon representation of all the inland waterways that are managed by Rijkswaterstaat

Some of the datasets needed some adjustments. The data of all the harbours in the Netherlands, for example. In the ideal situation, the dataset of Dutch harbours would contain polygons that cover only the waterbodies that were considered a harbour, so the waterway to reach the harbours would not be included. This dataset was not found. The closest available dataset was the dataset of OpenStreetMap, but this dataset was not complete. The harbour of Utrecht for example was marked as just a waterway, so this harbour could not easily be selected with a query. This could have been fixed by updating the OpenStreetMap data, but it was unknown how many more harbours would need to be fixed. The other option was the dataset of Rijkswaterstaat, which only consists of points and lines representing the harbours. This dataset does however include all the Dutch harbours and was used in combination with a buffer around these points and lines to better cover the harbour area. The navigability dataset also needed an adjustment. The dataset consisted of separate linestrings for every waterway. However, the dataset needs to represent a network, so at every intersection of two or more waterways the linestring is split. These split sections can then be turned into a network, where every waterway is connected correctly.

2.4 Substances Prohibited from Degassing by the CDNI

The CDNI has declared a ban on the 20 most transported VOCs in its operation area (CDNI, 2023; Geerlings & Kuipers, 2019). This ban will be rolled out in three phases, the first phase has the most transported and most dangerous substances and the last phase has the least transported and least dangerous substances of this list of 20 VOCs (CDNI, 2023). For this research, the substances of all three phases are included. The list of substances per phase can be found in Table 4.

Table 4: Substances that are prohibited from degassing by the CDNI divided by phase

Phase	UN-number	Description
1	1114	Benzene
	1203	Petrol or fuel for automotive engine
	1268	Petroleum distillates, petroleum products, N.S.O.
	3475	Ethanol and petrol, blended, or ethanol and fuel for automotive engines, blended, containing more than 10% ethanol
2	1267	Crude oil (containing more than 10% benzene)
	1993	Inflammable liquid, N.S.O. containing more than 10 % benzene
	3295	Liquid hydrocarbons, N.S.O. containing more than 10% benzene
3	1090	Acetone
	1145	Cyclohexane
	1170	Ethanol (ethyl alcohol) or ethanol in solution (ethyl alcohol in solution), aqueous solution containing more than 70% alcohol by volume
	1179	Ether ethylene butyl
	1216	Isooctanes
	1230	Methanol
	1267	Crude oil (containing less than 10% benzene)
	1993	Inflammable liquid, N.S.O. containing less than 10 % benzene
	2398	methyl tertiary butyl ether
	3257	Liquid transported when hot, N.S.O. (Including molten metal, molten salt, etc.) at a temperature equal to or greater than 100° C and below its flashpoint
	3295	Liquid hydrocarbons, N.S.O. containing less than 10% benzene
	9001	Substances with a flashpoint above 60° C handed over for transport or transported at a temperature within the range of 15 K below the flashpoint or substances the flashpoint of which > 60° C, heated to within less than 15 K of the flashpoint
	9003	Substances with a flashpoint greater than 60° C and less than or equal to 100° C which cannot be assigned to any other class or heading within class 9

2.5 Clustering

A clustering method is needed to eliminate redundant AIS data when a ship is not moving. As the AIS transmitter sends a record every 10 seconds, the amount of data recorded in approximately the same place can rapidly increase when a ship is not moving. To reduce the amount of data records and the size of the dataset, these points are clustered together with a clustering algorithm. The cluster algorithm will be used within a database with the PostGIS extension. The PostGIS extension provides multiple clustering algorithms, like DBSCAN, K-means clustering, intersect clustering and distance clustering. The K-means clustering algorithm is the only algorithm provided by PostGIS that can handle 3D geometry (Section 3.2 explains this further). Thus, the cluster K-means method will be used to cluster the points.

K-Means clustering is a so-called hill-climbing clustering algorithm, it was one of the earliest hill-climbing cluster algorithm and has been used in many clustering applications (Everitt et al., 2011; Likas et al., 2003). Hill-climbing algorithms are algorithms that use optimization techniques to find the optimal solution to the proposed problem by iteratively improving a solution. For the k-means clustering algorithm, this is the rearrangement of existing partitions and keeping the new one only if it provides improvement (Everitt et al., 2011). The k-means clustering algorithm is point-based and starts with arbitrary points as cluster means (cluster centres) (Likas et al., 2003). All the points are appointed to the cluster with the closest cluster mean. Once all the points are appointed, a new cluster mean is determined, after which the points are again appointed to the cluster with the closest cluster mean (Everitt et al., 2011; Likas et al., 2003). As mentioned before, this iterative process continues as long as it provides improvement (Everitt et al., 2011).

One of the disadvantages of this algorithm is its sensitivity to the initial (arbitrary) positions of the cluster means when multiple clusters are located close together (Likas et al., 2003). For detecting stop points this should not be a problem, as separate stop points are located far enough apart. Another big disadvantage of k-means clustering is the determination of the amount of clusters. In general the number of clusters has to be defined up front as input for the algorithm (Ming-Tso Chiang & Mirkin, 2010). However, the cluster k-means method of PostGIS also provides the option to cluster based on a maximum radius, independent of the number of clusters. In that case, the algorithm starts as a cluster k-means algorithm, but when a point is further than the determined distance from the cluster mean, it will not be included in that cluster and a new cluster is created.

2.6 Optimal Path Algorithm

In order to determine if a vessel deviated from the optimal route, the optimal route between the start and the end point of a trip needs to be determined. Optimal path computing over a network is one of the showpieces of real-world applications of algorithmics (Delling et al., 2009). The optimal path can be considered the path with the least resistance (Johner et al., 2022). In this case, the optimal route is the route with the shortest distance within the waterway network. A network is a graph with arcs and nodes (Wu & Shan, 2000). The network can be on all kinds of subjects, including communication or relatedness of the ingredients in a recipe, but the most used type of network is the transport network (Cai et al., 2011; Delling et al., 2009; Leigh et al., 2007; Wu & Shan, 2000). In a transport network, the nodes represent locations with significance, like an intersection (Yadav et al., 2020). The edges are lines connecting these locations, which represent ways or paths (Yadav et al., 2020).

Optimal path algorithms (also called route planning algorithms) determine the most efficient path between two or more locations while considering the principles or constraints of the graph/network (Cai et al., 2011; Golshani et al., 1992). There are many different kinds of algorithms for different kinds of situations (Delling et al., 2009). Most of the algorithms can handle a basic static network, but some algorithms are optimized for applications with dynamic networks or to avoid being trapped in a local solution (Delling et al., 2009; Leigh et al., 2007; Wu & Shan, 2000). One of the classic and most applied algorithms is the Dijkstra algorithm (Bast et al., 2016; Nha et al., 2012). This algorithm, together with the Bellman-Ford and Floyd-Warshall algorithm, are considered basic or uninformed algorithms (Bast et al., 2016; Cai et al., 2011; Johner et al., 2022). The difference between these algorithms is in the number of nodes each algorithm visits and the amounts of paths that are calculated (Delling et al., 2009; Johner et al., 2022; Yadav et al., 2020). The fastest algorithm of these three is the Dijkstra algorithm, as it stops when the destination is reached (Breed, 2021; Johner et al., 2022). The execution times of these basic algorithms increases rapidly with increasing graph sizes (Johner et al., 2022). To improve the efficiency of the algorithms the uninformed algorithm can be changed to an informed algorithm by applying speed-up techniques to them (Breed, 2021; Cai et al., 2011; Delling et al., 2011; Johner et al., 2022). Uninformed algorithms operate solely on the structure of the graph they explore, without any additional information on the distance or state of the target, so they will start searching without any bias (Johner et al., 2022). This will result in the optimal route, but has a long execution time, especially on large-scale graphs (Johner et al., 2022). By adding extra information about the direction, distance or state of the target to the algorithm, it becomes an informed algorithm. However, due to the relatively small size of the Dutch waterways network data, these informed algorithms do not seem necessary.

2.7 Degassing behaviour

There are currently no behavioural patterns that can be identified as degassing behaviour. In a previous research by Haskins et al. (2024), however, some patterns were identified that could be related to degassing. This research only included detours, loops and overlaps in the movement of the vessels. These three types of behaviour were selected based on the assumption that "ships behaving suspiciously are likely to follow paths which overlap [loop or detour]" (Haskins et al., 2024, p. 19). In this research, the optimal route will be calculated to compare with the taken route. By comparing the optimal route to the taken route, all the identified deviating patterns are included in the analysis and the research is not limited to only detour, looping or overlapping behaviour. However, this method also does not include all the deviating behaviour that could be related to degassing. For example, slowing down the speed of the vessel could also be identified as deviating behaviour related to degassing. This type of deviating behaviour is difficult to identify, as it is dependent on many different dynamic factors, and thus this kind of behaviour is not included in this research.

As there is currently no known behavioural pattern related to degassing, there is also no database containing vessel ids or movement where the relation to degassing can be certified. This makes it difficult to validate the results of the research, as there is no database to test the results against. However, other methods to still validate the results will be proposed in Section 3.5.

2.8 Density Estimation

Density estimation is a statistical method for visualizing the distribution of observations in a dataset (Kamilaris & Ostermann, 2018). It is used in various fields including archaeology, banking, climatology, economics, genetics, hydrology and physiology (Sheather, 2004). The density estimation is the construction of an estimate of the density function from the observed data (Silverman, 1998). The density function is a fundamental concept in statistics, Silverman (1998) explains it as follows: "Consider any random quantity *X* that has probability density function *f*. Specifying the function *f* gives a natural description of the distribution of *X*, and allows probabilities associated with *X* to be found from the relation." (p.1). Density estimations can be made on univariate and multivariate data. The multivariate methods of density estimation are similar to the univariate methods, as all the multivariate methods are generalizations of univariate methods (Silverman, 1998). The histogram is the oldest and most widely used density estimator (Silverman, 1998). Other examples are the naïve estimator, the kernel estimator, nearest neighbour method, the variable kernel method, orthogonal series estimators, maximum penalized likelihood estimators, general weight function estimators, bounded domains and directional data (Silverman, 1998).

In geospatial research, density estimation takes known quantities of the measured phenomenon at each location and examines the spatial relationship of the location of the measured quantities (Kamilaris & Ostermann, 2018). An important note to make is that density estimation is a method for exploring and displaying the spatial patterns of point data to show areas of high-concentration (Cromley & Mclafferty, 2002). So, it is not a cluster detection method as such, as it will only show the visual clustering and not the statistical or mathematical clustering (Cromley & Mclafferty, 2002). A particularly common output of density estimation in the geospatial domain is a heat map. A heat map assigns a density value to each raster cell and visualizes these values using a temperature gradient for the entire map (Kamilaris & Ostermann, 2018).

A heat map is also what will be created in this research. After the calculation of the optimal path and detecting deviating patterns, a dataset with deviating trips will be created. This dataset is filtered on explainable behaviour, after which only unexplainable deviating behaviour is left. This will be the input for the heatmap, which can be used to identify hotspots for deviating behaviour which could be related to degassing.

2.9 Emission Volume Analysis

The emission of degassing consists of a part of the liquid VOC that is vaporized during transport. The amount of liquid that is vaporized is dependent on several factors, like the density of the liquid, the vapour pressure of the liquid, temperature and the amount of cargo. The calculations are based on some assumptions, as there are many dynamic variables in the calculation of vaporized liquid (Loefen, 2017). The calculation is thus a rough estimation of the reality.

Loefen (2017) and Bolt (2003) already made calculations on the amount of emission from degassing. Loefen (2017) uses Equation 1 to calculate the emission of a single vessel transporting 1000 ton of benzene. This equation can however also be used to calculate the emission of different substances, as it uses variables that are also known for other substances. The equation, however, does not take temperature and residue into consideration, so these still have to be added to the formula.

$$Mass_{T=15°C}^{vapor} = Vapor\ pressure_{T=15°C}^{benzene} \times Vapor\ density^{benzene} \times \frac{Mass\ product}{\rho_{benzene}} \quad (1)$$

Bolt (2003) uses Equation 2 to calculate the evaporation factor (EF [kg/tonne]) of a substance in the tank of a vessel. The equation uses the standard air pressure (P_{air} [kPa]) and density of air at 20°C (ρ_{air} [kg/m³]) in the Netherlands. These values are constant and thus do not change for different substances. This is also the case for the saturation factor of the air-vapor mixture (S) and the amount of residue in the tank (RC [kg/tonne]). The variables relative density to air (ρ_{vapour}), density of the liquid (ρ_{liquid} [tonne/m³]), vapour pressure (P_{vapour} [kPa]) and temperature correction (CorrT) differ per substance. The relative density to air, density of liquid and vapour pressure are all characteristic of a substance and can be found in the literature. The temperature correction can be calculated with the Clausius-Clapeyron equation, see Equation 3 (Mondal et al., 2022).

$$EF = \frac{\rho_{air}}{P_{air}} \times \frac{\rho_{vapour}}{\rho_{liquid}} \times P_{vapour} \times S \times CorrT + RC \quad (2)$$

The Clausius-Clapeyron equation uses the enthalpy of vaporization (ΔH_{vap} [J/mol]), the pressures (p_1 and p_2 [atm]), temperatures (T_1 and T_2 [K]) and the gas constant (R [J/mol-K] to calculate the pressure at the temperature of interest. The gas constant is always equal to 8,314 J/mol-K, $p_1$and T_1 are the variables that are already known, ΔH_{vap} is a characteristic of a substance and T_2 is the temperature for which the pressure needs to be calculated. The equation will result in p_2 which is the vapour pressure of the substance at the temperature T_2. By converting the result to a relative difference value of the vapour pressure at the known temperature, the temperature correction is calculated.

$$ln\left(\frac{p_1}{p_2}\right) = -\frac{\Delta H_{vap}}{R}\left(\frac{1}{T_2} - \frac{1}{T_1}\right) \quad (3)$$

This calculation will only be done on the substances banned by the CDNI (see Section 2.3). For every substance in Table 4 the total amount of transported weight will be calculated in addition to the total amount of degassed cargo weight for that substance. Which vessels are included in the amount of degassed cargo weight is again based on the deviation analysis mentioned in Section 2.5.

2.10 Conceptual Framework

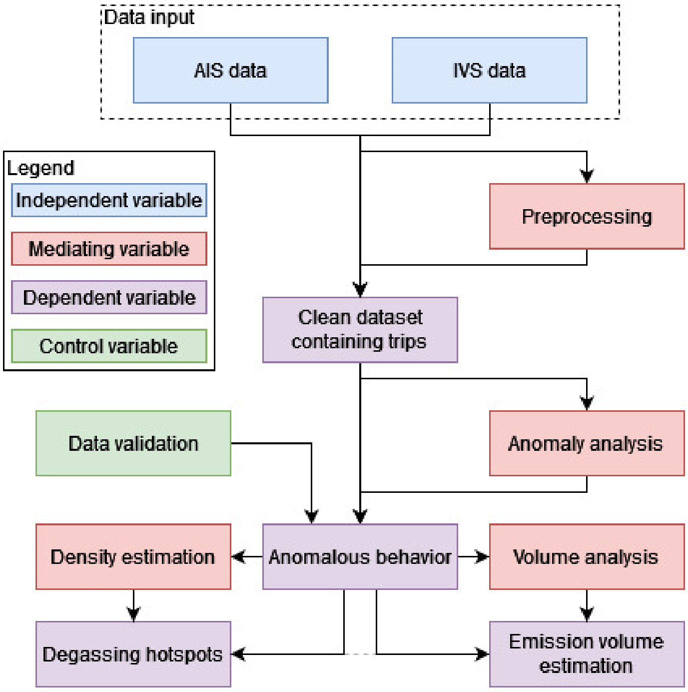

Figure 2: Conceptual framework for the research 'Degassing in Inland Shipping'

Figure 2 shows the conceptual framework of this research. The conceptual framework provides a theoretical overview of intended research and the order within that process (Leshem & Trafford, 2007). It is a way to model possible patterns and relationships which establishes or defines the boundaries of the research (Leshem & Trafford, 2007). This, however, does not mean that the framework is rigid, as the framework can evolve as long as the research progresses (Leshem & Trafford, 2007). The conceptual framework shows how the theories and concepts, explained in the theoretical framework, are connected and which results it will produce. It consists of independent, mediating, dependent and control variables (Swaen & George, 2024). The independent variables are the cause or input (Swaen & George, 2024), which is the AIS and IVS data in this case. The mediating variables are the link between the independent and dependent variables, it explains the relationship between them (Swaen & George, 2024). In this case, it is the different theories and methods that will be applied to the variables. The dependent variables are the effects or results of the independent variables plus the influence of the mediating variable (Swaen & George, 2024). This can both be intermediate results or the end results. The control variables are the factors that are monitored to ensure that the relationships between the variables are not influenced by external variations (Swaen & George, 2024). In this case, the data validation minimizes errors or biases that could affect the dependent or mediating variables. By adding the control variable, the subsequent analyses are based on more accurate and reliable data. The different mediating and control variables will be further explained in the Methodology chapter. The end results are a map containing the 'Degassing hotspots' and the 'Emission volume estimation'. These results are separate outcomes of the research and are not linked to each other.

The theoretical framework defines the different concepts and theories that are used in the research. It should give a sufficient explanation of these theories and concepts to understand the rest of the report. The next chapter, the Methodology, will explain the choices that were made on the theories and concepts elaborated in the Theoretical framework and expand on the Theoretical framework with a more focused explanation of the chosen algorithm or parameters.

3 Solution Design and Development

The Methodology expands on the Theoretical framework by presenting the research design and explaining the choices that were made during the research. The theories and concepts explained in the Theoretical framework will be put into practice in the Methodology. The methodology starts with the flowchart, which will be elaborated on in the remaining part of the chapter.

3.1 Flowchart

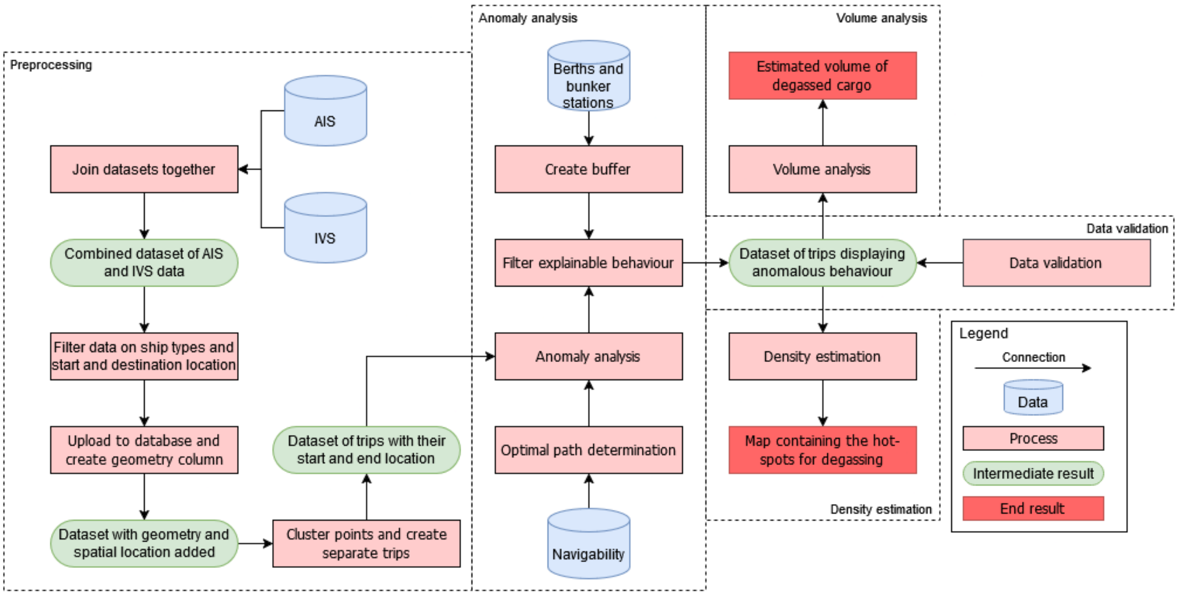

Figure 3: Flowchart for the research 'Degassing in Inland Shipping'

To expand on the conceptual framework, a flowchart is presented in Figure 3. While the conceptual framework only shows the different variables included in this research, the flowchart expands on these variables by showing the steps that will be taken for every variable and the dependency between these steps. The mediating variables of the conceptual framework are presented in the flowchart with a box of dashed lines. Within these boxes are the steps that belong to the mediating variable.

The flowchart shows the steps that will be taken in the research, the parts where data is added, some intermediate results to be obtained, and the end results of the research. The flowchart clearly shows that there are two separate end results and conclusions (estimated volume of degassed cargo and a map containing the hotspots for degassing).

The joining of the data, the first filter process, and the density estimation will be performed with Python. To help with the calculation time and storage capabilities, permission was asked and granted to work on the DelftBlue supercomputer. All the other steps are performed in a PostgreSQL database. To help with storage capabilities, permission was asked and granted to work on the server of the ABE faculty of Delft University of Technology. On this server, a PostgreSQL database is setup where the processes of this research can be performed.

The flowchart is also a guide for the Methodology chapter, as all the steps present in the flowchart will be explained in more detail in the remaining part of this chapter. The explanation is in chronological order of the research, so the start of the flowchart is at the top left corner with the input of IVS and AIS data and the last parts are the two end results on the right of the flowchart.

3.2 Preprocessing of the Data

The preprocessing of the data consists of multiple steps, where the first step is already taken before acquiring the data. At Rijkswaterstaat, the original IVS data and AIS data are pseudonymized with the same pseudonymization before the data is sent out. This step is done internally at Rijkswaterstaat to make sure the privacy regulations are followed and the data can be joined correctly. By using the same pseudonymization for both datasets, the datasets can still be joined after privacy sensitive data is anonymized. By using pseudonymized AIS and IVS data and aggregating the final results into a heat map, the privacy policy of AIS and IVS data should be satisfied throughout the whole process. The initial design of the proposed methodology was developed and tested using a one-week test dataset of AIS data between Amsterdam and Rotterdam from December 2020. Once a full-month dataset of AIS and IVS data became available, the method was adapted to the new format and finalized.

Joining and Filtering

After the data is pseudonymized at Rijkswaterstaat and send out, the data can be joined together. As both datasets contain the pseudonymized vessel identifier and a timestamp for each record, these two characteristics can be used to join the data. However, the interval frequency of the AIS data (approximately every 10 seconds) is not equal to the interval frequency of the IVS data (Every time a vessel reaches an object). With a python script, for every record in the AIS dataset a check is performed if the pseudonymized vessel identification is also present in the IVS dataset. If it is present in the IVS dataset, the timestamp of the AIS dataset is compared to the timestamp of the IVS dataset. If the timestamp of the AIS dataset is later than the recorded timestamp of the IVS dataset, the IVS data is joined to the AIS dataset. The result is a dataset where the most recent IVS record of a certain vessel is joined with the records of the AIS dataset where the timestamp is later than the IVS timestamp, until the AIS timestamp surpasses the timestamp of another IVS record (Appendix A1).

Once the data is acquired and joined, the data needs to be filtered to significantly reduce the size of the dataset and to remove unnecessary data. This is done in the same script. The filtering is done after joining the data, as some filter criteria are on IVS data characteristics and some on AIS data characteristics. Only the data about tanker vessels on inland waterways will be kept in the dataset. Besides, many columns are filtered out, as they do not contain relevant information for this research. Only 17 columns are kept of both datasets, these columns can be found in the python script (Appendix A1). The data will be filtered with a piece of Python code that streams the dataset through the script to ensure efficient memory usage. Besides, parallel processing is used to speed up the whole process, this however resulted in some extra code to make sure the parallel processing was working and to make sure the metadata was correctly provided for every separate process . As the AIS dataset is quite large and parallel processing is required, the DelftBlue super computer (DHPC) is used to speed up the process and to make sure sufficient memory is available to perform the Python processes in this research . Once the data is filtered, it can be uploaded to the database. The database is a PostgreSQL database with the PostGIS and pgRouting extensions installed, to make sure it can handle geometry data and the optimal route analysis. This database is running on a Linux-server of the TU Delft.

Database

Within the database, the first step is to create a table (Appendix B1) and add a geometry column from the latitude and longitude data (Appendix B2). Both 2D and 3D geometry are added to the database, where the third dimension is the time relative to the first recorded time of a vessel. Both geometries are added to the database, as the 3D geometry will be used for the clustering and the 2D geometry will be used for all the other spatial calculations. The 3D data is needed because the points will be clustered both on space and time (third dimension is not height but time in this case). The 3D geometry will not be used in the other spatial calculations, as the time dimension is not needed in these calculations and the usage of the 2D geometry is more time and resource efficient.

The dataset consists of a continuous stream of all the movements of tanker vessels in inland waterways in the Netherlands. To make the data more useful for further analysis, the movements need to be split into separate trips that were made by the different vessel, represented with a single linestring (Appendix B3). To accomplish this, first the points close to each other, for example when a vessel was waiting, need to be clustered to make sure the linestring does not store excessive data. As the dataset contains a 3D geometry consisting of the spatial and temporal position, the data can be clustered both in space and time. For the clustering, the K-means clustering method is used, which is part of the PostGIS extension. The K-means clustering method is used, as it is one of the provided clustering methods in the PostGIS extension which delivers the desired result and is the only clustering methods of PostGIS that supports 3D input. The K-means clustering requires as input the geometry, the amount of cluster as integer and the max radius of a cluster. After an iterative process, the following values were used as input: the geometry is the 3D geometry on time and space, the amount of clusters is set to 1 (max radius will be used to decide the clusters) and the max radius is set to 0.0001 degrees. If the query detects a cluster (of two or more points), the centre of all the points is returned with relevant information like the UN cargo number, the transported cargo weight, the earliest time included in the cluster, the ship type and the vessel identification. In first instance the DBSCAN cluster method was used, which is a different and more efficient way to cluster. But the DBSCAN method of PostGIS did not support 3D geometry and thus could not be used.

These points, together with the points that fall outside a cluster, are then split into separate segments of the movement of a single vessel. These segments should represent the different trips a vessel made over time. This is one of the most challenging parts of the research as there is no definitive point or method to split the movement into segments. Multiple methods are considered, which all have different results. The first sub-question of this research is focused on this problem and in Section 4.1 the different methods and considerations for this problem are explained in more detail. The chosen method for this problem is to split the data every time the vessel has a gap of more than 2 hours in the data (after clustering the points) or when the amount of transported cargo (IVS data) changes. This makes sure all the relevant changes in the IVS data are used as input to split the data into trips, but also includes stop points of over 2 hours, for when the vessel is not transporting any cargo.

For every trip, a simple cleaning process is applied. An error or corruption in the AIS data would place one or two AIS points of a single track somewhere, far away from the rest of the track, on land or near the water. There was no explanation for these errors found and only a few tracks had this error. The cleaning process removes some of these outliers that are present in the AIS data. The process will check if a point is within 5 kilometres of its previous and its next point in the track (based on timestamp). If it is not within 5 kilometres, the point will be not be used as input for the track and the track is based on the points that are left after the cleaning. Once all the movements are split into segments and outliers are removed, the points related to a single trip of a vessel are turned into a linestring in the order of the timestamps of the points. This will result in a dataset containing the vessel id (pseudo id), trip id (segment id) and the geometry of the trip (linestring) together with the characteristics of the vessel.

One of these characteristics is the previously transported cargo. This data together with the upcoming cargo is needed for the validation of the process. For the validation of the process the consecutive loads are examined to check if a ship needed to degas, to identify false-positive or negative results. This data is added to the database with a query, and to give more context the second to last cargo is also added to the trip (Appendix B4).

3.3 Shortest Path Analysis

The Theoretical framework gave a short introduction to shortest path analysis and its algorithms. There are many different algorithms, so the choice of algorithm will be explained step by step.

As this research includes a transport network (network of inland waterways in the Netherlands), the focus will be on graph-based route planning algorithm. The most classical approach is the Dijkstra algorithm (Cai et al., 2011; Cao et al., 2009; Delling et al., 2009). The Dijkstra algorithm functions by calculating the path from one node to all the other nodes in the graph until it reaches the target node (Delling et al., 2009; Yadav et al., 2020). The algorithm visits the nodes based on proximity with the start node and keeps a priority queue in which the shortest path known at the moment is stored (Delling et al., 2009; Schultes, 2008; Yadav et al., 2020). Due to its selection of nodes based on proximity and the algorithm stopping at the target node, the Dijkstra algorithm can be considered a greedy method (Johner et al., 2022).

As mentioned in the Section 2.6, there are various techniques that can be applied to reduce the computation time of the Dijkstra algorithm, like heuristic/goal-directed, data structure, area restriction, hierarchical, separator-based and bounded-hop techniques (Breed, 2021; Cai et al., 2011; Delling et al., 2011). However, as the network only includes the waterways of the Netherlands, the graph stays relatively small, so the speed-up techniques will have a minimal increase in execution time and thus are unfavourable. This also excludes the application of meta-heuristic algorithms, as they are optimized for dynamic, large-scale graphs (Nanayakkara et al., 2007).

The pgRouting extension for PostgreSQL databases can calculate the shortest path using the Dijkstra algorithm within a database. So, this is the application that is used in the research to calculate the shortest path over the network from the start to the destination location of the different trips.

The network data is created from the navigability dataset[1] provided by Rijkswaterstaat. With the use of GIS software, this dataset is split into separate segments and transformed back into a valid network with pgRouting (Appendix B5). Within the database, the start and end point of each trip is extracted and the closest node to these points in the network is used as input for the shortest route algorithm. The pgRouting application is set to use the length of the segments as the default value for the cost parameter, so it does not need to be specified later in the process. As the used dataset includes information on the navigability of the segments, this can be used to specify which part of the dataset will be used within the algorithm. This is useful to separately calculate the optimal route for different ship sizes. The result of this function is a table with all the separate segments that form the route from the start to the end point including the sequence in which they need to be followed. This result can be combined within the database to create a linestring with the right sequence of segments. By combining this linestring with the pseudo and segment id, the optimal route can be used in the deviation analysis to compare the optimal route to the taken route. The full query that performs these steps can be found in Appendix B6.

3.4 Deviation Analysis

The previous research of Haskins et al. (2024) detected three kinds of deviating behaviour that could be related to degassing. First, there is sailing in a loop for an extended period of time. Secondly, there is taking a wrong turn at an intersection and turning back around, or ‘overlap’ as Haskins et al.

[1] https://www.vaarweginformatie.nl/frp/main/#/geo/map?layers=NAVIGABILITY

(2024) called it. Lastly, vessels could make a small trip around the harbour to degas and eventually return to the same harbour to collect their new cargo.

The deviation analysis is quite simple due to all the preprocessing that is already done on the data (Appendix B7). The first part of the deviation analysis is a simple comparison of the length of the taken route against the length of the optimal route. This should detect the first two identified deviating behaviours. As the optimal route is perfectly straight and in the centre of the waterway, an increase of 25% is added to the total length of the optimal route. The 25% is determined through testing with several values and should exceed the extra length added to the taken route due to oncoming traffic, drifting and other small detours. Besides, the start and end points of the taken and optimal routes are not perfectly aligned, so this is also taken care of in the analysis. By checking if the taken route is longer than the optimal route (+25%), the trips of a vessel where a significant detour was made are selected from the dataset. These trips have a high probability of floating degassing and are thus added to the deviating behaviour dataset.

The second part of this list consists of the trips identified with the third deviating behaviour. To identify the trips displaying this behaviour, the trips are tested against various conditions. First off, because the start- and endpoint are located very close to each other, the shortest route analysis could not be completed. So, the first filter is all the routes where the shortest route calculation could not be completed. This filtered dataset is then tested against the condition that the start and end location are within 1 kilometre of each other. Lastly, to filter out all the short trips within the harbour, the filtered dataset is tested against the conditions of an interval of more than 7,5 hours between the start and end time of the trip and a minimal length of 10km of the trip. After applying these filters, the dataset returns the trips with the deviating behaviour of making a small trip around the harbour to degas and eventually returning to the same harbour to collect their new cargo.

The deviation analysis is an iterative process. After performing the deviation analysis, the results are evaluated and described. In Section 4.2 the different identified deviating behaviours are presented and analysed. The question 'Is the identified behaviour actually deviating behaviour or can the movements of the vessel be explained?' is asked on a sample set of results. If the behaviour can be explained, the characteristics of this behaviour are analysed. If certain characteristics are unique to this behaviour, the filter of the deviation analysis is tightened to make sure that the explainable behaviour is not identified as deviating (Appendix B8). Some explainable behaviour that could be marked by the deviation analysis is resting or waiting at a (free) berth location until the next cargo is received or the allotted time at the berth location has passed, refuelling at a bunker station or mooring the vessel. By checking if the trips approached a bunker station, this explainable behaviour can be filtered out of the deviating behaviour dataset. This is done by loading the bunker stations into the database, creating a buffer of 25 meters around the bunker station and checking if a trip intersects with this buffer. For checking if a trip used a berth location to rest or replenish, the berth locations are also loaded into the database. As some berth locations are located on the waterway, especially at locks, the method used for bunker stations cannot be used. Instead, the trips that intersect with a berth location are turned back into points and the amount of points of a single trip on a single berth location are counted. If the amount of points is higher than 90, the trip is filtered out of the deviating behaviour dataset. As AIS signals are send approximately every 10 seconds, 90 points is equivalent to 900 seconds or 15 minutes. So if a trip stopped for longer than 15 minutes on a single berth location, the trip is removed from the deviating behaviour dataset. After several iterations, the deviating trips of the first and second deviation analyses can be combined into one dataset and can be used in the following steps of the analysis.

3.5 Data Validation

Since the classifications based on the data do not reflect reality, random errors in the classification may lead to inaccurate conclusions (Banerjee et al., 2009). By checking the results against certified results, wrongly classified results can be identified. This will not eliminate the uncertainty completely, but it can quantify the uncertainty (Banerjee et al., 2009) and give an indication on the validity of the results. However, as degassing is quite a new restriction for inland shipping, there is no concrete dataset to test the result of the research against. To still validate the research, the following methods are considered.

The first option is to discuss the results of the research with experts on the subject of degassing. This includes employees of Rijkswaterstaat who work on the subject of degassing or inland shipping and employees, including inspectors, at the ILT who carry out the supervision on degassing. This can also be in the form of some fieldwork with an inspector of the ILT, to check the results in the field.

A second option is to check the results against the results of the e-nose network in the Netherlands. E-noses are electronic noses that reproduce the structure and principle of olfactory sense (sense of smell) to distinguish complex volatiles, mostly used in the food industry, medical treatment and environmental detection (Arakawa et al., 2023). As Arakawa et al. (2023) state: “In the detection and classification of gas and odor, electronic nose has the characteristics of high sensitivity, high efficiency, and high recognition.” The data of this e-nose network is however managed by the environmental services in the Netherlands. In total, there are 28 environmental services in the Netherlands, which means that the data is not managed and collected by one organization. Besides, the ILT and Rijkswaterstaat do not have permission to use the data without making a request. This makes it difficult to access the full dataset of all e-noses in the network, and the full network is needed to make a useful statement about the validity of the results of the research. The ILT does, however, have some data on illegal degassing activity. However, this dataset is small and contains privacy-sensitive information, which makes it difficult to use for the validation of the results. Besides, the known degassing behaviour changes over time and keeps on changing as the law and the enforcement adapts to the current situation.

Lastly, there is a more statistical approach, in which (a part of) the false-positive and false-negative results are determined. A false-positive result is a result where the hypothesis is classified as true, while in reality, it is false (type I error) (Banerjee et al., 2009). A false-negative result is a result where the hypothesis is classified as false, while in reality, it is true (type II error) (Banerjee et al., 2009). By examining the substance that is being transported at a certain trip identified as deviating, as well as the substances in the trip before and/or after that trip, it can be determined whether a ship needed to be degassed or not. If a trip is identified as deviating, but the succession of substances indicates that degassing of the vessel was not needed, the trip can be identified as a false-positive. The same can be done to identify false negatives. For each identified trip, the transported substance is determined, along with the substance transported in the subsequent trip (Appendix B9). If degassing is needed between two trips, but the movement between those trips was not identified as deviating, it can be considered a false-negative.

With the use of a confusion matrix, an accuracy measure based on the proportion of correct classifications can be calculated. A confusion matrix is a table that summarizes and visualizes the performance of a classification algorithm (P. Singh et al., 2021). It creates a characterization of the data by comparing the classifications made by an algorithm to the known classifications (Lewis & Brown, 2001). As mentioned before, there are no known classifications available for this research. So, the identified false-positive and false-negative results, based on the data characteristics, are used as input for the confusion matrix instead of the known classifications. With the confusion matrix, a variety of descriptive and analytical measures that summarize the accuracy of classification can be calculated (Lewis & Brown, 2001). By comparing the proportion of correct classifications in the confusion matrix, a precision, recall and accuracy measure can be calculated (Lewis & Brown, 2001). Precision is the ratio of correct positive classifications to the total predictions of positive classifications and is calculated using Equation 4 (P. Singh et al., 2021).

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Recall is the ratio of correct positive classifications to the total number of actual positive classifications and can be calculated using Equation 5 (P. Singh et al., 2021).

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

Accuracy is the ratio of correct classifications to the total number of classifications and is calculated using Equation 6 (P. Singh et al., 2021).

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (6)$$

In Equation 4, 5 and 6 the true-positive (TP) represents the trips identified as deviating from vessels that also did need degassing. True-negative (TN) represents the trips not identified as deviating from vessels that also did not need degassing. False-positive (FP) represents the trips identified as deviating from vessels that did not need degassing. False-negative (FN) represents the trips not identified as deviating from vessels that did need degassing.

Identifying false-positive and negative results

In the preprocessing phase, a table is created with the current and the subsequent cargo transported by a vessel (Appendix B9). By comparing the current and following UN number, a decision can be made if degassing was needed or not and thus if true-positive or negative results are present in the data. In Table 5 the compatibility of different banned substances is presented. This table originates from the research by Koop (2016), and is based on the data from the VNPI (Netherlands Petroleum Industry Association) in the research of Buck et al. (2013) and the data present in EFOA (2008). Outside Table 5, no further information on compatibility of banned substances has been found in this research.

Table 5: Compatibility for several of the substances banned for degassing by the CDNI. Substances that are not present in the table have zero compatibility or no information is present about these substances.

Previous cargo → (UN number) → Next cargo ↓(UN number)↓	1114	1267	1268	1863	1993	3295	3475
1114							
1203							
1223							
1267							
1268							
1863							
1993							
2398							
3295							
3475							

Table 5 is created based on two assumptions:

- “Changes to the same UN-number are regarded 100% compatible when the UN-number refers to homogenous bulk chemicals. Consignors in some cases do distinguish products within the same product category from one producer to another. It is however assumed that the logic of full compatibility between homogenous bulk chemicals is valid for most cases” (Koop, 2016, p. 9);
- “Changes to the same UN-number are regarded 100% compatible when the UN-number refers to a mixture (for example: UN 3295 Hydrocarbons, liquid, not otherwise specified). Compared to the ‘pure’ products, this will less often be the case. However, no general information nor specific cargo change information is available” (Koop, 2016, p. 9);

Besides, cargoes that are considered not to be compatible with any of the preceding cargoes are omitted from Table 5 (Koop, 2016). Even though the table is not complete, as not for every substance sufficient information could be found on the compatibility, for substances not present in the table it can be considered that the substance is not compatible with other substances. In Appendix B10 the compatibility table is turned into a query to check if degassing was needed for certain vessels.

3.6 Density Estimation

As mentioned in the Theoretical framework, the most widely used density estimation is the histogram. When the data is in two or more dimensions, the arguments for using different methods become much stronger. The construction and presentation of a multivariate histogram have severe presentational difficulties and use more parameters, like the size of the bins, the origin of the system of bins and the orientation of the bins (Silverman, 1998). The second most used method is the kernel density estimation, which is also the method that will be used to create the heat map in this research. The kernel density estimation approach is simple and intuitively appealing and its mathematical properties are well-understood (Silverman, 1998). Besides, ‘undersmoothing’ in the tails of the data is not a problem, as the research is aimed at hotspots, so the kernel approach is chosen over the adaptive kernel approach.

The kernel density estimation is a nonparametric method that infers the probability distribution of a dataset by examining the data itself, without relying on predefined assumptions about the shape of the distribution model (Chen et al., 2024). This technique allows for flexible and data-driven estimation that captures the distribution’s inherent characteristics without the limitations of parametric assumptions (Chen et al., 2024). The kernel density estimation moves a kernel (or spatial window) across the study area and calculates the density of points within the kernel (Cromley & Mclafferty, 2002). Usually, the kernel has a constant size and the events within the kernel are weighted according to the kernel function. The kernel function describes mathematically how the weights vary over the distance from the centre of the kernel (Cromley & Mclafferty, 2002). By giving the events near the centre a greater weight than those distant from the centre, the kernel density estimation reflects the underlying geographical locations of events within each kernel (Cromley & Mclafferty, 2002). After computing the density of each kernel, the result can be presented as a heat map in which the value (or colour) of each cell reflects the intensity of the phenomenon at that location (Delmelle et al., 2014).

The Python library ‘Seaborn’ is a statistical data visualization library and contains a function called kernel density estimate plot. This function is used by the kernel density estimation plot (KDE-plot) function of the Geoplot library, which creates KDE-plots for geospatial databases. The KDE-plot function of the Geoplot library is used in this research to create a heatmap of the results of the deviation analysis. This function only takes points as input data, so the results of the deviation analysis need to be converted back to points before they can be used as input for the KDE-plot function. As most of the identified trips start and end in a harbour, these locations will be identified as hotspots while it is very unlikely that a harbour is a hotspot for degassing due to the busy environment (this however does not mean it does not happen). To make sure the harbours do not outclass the actual degassing hotspots, the points within harbours will not be used as input for the density estimation. The conversion from linestrings to points, including the location filter, will be the last query within the database before the results are exported to Python (Appendix B11).

The KDE-plot function takes the x and y values of the points, and plots them on a multivariate graph. The distribution of points over the graph is used as input for the kernel density estimation method of the seaborn library to create the heatmap. The KDE-plot function gives the ability to set a threshold and the number of levels that need to be included in the calculation by calling back on the kernel density estimate function of seaborn. Besides, it also gives several visualization options, including clipping, colour map selection and if the contours need to be filled in or not. The choice of colormap is important in heatmaps, as it is important to choose colours for each pixel that lead the viewer to perceive the data faithfully (*Plotting Pitfalls — Datashader*, 2024). Most of the colormaps provided in python libraries are highly nonuniform (*Plotting Pitfalls — Datashader*, 2024; Walt & Smith, n.d.). Fortunately, the matplotlib library (which is already in use in the Python script) has four uniform colormaps included in its library (Walt & Smith, n.d.), of which the 'inferno' colormap will be used in this research. The results of the kernel density estimate will be clipped on the contours of the major waterways and bodies in the Netherlands. This dataset is provided by Rijkswaterstaat[2]. The other settings used in the KDE-plot function are for cosmetic purposes. The full script can be found in Appendix A2.

3.7 Volume Analysis

In the Theoretical framework, two equations are presented that can be used to calculate the emission of a vessel as a result of degassing. Both equations use vapour pressure, vapour density and relative density of the substance as input, but use different methods to calculate the emission. The maximum difference between the results of the equations for the same substance is around 88%, where the equation used by Loefen (2017) is always higher than the equation used by Bolt (2003). This is probably because the equation used by Loefen (2017) initially does not take the temperature correction, saturation air mixture and the residual liquid into consideration. The residual liquid of a tank is added later in the report of Loefen (2017). As the equation used by Bolt (2003) takes more factors into account, this equation will be used to calculate the emission of a vessel due to degassing.

This equation requires multiple inputs which need to be determined. First off, the standard air pressure and the density of air at 20 °C need to be determined. The standard air pressure in the Netherlands is on average 1015.5 hPa (or 101.55 kPa) as determined by the KNMI[3]. The density of air at 20°C in the Netherlands is between 1.20 and 1.23 as determined by the KNMI[4]. The value exactly in between the interval is used as the average, so the density of air used in the equation is equal to 1.215 kg/m^3.

Secondly, the relative density to air, density of liquid and vapour pressure need to be determined for every banned UN number (Table 3). Most of these values can be found on the International Chemical

[2] https://maps.rijkswaterstaat.nl/dataregister/srv/eng/catalog.search#/metadata/rws1680f-68b5-4ff3-94a4-9c24109ffd5e
[3] https://wow.knmi.nl/nieuws/nieuws-nieuws-item50
[4] https://www.knmi.nl/over-het-knmi/nieuws/lichte-lucht-zware-lucht

Safety Cards (ICSCs)[5] developed by the International Labour Organization (ILO, an agency of the United Nations) and the World Health Organization (WHO). Any missing information can be supplemented with information from the PubChem compound summary[6] developed and maintained by the National Center for Biotechnology Information. For UN number 3475 the same values as for UN number 1170 are used as they referenced to the same page. These UN numbers have the same base substance, but the relative proportions of the substance contents are different. For UN number 1267 (crude oil) some averages of the variables are taken. Crude oil can be found in different places around the world, which results in different kinds of crude oil, with different vapor pressures and densities. By using the average, most of the crude oils are represented in the variables.

Also for UN number 1993 the average is taken as this UN number is a collection of substances. In total, 25 substances were used to calculate the average vapour pressure and densities for UN number 1993. Lastly, for UN numbers 3257, 9001 and 9003 no data could be found on the characteristics of these substances that are used in the equation. These UN classes use broad descriptions and thus can be related to many substances. The classes do not specify a substance or a mixture of substances but use temperature or characteristics of a substance. Therefore, these UN numbers will unfortunately be left out of the emission calculation.

The next variable is the saturation factor of the air-vapor mixture. The saturation of the air-vapour mixture near the bottom of the tank is around 100% (Bolt, 2003). However, higher in the tank the saturation decreases. Bolt (2003) concludes that, as an average for the whole tank, a saturation factor of 0.56 can be used.

The temperature correction factor can be calculated based on the Clausius-Clapeyron equation, as mentioned in the Section 2.9. This requires some more characteristics of the substances. The temperature of the initial vapour pressure is mentioned on the ICSCs, the temperature for which the vapour pressure needs to be calculated is 10°C as this is approximately the average temperature in the Netherlands (Bolt, 2003; The World Bank Group, 2021). Lastly, the enthalpy of vaporization needs to be determined which can be found on PubChem compound summary for most substances. By dividing the new vapor pressure by the initial vapor pressure, the temperature correction is calculated. If the enthalpy of vaporization is not known, a generally applicable correction factor of 0.75 can be used to correct for the temperature difference (Bolt, 2003).

Lastly, the residual cargo needs to be determined for the equation. Bolt (2003) estimates the residual cargo at 0,07‰ of the transported cargo. Loefen (2017) has an indicated residual cargo that is very close to this number, so 0,07‰ of the transported cargo can be used as a rule of thumb for the residual cargo variable.

The above-mentioned characteristics of substances, together with IVS data, which contains information on the vessel type, cargo amount, and cargo type, can be used to make an estimation of the amount of vapour that is present after discharging the vessel's cargo. By adding all the amounts of vapour of each vessel, an estimation of the total emission due to degassing can be determined. As mentioned before, this is a very rough estimate of the total emission produced due to degassing.

The information on the vessel type, cargo amount and cargo type can be queried from the database (Appendix B12). The query first adds a column in which the cargo weight is converted from kilograms to tonnes and the calculated evaporation factor is applied. Essentially, this column contains the amount of emission created by the vessel in question. This is only applied on the vessels identified as

---

[5] https://webapps.ilo.org/dyn/icsc/showcard.home
[6] https://www.ncbi.nlm.nih.gov/pccompound/?term=

displaying deviating behaviour and the transported UN number is taken into consideration, to make sure the right evaporation factor is applied. The are added together to calculate the amount of emission for every UN number that will be banned by the CDNI and counts the number of vessels that were identified as displaying deviating behaviour per UN number. The result of this query thus will be a table which quickly displays the amount of emission in kilograms and number of vessels for every UN cargo number that is banned by the CDNI.

3.8 Responsible Data and Algorithm Usage

Part of the data used in this research is considered personal data and thus privacy legislation must be followed when working with the data. By pseudonymizing the data, the personal aspect of the data is removed. The pseudonymizing of the data is done internally at Rijkswaterstaat and is outside of the scope of this research. However, as the data contains location data, it can be traced back to a company or even an individual as certain locations can be private, a shipyard for example. To make sure none of the personal data is shared through this research, the data is additionally aggregated by the creation of a heat map of the results of the deviation analysis. Some of the intermediate results will be shared in the Chapter 4, but these results will be shown without spatial context.

Besides the pseudonymization of the data, every step taken in this research is explained in detail by explaining the choices that were made and the theory behind the used methods. The process should be completely repeatable based on the information given in this report, ensuring transparency and reproducibility. Additionally, responsible data handling practices are followed to prevent misuse, and any limitations of the approach are explicitly acknowledged. This ensures a transparent, explainable, and ethically sound process, leading to trustworthy results.

4 Results

In the results chapters, the explained methodology will be applied and the results are discussed. It will contain the answers to the sub-questions of the research. The structure of this chapter follows the order of the sub-questions, as the answers to the first sub-questions will supports the results of the following sub-questions and so on.

The initial aim for this project was to analyse the AIS and IVS data for the whole of 2022 and the first half of 2024 over all the inland waterways in the Netherlands. Due to processing limitations at Rijkswaterstaat on the data, the data of only one month (February) of 2024 between Amsterdam and Rotterdam could be provided and is used as input for the analysis. This can result in less valuable results, as only a small portion of the initial data is used as input, but can work as a proof of concept to test if the proposed method works as expected and if it delivers correct results. Thus, all the results presented in this chapter are extracted from the AIS and IVS data of February 2024 between Amsterdam and Rotterdam. Table 6 shows the amount of records and the timespan of the different datasets. Due to filtering and aggregation the amount of records in the dataset decreases as the research progresses.

A comment that must be made on the joining process of the AIS and IVS data is that the joining of the two datasets is based on the (pseudonymized) Erinumber (in the AIS data) and shipnumber (in the IVS data) of a vessel. However, not all records in the AIS data had a Erinumber. Besides, there were problems with the identification of tankers as the AIS and IVS data not always identified the same vessel as a tanker vessel. Also within the IVS data itself there were sometimes tanker vessels that transported dry cargo (based on the provided cargo type) and vice versa. Both these characteristics of the data resulted in the inclusion of all the vessels that were identified as tanker by the IVS data, but also the ships identified as a tanker within the AIS data without an Erinumbers. This means that there are AIS records within the dataset that are used as input for the analysis, but do not contain the voyage related information from the IVS data. This data may be needed later in the methodology, when the emission is calculated and the results are validated, which means that these records become useless after the first half of the analysis. Other vessel identification values, like MMSI- or ENI-number, were not included in the IVS data, so this was the only option to join the two datasets.

Table 6: An overview of the amount of records and the timespan of the different datasets

Dataset/processing step	Amount of records / lines in dataset	Timespan
AIS dataset	170.598.624	2024-02-01 14:57:00 - 2024-02-25 08:37:16
IVS dataset	34.420	2024-01-01 00:28:00 - 2024-06-30 23:52:00
Joined dataset	21.461.063	2024-02-01 14:57:00 - 2024-02-25 08:37:16
Dataset of identified trips	45.616	2024-02-01 14:57:00 - 2024-02-25 08:37:16
Dataset of deviating vessel trips	504	2024-02-01 14:57:00 - 2024-02-25 08:37:16
Dataset containing all the unexplainable deviating vessel trips	160	2024-02-01 14:57:00 - 2024-02-25 08:37:15

4.1 Splitting the Data

The AIS data just consists of a continuous stream of records about the status of the vessel, like location, speed, heading, etcetera. It does not contain any voyage related information, like the start or end location, the amount of cargo and the type of cargo on board. This makes it difficult to identify separate trips from the data, which is needed to identify deviating behaviour of vessels on inland waterways. The first sub-question focusses on this topic. Different methods were considered, which all have their own advantages and disadvantages. These methods and their advantages and disadvantages will be discussed in the following section.

Different Methods and their Characteristics

The IVS data does contain voyage related information and by joining the two datasets together, this data is added to the AIS data. By joining the data, the IVS data can be used to split the AIS data. But as the data is differently formatted and needs to be joined together, it creates room for errors. For example, the interval between timestamps of AIS data and IVS data is not equal. So, if a new trip is already started but the IVS data is not updated as the vessel did not yet reach an object, the first records of AIS data of the new trip still have the IVS data of the previous trip.
Additionally, the movement of a vessel after unloading the cargo will be classified as a single trip. Whilst this might be true for some trips, most of the time the vessel does not move straight from the unloading location to the next loading place. This means that the IVS records cannot be used to split the AIS data into separate trips, or at least not as the sole condition.

Besides voyage related characteristics, spatial or temporal characteristics can also be used to split the AIS data into trips. The first proposed method to split the AIS data was to split on a stop time of more than 2 hours within a harbour or berth location or a stop time of more than 12 hours outside the harbours or berth locations. This method was based on the assumption that discharging of the cargo takes at least 2 hours and outside of harbours and berth locations there was not really a reason to stop, so only prominent stop times were used to split the data. This, however, neglected the vessels that lay at anchor on an inland lake for example. As a relatively high threshold was set for time outside harbours and berth locations, these trips were not split even though it is better to also split the trips outside the harbours and berth locations when a ship stopped for more than 2 hours, for example.

Another consideration was to split the trips on places where a ship could load or unload its cargo. As most of the times, a trip starts when the cargo is loaded and ends when (part of) the cargo is unloaded. For example, in a harbour or on berth locations of the category ‘load and discharge’. When a ship stopped for more than 2 hours on a place where it could load or unload its cargo, the trip would be split. Again, 2 hours was used, as the unloading of cargo was considered to take at least 2 hours. There were several problems with this method, as the provided data on berth locations of the category ‘load and discharge’ was incomplete. Multiple berth locations were found in the data that were located around an unloading facility but were not of the category ‘load and discharge’. Besides, a vessel being in a harbour for more than two hours does not always mean that it is unloading. And, like is mentioned above, only looking at berth locations of the category ‘load and discharge’ was not an option.

Lastly, a method that purely uses a temporal condition was proposed. This method is essentially the same as the method of splitting on a stop time of more than 2 hours inside a harbour, but the geographical condition is removed. So, for this method, every stop time of more than 2 hours is used to split the data. The threshold of 2 hours is based on the considered minimal time that is needed to unload a ship, but also on the target time of 30 minutes of waiting time at a lock set by Rijkswaterstaat (Koedijk, 2020). By extending this time significantly, the time needed to pass the lock and the occasional exceeding of the target time is included. This made sure the waiting time at a lock or a bridge was not the reason a trip was split. The problem with this method is that it will probably split one trip into multiple segments, as a vessel could stop for more than 2 hours without loading or unloading its cargo. At least it will make sure that every trip between two stop points is included in the dataset. This is especially useful for when a ship is waiting for new cargo and is attending multiple locations, like resting places, bunker stations or berth locations.

Method used in this research

By combining various aspects of the previous mentioned methods, some of the mentioned problems can be irradicated. As the split based on the IVS data is the closest to the actual situation, this is used as the base for the method. To make sure the movement after unloading is also split accordingly, the temporal condition of stop times of more than two hours is also added to the final method. This way, trips made to resting places, for example, are also identified as separate trips. To make sure no IVS data of the previous trip is added to the current trip due to slight misalignment in the joining process, the most frequent value is used as the voyage data of the trip.

This classification of trips is based on the input of experts on the field of inland shipping. At the start of the research, the method based on geographical and temporal conditions was used to identify separate trips. After discussing the results from the whole process on the test dataset with the supervisors and experts, the feedback was given to split the data more frequently outside of the harbours, as a vessel can engage in different activities after the cargo is unloaded. The suggestion was to split the data after every stop point, so this is what is being implemented. After some research, like the target waiting time at locks, and the input from the meetings, a stop time of at least 2 hours was used as the condition to split the data.

In total 45.620 different trips from 1.064 different vessels could be identified over the month of February 2024. Four trips had to be manually filtered out, due to significant gaps or errors in the AIS data. There are still some trips where the route is cut off over land, but this is probably only a small deviation from the actually taken route. Figure 4 shows all the identified routes and there geographical distribution. A clear geographical constraint on the different trips can be seen in Figure 4, as all the trips are perfectly cut off at certain points. This is the geographical scope that was applied to the AIS data by Rijkswaterstaat, so all the data outside of this geographical scope was not available for the analysis. Of the 45.620 trips, there are also a lot of short trips of only a few points, that never left the harbour for example. These trips will be automatically filtered out in the further processing of the data.

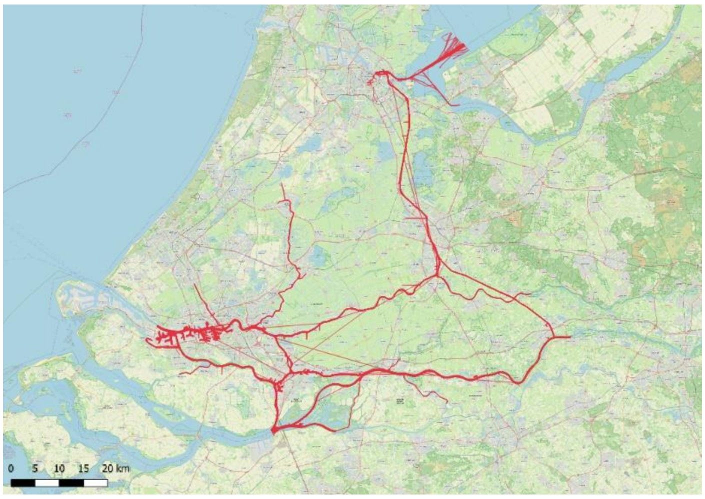

Figure 4: Geographical distribution of the identified trips between Amsterdam and Rotterdam in February 2024.

4.2 Deviating Behaviour

The deviations from the optimal route based on the method explained in the methodology will be presented in the following chapter. To quickly summarize the methodology, deviating behaviour is based on a comparison between the length of the taken route against the length of the optimal route + 25% and on trips that have the same start and end point. After applying this filter, only 504 of the 45.620 trips returned. These trips contain all kinds of deviating behaviour, but some of this behaviour can be explained. For example, a trip to a shop or to fuel up a vessel at a bunker station are also returned in the deviating behaviour analysis. Figure 5, for example, shows all the movement of trips classified as deviating around two bunker stations. Noteworthy are all the turns that are made around the bunker stations. It is highly probable that these are trips of vessels that fuelled up at one of the bunker station and returned back to their origin location. The pink colour stands out in Figure 5. Every colour indicates a different vessel, so the pink vessel has returned many times to one of these bunker station. This probably indicates that this is a bunker vessels of one of the bunker stations, which can fuel a ship whilst it is moving. Most of these trips need to be filtered out of the results as they (probably) have nothing to do with degassing.

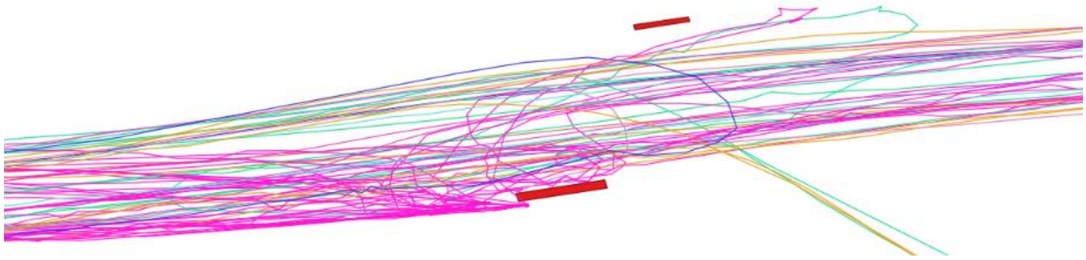

Figure 5: Zoomed-in cut out of the vessel movement of trips identified as displaying deviating behaviour around two bunker stations (red rectangles). The trips are displayed as lines, where every colour identifies a different vessel. The pink colour stands out in this figure, this is probably a bunker vessel. Due to confidentiality of the data, the geographical context is left out.

By further filtering the deviating trips, most of the explainable behaviour is removed from the results. There are still some cases, like in Figure 6, where some of the explainable behaviour is still in the results. In this case, this is due to the vessels not quite reaching the bunker station. The figure displays three different trips of a single vessel around a bunker station. As shown in the figure, the vessel does not quite reach the bunker station, which mean the applied filter does not identify it as explainable behaviour. Another example can be seen in Figure 7, but here the geographical scope of the AIS data prevents the filter of identify these trips as explainable behaviour. Just outside the geographical scope of the AIS data a bunker station was situated. This is probably the movement of one of their bunker vessels. But as the bunker stations is situated outside of the geographical scope, these three trips did not end on the bunker station and thus did not satisfy the filter threshold and remain in the final dataset. These are the two examples of explainable behaviour that could be found in the final dataset containing ‘unexplainable’ deviating behaviour. In total, 160 separate trips were identified as unexplainable deviating behaviour.

Figure 6: Three trips of a single vessel around a bunker station. These trips were not filtered out of the final dataset, due to their distance to the bunker station. Due to confidentiality of the data, the geographical context is left out.

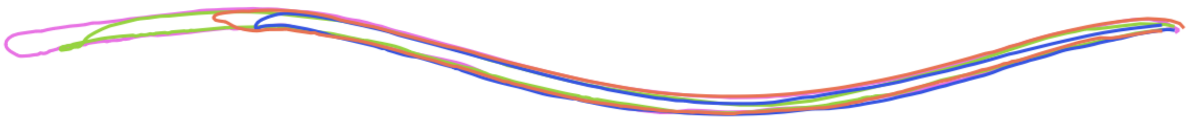

Figure 7: Vessel movement at the edge of the geographical scope of the AIS data. This is probably a bunker vessel, as there is a bunker station just outside the geographical scope of the data. Due to confidentiality of the data, the geographical context is left out.

Besides the few trips displaying explainable behaviour, there are four other types of behaviour that can be identified. These types of behaviours were also identified in the test dataset (one week of AIS data between Amsterdam and Rotterdam in December 2020). The identified behaviours were:

1. Making a small detour by taking a wrong turn and turning back around
2. Taking an extended route which is less populated or takes more time
3. Making a round trip around or near a harbour
4. Trips showing looping behaviour.

Three of the four descriptions of behaviours match with the degassing behaviours identified in the research of Haskins et al. (2024).

Figure 8 shows an example of the round trip behaviour. The figure displays a single trip of a vessel, which starts and ends in the same location. Part of the trip, in the left bottom corner of the figure, is cut of due to the geographical constraint of the AIS data. As the trip is still identified as a single trip, it can be concluded that not more than two hours of trip is cut off. Still, the rest of the trip displays the round trip behaviour quite clearly. Due to the confidentiality of the data, the geographical context of the trip is left out of Figure 8. But the trip starts in the top left corner of the figure in a more populated area, leaves this area for a less populated area, especially in the lower half of the figure, and then returns to the location it departed from. The approximate length of this route is 100 kilometres, without the part that is cutoff due to the geographical scope. In this area, multiple of these trips can be identified.

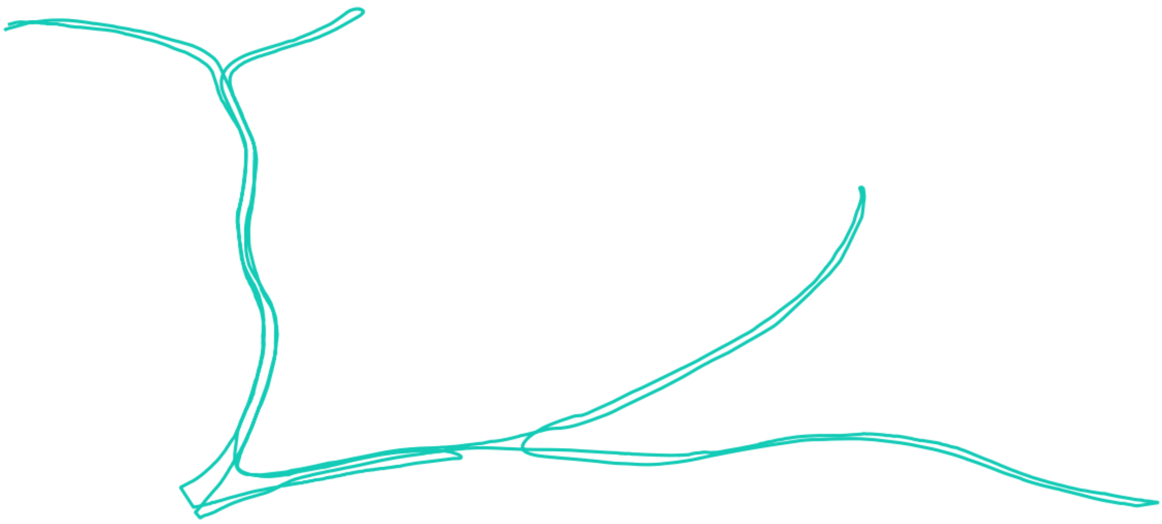

Figure 8: The round trip behaviour which could be identified in the AIS and IVS dataset of February 2024. Due to confidentiality of the data, the geographical context is left out.

The deviating behaviour of taking a wrong turn and turning around to return to the same waterway can be seen in Figure 9. Due to the geographical constraints of the AIS data, it is unclear where the vessel actually went for the detour. It could well be that the vessel went to a berth or bunker location. However, it returned back into the geographical scope within two hours as it is identified as a single trip. This removes a lot of possible actions it could have performed outside of the geographic scope. However, from Figure 9 it is clear that the vessel started somewhere in the left, made a detour and returned to the same waterway it started from and continued to the right. This fits the description of taking a small detour and continuing with the route perfectly.

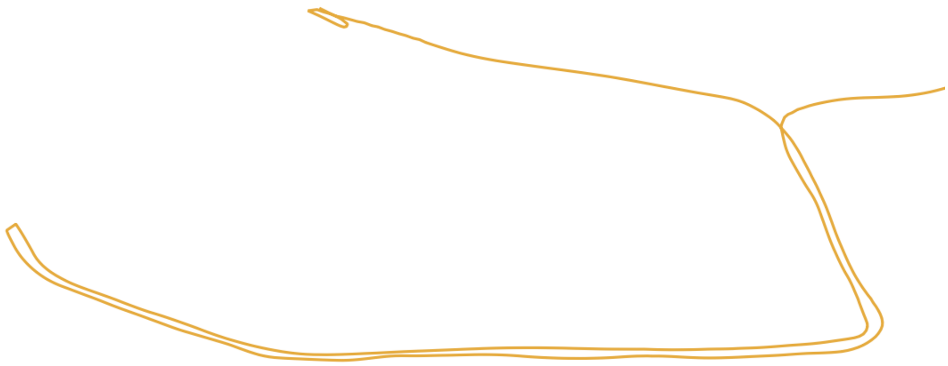

Figure 9: A vessel displaying the deviating behaviour of taking a small detour, which could be related to degassing. Due to confidentiality of the data, the geographical context is left out.

Another identified deviating behaviour was the loop behaviour. An example of this type of behaviour can be found in Figure 10. There is clearly some loop behaviour at the right side of the figure, so the trip is correctly identified. In this case the trip shows multiple loops in the movement at a single location, but there were also trips identified which only had a single loop in their movement. The location of the loops is mostly in less populated areas. After sufficient loops are made, the vessel continues on the route it originally planned to take to reach its destination.

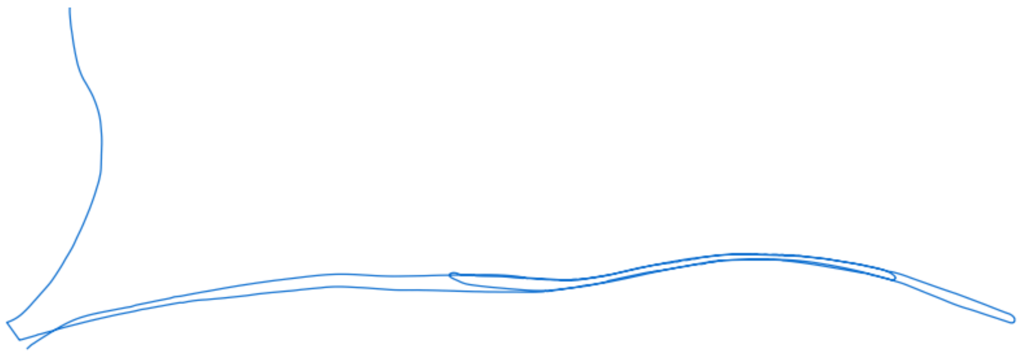

Figure 10: A trip displaying the looping behaviour. In this case there is only one loop, a vessel could also loop multiple times on the same place. Due to confidentiality of the data, the geographical context is left out.

A fourth type of deviating behaviour that can be identified from the results is taking a extended route. This behaviour can be seen in Figure 11 and could be due to two reasons. First, by following an extended route, there is more time to degas the vessel. Or second, the extended route is in a less populated area and thus they cause less nuisance and the chance of being noticed is smaller. The figure shows that instead of taking the left, shorter route, the vessel took the right, extended route. A reason for this deviation could be to attend a harbour or berth location, but the route in the figure does not show this kind of behaviour. This type of behaviour was not yet identified or mentioned in the research of Haskins et al. (2024).

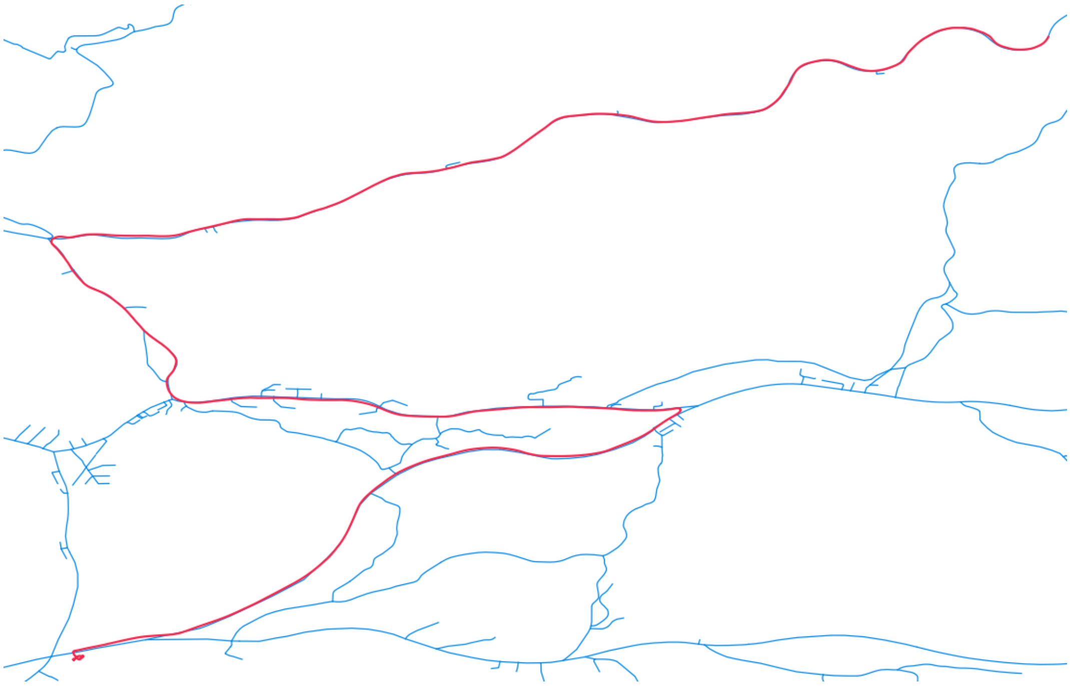

Figure 11: Deviating behaviour of taking an extended route. The red line is the taken route and the blue lines are the waterways. Due to confidentiality of the data, the geographical context is left out.

Figure 12 shows all the trips identified as deviating, after the ‘explainable behaviour’ filter is applied. It shows that most of the trips are located in or around a harbour. This is as expected, as the cargo is discharged at a harbour, after which degassing might be needed. Besides, a harbour is a place where all the different trips from different locations concentrate as they need to load or unload their cargo in or near a harbour.

Figure 12: The distribution of the identified deviating trips that could be related to degassing.

4.3 The Degassing Hotspots

Figure 13 shows the results of the kernel density estimation analysis on the identified deviating trips. The full size image can be found in Appendix C1. Also, a full size image were the kernel density estimation analysis is not clipped on the waterways can be found in Appendix C2. In this figure, the part of a trip that is inside of a harbour is not included, as it is unlikely (but not excluded) that a vessel is degassing within a harbour. If the points inside the harbour are not included in the kernel density estimation analysis, a clear hotspot can be found at the intersection of the Dordtsche Kil and Nieuwe Merwede/Hollandsch Diep and the area called Biesbosch located directly besides it. This is the most apparent hotspot that can clearly be found in Figure 13. The figure presented in Appendix C2, which is not clipped on the waterways, shows there are some more spots where the distribution of points, related to the trips identified as deviating, are concentrated. Figure 14 shows a cutout of Appendix C2.

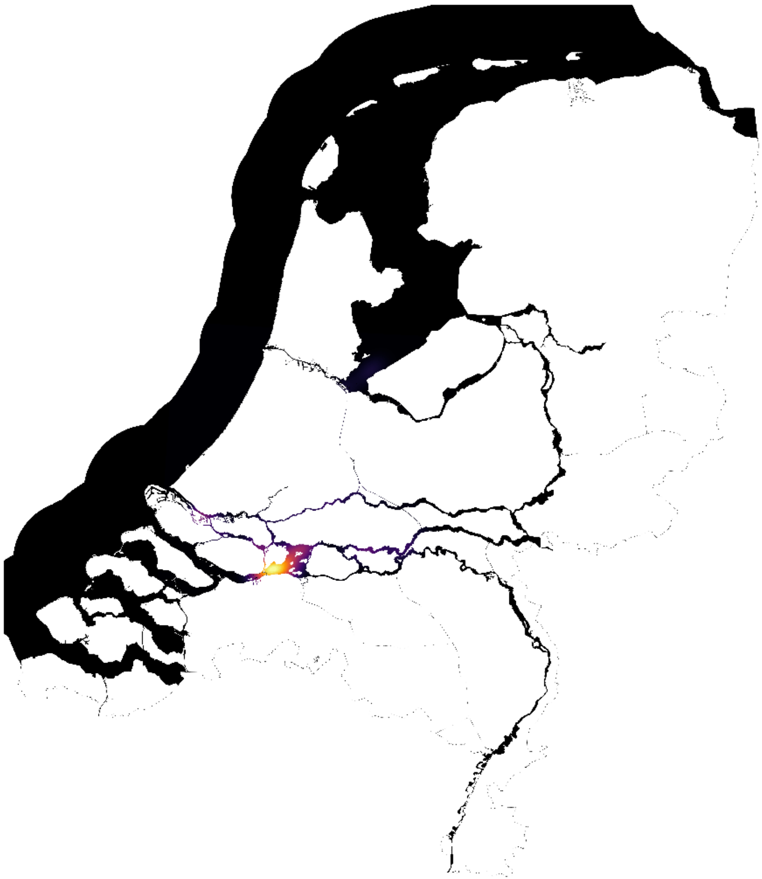

Figure 13: The result of the kernel density estimation analysis on the identified deviating trips

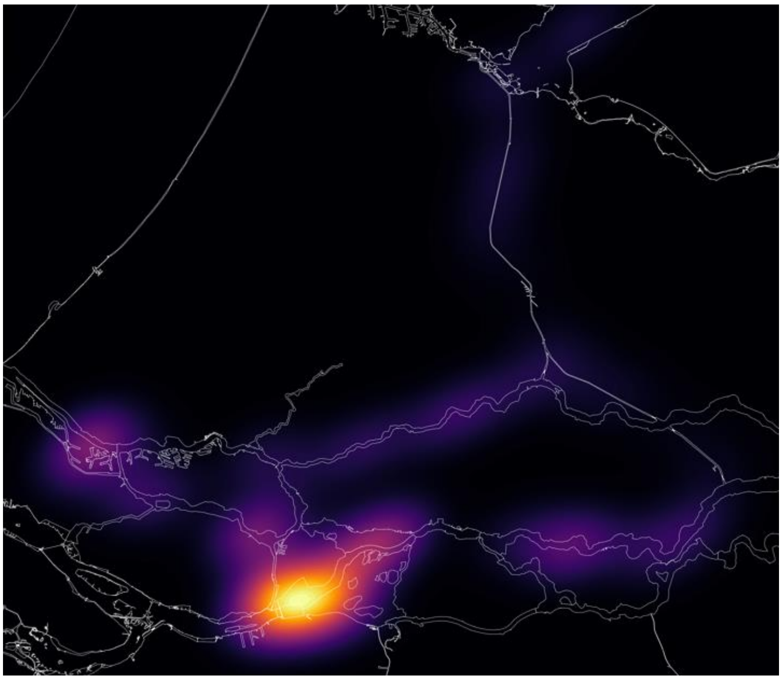

Figure 14: Cutout of Appendix C2, displaying the kernel density estimation plot without clipping the results on the waterways.

Figure 14 shows, just like Figure 12, that most activity is around the harbours, especially the Port of Rotterdam. Two noteworthy spots located around the Port of Rotterdam are: around the place of Zwijndrecht and around the Hartelkanaal, located just beneath the Port of Rotterdam. These spots could still result from the concentration of trips within a harbour, however, Zwijndrecht is not directly located at an access point to the Port of Rotterdam. Similarly, the Hartelkanaal is separated from an access point to the port and was not considered part of the Port of Rotterdam. All other parts of the trips within the Port of Rotterdam were excluded from the analysis, as they are situated within the harbour itself, but the points at the Hartelkanaal are included. Other locations marked as spots with a concentration of unexplainable deviating behaviour, outside of the main hotspot around the Biesbosch, are a part of the Amsterdam-Rijnkanaal just outside the Port of Amsterdam, the Waal around Zaltbommel and a big part of the river Lek.

Most of the hotspot of Figure 14 are located in a less populated area a few kilometres from a major harbour. This may be the reason why this location is preferred for degassing as the building density is low, which reduces the chance of getting noticed or caught degassing. Besides, due to the proximity to the harbour, the vessel can be cleaned quickly and new cargo can be collected as quickly as possible. This probably means that similar kinds of locations around other major harbours are also hotspots for degassing and thus should be the focus of the enforcement strategy.

If the parts of the deviating trips inside the harbours are included in the kernel density analysis, the analysis only covers the Port of Rotterdam and its surroundings. Other places, like the spot in the Lek or in the Waal are not concentrated enough to be clearly marked as a hotspot in the analysis. Figure 15 shows the results of the kernel density analysis if harbours were included in the analysis. As the figure shows, the Port of Rotterdam is the clear hotspot in this analysis. The other spots nearly disappear in this analysis. Around Dordrecht and the Biesbosch there is also a slight highlight. This is also the hotspot that can be seen in Figure 14, without the points of the harbour included. So, from the Port of Rotterdam, the river Oude Maas is slightly highlighted continuously until, around Zwijndrecht, where the density increases a bit. The highlight of the Oude Maas is barely noticeable in this small figure. In Appendix C3 the full size figure is included, this figure shows all the slight highlights better.

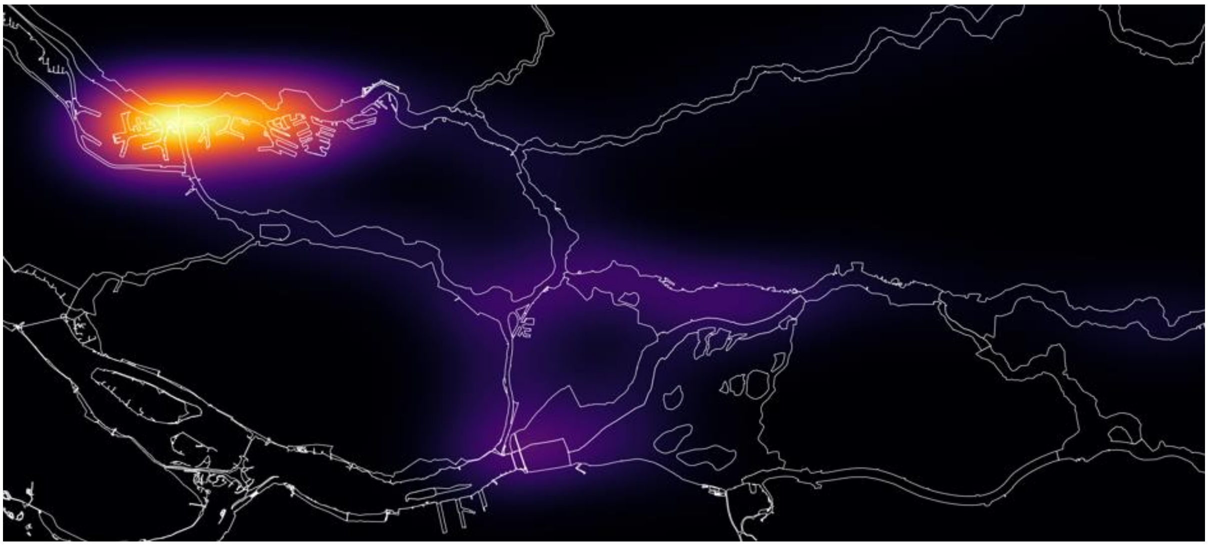

Figure 15: The result of the kernel density estimation when the parts of the identified deviating trips in the harbours is included.

4.4 Estimating the Emission of Degassing

The calculation of the emission estimation consists of two parts. First, for every substance the evaporation factor needs to be calculated. This is an equation based on several characteristics of the substance. These characteristics are provided by the KNMI, WHO, UN and NIH, like mentioned in Section 3.7. The second part is to calculate the amount of transported weight on the deviating trips for every substance. By combining these two values, (an estimation of) the amount of emission in kilograms due to degassing can be calculated.

Table 7 shows all the used values for the calculation of the evaporation factor and the calculated evaporation factor itself. The evaporation factor is calculated using Equation 2. If the enthalpy of vaporization was known, Equation 3 was used to calculate the temperature correction, otherwise a standard temperature correction of 0.75 was used based on the methodology explained by Bolt (2003). With the use of these equations and the characteristic of the substances, the evaporation factor at 10 °C could be calculated for most of the substances that will be banned by the CDNI. Like mentioned in Section 3.7, for the grouped substances UN3257, UN9001 and UN9003 no characteristics were available, thus the evaporation factor could not be calculated for these UN numbers.

Table 7: Characteristics and the evaporation factor for all the UN Numbers that will be banned by the CDNI.

UN-NR.	ρ AIR	P AIR [KPA]	ρ VAPOR	ρ LIQUID	P VAPOR [KPA]	S	ΔH_VAP [KJ/MOL]	CORRT	RC	EF
1114	1.215	101,55	2,70	0,88	10,00	0,56	33,83	0,61	0,07‰	0,20
1203	1.215	101,55	3,50	0,75	30,00	0,56	-	0,75	0,07‰	0,77
1268	1.215	101,55	4,75	0,78	0,75	0,56	38,07	0,58	0,07‰	0,09
3475	1.215	101,55	1,60	0,79	5,80	0,56	42,32	0,54	0,07‰	0,11
1267	1.215	101,55	11,03	0,85	5,33	0,56	-	0,75	0,07‰	0,42
1993	1.215	101,55	3,83	0,92	4,54	0,56	-	0,75	0,07‰	0,17
3295	1.215	101,55	2,35	0,73	26,75	0,56	-	0,75	0,07‰	0,51
1090	1.215	101,55	2,00	0,80	24,00	0,56	29,10	0,66	0,07‰	0,33
1145	1.215	101,55	2,90	0,80	10,30	0,56	29,98	0,65	0,07‰	0,23
1170	1.215	101,55	1,60	0,79	5,80	0,56	42,32	0,54	0,07‰	0,11
1179	1.215	101,55	3,50	0,75	12,80	0,56	32,18	0,63	0,07‰	0,32
1216	1.215	101,55	3,87	0,71	3,36	0,56	38,90	0,57	0,07‰	0,14
1230	1.215	101,55	1,10	0,79	12,90	0,56	37,34	0,58	0,07‰	0,14
1267	1.215	101,55	11,03	0,85	5,33	0,56	-	0,75	0,07‰	0,42
1993	1.215	101,55	3,83	0,92	4,54	0,56	-	0,75	0,07‰	0,17
2398	1.215	101,55	3,00	0,70	27,00	0,56	29,29	0,65	0,07‰	0,58
3257	1.215	101,55	-	-	-	-	-	-	0,07‰	-
3295	1.215	101,55	2,35	0,73	26,75	0,56	-	0,75	0,07‰	0,51
9001	1.215	101,55	-	-	-	-	-	-	0,07‰	-
9003	1.215	101,55	-	-	-	-	-	-	0,07‰	-

When checking the previous cargo of the identified unexplainable deviating behaviour, it is noticeable that only four of the UN numbers that will be banned by the CDNI are included in the results (see Table 8). The previous transported cargo is used as the degassing happens after the cargo is discharged, so the degassing happens on the substance of the previously transported cargo. Interestingly, for most of the vessels of the identified trips, the subsequent cargo is the same. This is a pattern that can be identified for a lot of vessels and is called dedicated shipping. This will be further explored in Section 4.5 on the validation of the results.

Table 8: Different characteristics of the trips that happened before the trips that were identified as possibly degassing. Only the first 10 characters of the pseudonymized id's are shown.

PSEUDO ID	TRIP ID	UN NUMBER	CARGO WEIGHT	NEXT UN NUMBER
0D6BD2B326…	1	1202	1350000	1202
0E0D376960…	1	1268	0	
40B3AE696D…	122	3082	1624000	3082
46F1B84C58…	334	1268	2270000	1268
472301C2A7…	47	1202	0	1202
5EB930194F…	61	3475	792008	3475
66A70BC25B…	55	1202	0	1202
A379C891CC…	86	2398	2100000	2398
C096BCB8E1…	14	1203	0	1203
CCD7B0633C…	434	1268	5640000	1268
EDE89CFA85…	40	3295	2500	

Of the 160 trips identified as possibly related to degassing, only 11 trips had a previous trip where cargo with a UN number was transported. For 10 of the 12 trips, the characteristics stayed the same, which could mean that a single trips was split into multiple trips by the algorithm or that the IVS data was not updated after the vessel discharged the cargo. Only 12 vessels had a previous trip where cargo with a UN number was transported, because, many of the identified trips which could be related to degassing were one of the first identified trips of a vessel, so most of the segment ids were 5 or lower. Besides, there were a lot of vessels which did not have any IVS data related to them. This is because vessels identified by the AIS data as tankers which did not have any IVS data were also included in the dataset. Because no IVS data is known for these vessels, there was also no UN number connected to the transported cargo. Therefore, it is unknown what the previously transported cargo for these vessels is.

One clear example is the vessel with pseudo id 'ede89cfa85…'. It transported UN number 3295 for several trips, after which it transported no UN number for some trips. One of these trips is also the trip that was identified as possibly related to degassing. On the second trip after the trip identified as possibly related to degassing, the vessel started transporting UN number 1268. Looking at the compatibility table (Table 5), it shows that UN number 3295 and UN number 1268 are not compatible with each other. So, the ship had to degas somewhere in between these cargo's. This can be both controlled degassing at an official degassing station or uncontrolled degassing into the atmosphere. Interestingly, in the few trips before the possible degassing trip, where no UN number was transported, the vessel was heading in the direction of Moerdijk, where an official degassing station is. However, Moerdijk was not included in the geographical scope, so this statement cannot be confirmed or denied by the data unfortunately.

Based on the current available data, a clear estimation on the amount of emission due to degassing cannot be made. The emission factor for most of the UN numbers that will be banned by the CDNI are calculated and ready to use. The script that selects the characteristics of the trip before the trip identified as displaying potential degassing behaviour is also prepared. But with the current data and applied methods only two vessel showing somewhat clear degassing behaviour can be identified, of which one did not transport any cargo weight based on the data. So currently, based on the currently available data and applied methods, the only statement that can be made is that 1,28 kilogram of UN number 3295 (Liquid hydrocarbons, N.S.O. containing less than 10% benzene) was emitted into the atmosphere due to degassing in February 2024. This could be accurate, but it is highly likely that also different substances where emitted into the atmosphere due to degassing.

4.5 Validating the Results

As mentioned in Section 2.7, the validation of the results is challenging as there is no 'ground truth' dataset which can be used to validate the results. However, several options were suggested to perform a validation on the results. One of these options was to discuss the results of the analysis with an inspector of the ILT, who can better identify if certain behaviours or locations are known for degassing and if the results seem logical. Also they could identify even more explainable behaviour that was still present in the results of the analysis. Unfortunately, due to the acquisition of the data only two weeks before the deadline of the project, there was no time to discuss the results with a professional. Visiting one or more hotspots in the field was also a proposed validation method, but this was also not possible in this short period of time.

However, in the test phase of the process, the first results on a test dataset of AIS data have been discussed with an inspector at the ILT. In the first meeting, there were no filters applied on the deviating behaviour, so every track that did not follow the optimal route was used in the final results. At this meeting, some clear examples of explainable deviating behaviour were mentioned, which were used to filter the data further. After this filter was added, a new meeting took place. In this meeting a certain case became the focus, as this case was identified as a degassing activity, and it had the characteristics of a degassing activity, but the place was very unusual. The activity took place too close to the shoreline according to the inspector, which resulted in some scepticism. At the same time, it was also unlikely to be an bunker boat, as it did not visit a bunker station.

This indicates that it is very difficult to identify degassing activities and behaviour purely on AIS data. Some extra context, outside the location and taken route is needed to strengthen the claim if a vessel was degassing or not. This could be in the form of IVS data, but, as mention before, due to the data quality of IVS data, this is not a complete and reliable representation of the actual situation. An other option could be to include some context on the last attended discharge location. This context may consist of the most common type of cargo that gets unloaded at this location, or the kind of industry that is present in the direct neighbourhood of the discharge location.

Another option to validate the results is to test the results against the data from the e-nose network around the inland waterways in the Netherlands. However, this data was not available at the time of the analysis. The same applies for the dataset containing known locations and times of degassing activities from the surveillance by inspectors or other means of the identification of degassing activities. Besides, this dataset was still quite small and may not contain any data for February 2024, as the ban was not yet in place in the Netherlands at that time.

The last method to validate the results, which is not without its caveats, is to perform a statistical analysis on the results with the use of a confusion matrix. Like mentioned in the methodology, the false-positive and false-negative results will be identified based on the sequence of transported goods. But due to the data quality of the IVS data and the used joining method, the results of this analysis will probably not represent the actual accuracy, recall and precision of the analysis, it will merely give an indication on the accuracy, recall and precision of the analysis.

Section 4.4 already briefly touched upon the validation of the results. As mentioned, there was only one clear example of a vessel that needed to degas, as it started transporting a different substance that was not compatible with the previously transported substance. Most of the other trips did not have IVS data joined to the AIS data or they only transported one type of substance. The transportation of only one kind of cargo is also called dedicated transport. This means that from a quality perspective the vessel does not have to degas (Buck et al., 2013), but it does not mean the vessel did not degas. This is one of the problems of identifying false-positive and false-negative results based on the data. As it is impossible to identify if a vessel has degassed when it is only used for dedicated transport. For now it is assumed that these trips have been wrongly identified as degassing. However, the trips which do not have any voyage related data, as they only consist of AIS data, are assumed to be correctly identified. This results in 102 true-positive and 58 false-positive results.

For the true-negative and false-negative results, all the subsequent cargo types are evaluated to see which vessel needed to degas and which vessel transported compatible cargo's. The compatibility of cargo is based on Table 5. For the trips, not identified as deviating behaviour, it is assumed that degassing was done into the atmosphere and not at a degassing station, as degassing of most of these substances was not banned in February 2024. However, the degassing could have also happened outside of the geographical scope of the provided data, which means it could not be identified as degassing behaviour with the current method and has implications on the result of the validity analysis. This results in a total of 140 false-negative results, which will leave 45.481 true-negative results. However, this also includes trips less than 5km, or the small movement of a vessel when it is not sailing. When a filter of a minimum length of 5km for a trip is applied, only 10.883 trips are left. This is probably a better representation of the actual situation, so this value (minus the false-positive, false-negative and true-positive results) will be used as the number of true-negative results. This results in the confusion matrix presented in Table 9.

Table 9: Confusion matrix of the results from the performed analysis

	ACTUAL POSITIVE	ACTUAL NEGATIVE
PREDICTED POSITIVE	102	58
PREDICTED NEGATIVE	140	10583

The values from the confusion matrix can be used to calculate the precision, recall and the accuracy of the analysis. Equation 7 shows the calculation of the precision of the performed analysis. The current proposed method has a precision of 63.8%, this indicates that about three-fifth of the predicted positive results are also actually positive.

$$Precision = \frac{TP}{TP + FP} = \frac{102}{102 + 58} = 0.638 \quad (7)$$

Equation 8 shows the calculation of the recall of the performed analysis. The current proposed method has a recall of 42.1%, this indicates that it correctly classified only about a two-fifth of the 'actual' trips that were related to degassing.

$$Recall = \frac{TP}{TP + FN} = \frac{102}{102 + 140} = 0.421 \quad (8)$$

Equation 9 shows the calculation of the accuracy of the performed analysis. The current proposed method has an accuracy of 98.2%, but this is heavily skewed by the large amount of true-negative results. As most of the vessel probably also did not degas if it was not necessary, it is logical that this value is high.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} = \frac{(102 + 10583)}{(102 + 58 + 140 + 10583)} = 0.982 \quad (9)$$

The precision of 63.8% indicates that there is still some more explainable behaviour that can be filtered out of the results. However, the biggest take away from the validation results is that only 42.1% of the actual vessels that needed to degas was correctly classified as potentially degassing. This may imply two causes: first, 57.9% of all the degassing happens without any deviation of the optimal route. It could be that this is only true for the trips within the area of Amsterdam to Rotterdam, but this could also be a nation-wide demeanour. The second cause could be due to the strict filter for removing the explainable behaviour. This filter removes trips of vessels that attended a bunker station or berth location, but the degassing behaviour could be related to one of these behaviours. A quick search in the data however implies that scenario one is more likely, as only two vessel ids could be found in both the actual degas behaviour and the identified degas behaviour (without the explainable behaviour filter) dataset. Besides, one of these vessel ids was the id of the vessel mentioned in Section 4.4, and thus was correctly classified.

Again, these validation results are based on multiple assumptions, so this is only an indication on the validity of the proposed method. However, these results already show that the proposed methodology is far from optimal. This is partly due to the used amount of input data, partly due to the amount of time that was available to perform the method on an actual dataset and partly due to the high amount of uncertainty around the behaviour of degassing.

5 Discussion

The results of this research give an indication on the potential locations where degassing activities take place. Not only the locations, but also an indication on the total amount of emissions of these degassing activities is calculated. However, there are some limitations to the research, and thus the results should be interpreted with careful consideration. This chapter provides a discussion on the research process and the produced results.

The research proposed several steps and methods which show potential to identify degassing behaviour based on AIS and IVS data. Some of these methods have been applied on an AIS and IVS dataset of February 2024. Unfortunately, due to the lack of input data and time, the analysis did not yield directly valuable results. However, the research did provide several important insights into the applicability of AIS and IVS data for the detection of degassing behaviour.

The analysis did identify specific deviations in vessel movement that may indicate degassing activity, but, just as important, it also identified deviations in vessel movement that were not related to degassing activities. And, while the overall precision of the methodology was 63.8% and the recall 42.1%, the clustering of deviation patterns and the presence of clear hotspots indicates that some locations are more favourable for degassing than others. This gives a first indication of potential degassing hotspots, even if further validation of the method is needed.

The initial aim and formulated research question was, in retrospect, overly ambitious. The current research question was too ambitious, as the only previous research in this area was the short orientation on degassing behaviour in AIS data by Haskins et al. (2024). If the current research was to be done again, it would be better to flip the methodology around. So, use the current validation method to identify trips that need to degas. Inspect those trips to clearly identify what kind of behaviour is related to degassing. Based on these results, create a model that identifies degassing behaviour based on AIS and IVS data. By still creating a model that identifies vessels that are degassing based on vessel movement, the incompleteness and inaccuracy of the IVS data is bypassed. To further improve the model’s precision, external data can be incorporated, such as the type of industry surrounding the harbour or the number of buildings near a specific section of the waterway. This additional information could help correctly classify a vessel performing degassing activities and a bunker vessel refuelling another vessel. Another approach to identify bunker vessels is to analyse AIS data for instances where a suspected bunker vessel was sailing parallel to another vessel, which may indicate bunker activity.

The research of Haskins et al. (2024) was the basis for this research as it was the first exploration of identifying degassing activity with AIS data. The approach used by Haskins et al. (2024) was different from the method presented in this research, but in both research projects, similar kinds of patterns that could be related to degassing have been detected. Unfortunately, the problems related to the data quality of the AIS data mentioned by Haskins et al. (2024) could not be resolved in this research.

Haskins et al. (2024) only researched the AIS data of vessel movements between Amsterdam and Rotterdam of 1 month. The aim of this research was to increase the data amount, not only temporally, but also geographically. The aim was to use the AIS data of the full year of 2022 for the whole Netherlands and half a year of AIS data of 2024 for the whole Netherlands. However, due to the big size of these files and the first time at Rijkswaterstaat such a big file needed to be pseudonymized internally, not all the data could be delivered and the data was delivered later than expected. Besides, due to a miscommunication somewhere in the process, the first delivered files were only for the vessel movement between Amsterdam and Rotterdam. This was later clarified, but the data for the whole Netherlands could not be delivered on time, due to a significant data size increase. Unfortunately, in the end only one month of data of 2024 between Amsterdam and Rotterdam was used as input for the research.

Haskins et al. (2024) also mentioned the limitations of AIS and IVS data due to quality issues. Because some of the data, especially the IVS data, is reliant on human input, it is prone to human error or deliberate inaccuracy. A good example is the difference between the identified ship type in AIS data, IVS data and the type of cargo that was transported. The ship type identified in the IVS data did not always match with the type of cargo that was transported, which was also identified in the IVS data. Especially for less common ship types, this inaccuracy was high. After joining the two datasets on a pseudonymized vessel identity, the same kind of pattern could be found between the AIS and IVS data. Not every tanker vessel in the AIS data was a tanker vessel in the IVS data, and vice versa. This was one of the inaccuracies that could be explicitly checked within the data. Other inaccuracies, like the specified cargo weight or the cargo type are more difficult to validate.

Inaccuracy is also related to another limitation of this research, as, due to the only recently introduced ban on degassing, no validation dataset was present to validate the results of the research. Some alternative methods were proposed to perform a validation check, but due to the lack of data and the time constrains, the results could not be validated as thoroughly. Besides, the proposed methods are not the best suited validation methods for this type of research, but were the best option at this time.

Further research in this topic is needed to establish a better way to validate the results. For example, the e-nose network around the waterways, for which an increase in coverage is planned, can give valuable data that could be used to validate the results. Alternatively, when, due to the ban, more fines are being registered by the ILT, a bigger dataset of 'ground truth' with location and time data can be used to validate the results. A better validation method further outlines the limitations of the current analysis method and could give valuable insight into optimizing the current proposed analysis method. An optimized method could better identify degassing behaviour and can give better insight into the locations where degassing activities take place. This could also be achieved by increasing the input data, both on temporal and geographical scale. Especially the geographical scope on the AIS data resulted in a lot of uncertainty in the results. These are the main limitations that have to be solved in future research in detecting degassing patterns with AIS and IVS data.

6 Conclusion

This research aimed to identify hotspots for illegal degassing activities on Dutch inland waterways using historic IVS and AIS data and to estimate the quantities of gas emitted through these activities. The formulated research question was as follows:

> What sections of the Dutch inland waterway can be identified as 'hotspots' for illegal degassing activities based on historic IVS and AIS data, and what are the estimated quantities of gas emitted with these activities?

With the current method and limited data available, this question cannot be answered. The question can be answered for the waterways between Amsterdam and Rotterdam, but this answer is based on multiple assumptions. This is reflected in the validation of the method. Based on the confusion matrix, the indicated precision of this method is 63.8%, while the indicated recall is only 42.1%. However, the insights created in the research can be valuable for the ILT and Rijkswaterstaat, as it outlines both the potentials and the limitations of using AIS and IVS data to identify degassing activity. Besides, the methodology to detect deviating behaviour is developed, which includes the joining of the datasets, the splitting of the data in separate trips and the detection of deviating behaviour. The current results indicate that there are four different types of behaviour related to degassing, as discussed in Section 4.2. And, based on the hotspot analysis in Section 4.3, the conclusion can be made that less populated areas close to a major harbour are preferred for degassing activities. A clear estimate of the amount of emission related to degassing could not be made with the current method and limited data quantity.

More time is needed to further finetune the methodology and to discuss the intermediate results with experts in this field of expertise, as the methodology did shows some potential. There were, for example, some clear examples that were highly likely to be related to degassing, which were also classified as degassing. Besides, the identified deviating behaviour matches with some of the mentioned degassing behaviour by Haskins et al. (2024). Not all the mentioned behaviour was as clearly visible in the February 2024 dataset as shown in the research of Haskins et al. (2024), but some of the identified behaviour, especially the round trip behaviour, was also clearly present in the February 2024 dataset and was identified as deviating behaviour by the proposed method. Besides, an extra type of degassing behaviour was identified in this research, on top of the already identified behaviour by Haskins et al. (2024). An important note to make here is that, due to a recall of only 42.1%, only around two-fifths of the actual vessels that needed to degas have been identified correctly. This means that three-fifths of the vessels that needed to degas did not deviate from the optimal route more than 25%. This gives a clear indication of the difficulty of identifying degassing behaviour purely based on movement data.

As well as extra time, a complete dataset is also required to better test the proposed method. The currently available data, that was used as input, was inaccurate and incomplete, as a considerable part of the data was cut off and left out of the dataset due to the geographical scope that was applied on the data. To get a better insight into the results of the method and its validity, the method needs to be applied on a complete dataset for the whole of the Netherlands to remove some of the uncertainties that were currently present.

There are, however, several uncertainties that cannot be removed with the inclusion of a complete dataset. Haskins et al. (2024) already mentioned in their research several problems with the data quality of the AIS data. The IVS data also has problems with the data quality. It is partially reliant on human input. Data quality problems are challenging to deal with, as removing incomplete or inaccurate data is undesirable as this can remove relevant information. Filling in this information is nearly impossible as this would require a lot of effort for every record. This is also clearly stated in the research of Emmens et al. (2021). Besides, it is still unclear what exactly classifies as degassing behaviour and what the difference is between the movement of a vessel that is bunkering and a vessels that is degassing. A good example of this can be found in Section 4.3 on the validation of the results, where even the inspector of the ILT is doubting if a vessel was degassing or if it was a bunker vessel. Besides, a recall of less than 50% indicates that more than half of the actual degassing behaviour could not be identified by a minimum deviation from the optimal route of 25%. Due to these uncertainties, it will be challenging to reach a high precision, high recall and produce significant results with this method and thus should mainly be used to get more insight into the degassing behaviour. Methods like the e-nose network have more potential to be used as a monitoring tool. Not only due to precision, but also due to the privacy regulations around AIS and IVS data.

Ultimately, this research underlines the complexity of detecting degassing activities based on vessel movement data. While the whole process and the developed methodology offer some valuable insights, more time is needed to further develop the proposed method. The currently proposed method could not be tested sufficiently due to the lack of input data and time. For further validation of the results, more input data is needed. However, Rijkswaterstaat will need more time to handle the increasing size of the datasets when the temporal or geographical scope is increased. The research has also contributed to topics outside the direct research scope, such as the connection between Rijkswaterstaat and the ILT, the AIS data delivery process of Rijkswaterstaat and the process of joining the IVS data to the AIS data. Besides, some valuable insights, which should be taken into consideration in future research, can still be taken from this research and the current provided methodology and scripts can be used on other or bigger datasets. Future research can also further optimize these methods and scripts to increase its accuracy, as the current method shows potential of detecting degassing activity from AIS and IVS data.

References

Arakawa, T., Iitani, K., Toma, K., & Mitsubayashi, K. (2023). Biosensors: Gas Sensors. In R. Narayan (Ed.), *Encyclopedia of Sensors and Biosensors: Volume 1-4* (First Edition, Vol. 1, pp. 478–504). Elsevier. https://doi.org/10.1016/B978-0-12-822548-6.00066-2

Autoriteit Persoonsgegevens. (n.d.). *De AVG in het kort*. Retrieved December 18, 2024, from https://www.autoriteitpersoonsgegevens.nl/themas/basis-avg/avg-algemeen/de-avg-in-het-kort

Banerjee, A., Chitnis, U. B., Jadhav, S. L., Bhawalkar, J. S., & Chaudhury, S. (2009). Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal*, *18*(2), 131. https://doi.org/10.4103/0972-6748.62274

Bast, H., Delling, D., Goldberg, A., Müller-Hannemann, M., Pajor, T., Sanders, P., Wagner, D., & Werneck, R. F. (2016). Route Planning in Transportation Networks. In P. Kliemann Lasse and Sanders (Ed.), *Algorithm Engineering: Selected Results and Surveys* (pp. 19–80). Springer International Publishing. https://doi.org/10.1007/978-3-319-49487-6_2

Bolt, E. (2003). *EMS-protocol Emissies door Binnenvaart: Ontgassing van ladingdampen naar lucht*.

Breed, M. (2021). *Scalability of Customizable Route Planning* [Utrecht University]. https://studenttheses.uu.nl/handle/20.500.12932/41294

Buck, A. de, Hoen, M. ’t, & Boer, E. den. (2013). *Update estimate emissions degassing inland tank vessels*. www.cedelft.eu

Cai, W., Zheng, Y., Shi, Y., Zhao, N., & Zhang, M. (2011). Comparing study of route planning algorithms based on hierarchical strategy. *International Conference on Management and Service Science, MASS 2011*. https://doi.org/10.1109/ICMSS.2011.5998605

Cao, W., Shi, H., Zhu, S., & Zhu, B. (2009). Application of an Improved A* Algorithm in Route Planning. *2009 WRI Global Congress on Intelligent Systems*, *1*, 253–257. https://doi.org/10.1109/GCIS.2009.76

CDNI. (n.d.). *Tenuitvoerlegging - CDNI*. Retrieved October 8, 2024, from https://www.cdni-iwt.org/tenuitvoerlegging/?lang=nl

CDNI. (2019). *CONVENTION ON THE COLLECTION, DEPOSIT AND RECEPTION OF WASTE GENERATED DURING NAVIGATION ON THE RHINE AND OTHER INLAND WATERWAYS* (CDNI Secretariat, Ed.). CDNI: Central Commission for the Navigation of the Rhine. www.cdni-iwt.org

CDNI. (2023). *VERDRAG INZAKE DE VERZAMELING, AFGIFTE EN INNAME VAN AFVAL IN DE RIJN EN BINNENVAART*.

CDNI. (2024). *Part B – Degassing/Venting - CDNI*. https://www.cdni-iwt.org/part-b-degassing/?lang=en

Chen, J., Huang, H., Rui, Y., Pu, Y., Zhang, S., Li, Z., & Wang, W. (2024). Enhancing microseismic/acoustic emission source localization accuracy with an outlier-robust kernel density estimation approach. *International Journal of Mining Science and Technology*, *34*(7), 943–956. https://doi.org/10.1016/J.IJMST.2024.07.005

Cromley, E. K., & Mclafferty, S. L. (2002). *GIS and public health* (Vol. 9). Guilford Press.

David, E., & Niculescu, V.-C. (2021). Volatile Organic Compounds (VOCs) as Environmental Pollutants: Occurrence and Mitigation Using Nanomaterials. *Public Health*, *18*, 13147. https://doi.org/10.3390/ijerph182413147

Delling, D., Goldberg, A. V., Pajor, T., & Werneck, R. F. (2011). Customizable Route Planning. In S. Pardalos Panos M. and Rebennack (Ed.), *Experimental Algorithms* (pp. 376–387). Springer Berlin Heidelberg.

Delling, D., Sanders, P., Schultes, D., & Wagner, D. (2009). *Engineering Route Planning Algorithms*.

Delmelle, E. M., Zhu, H., Tang, W., & Casas, I. (2014). A web-based geospatial toolkit for the monitoring of dengue fever. *Applied Geography*, *52*, 144–152. https://doi.org/10.1016/J.APGEOG.2014.05.007

(DHPC) Delft High Performance Computing Centre. (2024). *DelftBlue Supercomputer (Phase 2)*. https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2

EFOA. (2008). *The European Fuel Oxygenates Association Guidelines MTBE / ETBE Transport over inland waterways*.

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

Erol, A., & Arcuri, A. (2023). *Floating Degassing in the Netherlands: Rights and Obligations under International Law*.

European Commission. (n.d.). *Legal framework of EU data protection*. Retrieved December 18, 2024, from https://commission.europa.eu/law/law-topic/data-protection/legal-framework-eu-data-protection_en

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Optimization clustering techniques. In *Cluster Analysis* (5th ed.). John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/uunl/detail.action?docID=661789

Geerlings, H., & Kuipers, B. (2019). ONTGASSEN DOOR DE BINNENVAART; EEN ZAAK MET EEN LUCHTJE. *TIJDSCHRIFT LUCHT*, *4*, 22–25.

Golshani, F., Cortes-Rello, E., & Ahluwalia, S. (1992). RUTA-100: A dynamic route planner. *International Conference on Tools with Artificial Intelligence, ICTAI*, *1992-November*, 418–423. https://doi.org/10.1109/TAI.1992.246440

Haskins, V., Reith, M., Van Krieken, L., Wolf, J., Bui, D., & Bongers, N. (2024). *Automated Detection of Illegal Degassing*.

Johner, R., Lanaia, A., Dornberger, R., & Hanne, T. (2022). Comparing the Pathfinding Algorithms A*, Dijkstra's, Bellman-Ford, Floyd-Warshall, and Best First Search for the Paparazzi Problem. In H. and B. K. and K. J. H. and B. J. C. Saraswat Mukesh and Sharma (Ed.), *Congress on Intelligent Systems* (pp. 561–576). Springer Nature Singapore.

Kamilaris, A., & Ostermann, F. O. (2018). Geospatial analysis and the internet of things. *ISPRS International Journal of Geo-Information*, *7*(7). https://doi.org/10.3390/IJGI7070269

Koedijk, drs. O. C. (2020). *Richtlijnen Vaarwegen 2020*. https://binnenvaartcijfers.nl/wp-content/uploads/2020/10/richtlijnen-vaarwegen-2020-def-webversie.pdf

Koop, K. (2016). *Effects of future restrictions in degassing of inland tanker barges*.

Leigh, R., Louis, S. J., & Miles, C. (2007). Using a genetic algorithm to explore A*-like pathfinding algorithms. *Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Games, CIG 2007*, 72–79. https://doi.org/10.1109/CIG.2007.368081

Leshem, S., & Trafford, V. (2007). Overlooking the conceptual framework. *Innovations in Education and Teaching International*, *44*(1), 93–105. https://doi.org/10.1080/14703290601081407

Lewis, H., & Brown, M. (2001). A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, *22*, 3223–3235. https://doi.org/10.1080/01431160152558332

Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, *36*(2), 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2

Loefen, L. M. H. (2017). *Reducing benzene emissions by degassing to the atmosphere in a transport network of a petrochemical company*. Eindhoven University of Technology.

Ming-Tso Chiang, M., & Mirkin, B. (2010). *Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads*. https://doi.org/10.1007/s00357-010

Ministry of Infrastructure and Water Management (2023). https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2023D20902&did=2023D20902

Mondal, B., Mukherjee, T., Finch, N. W., Saha, A., Gao, M. Z., Palmer, T. A., & DebRoy, T. (2022). Vapor Pressure versus Temperature Relations of Common Elements. *Materials*, *16*(1). https://doi.org/10.3390/MA16010050

Nanayakkara, S., Srinivasan, D., Lup, L., German, X., Taylor, E., & Ong, S. (2007). *Genetic Algorithm based route planner for large urban street networks*. 4469–4474. https://doi.org/10.1109/CEC.2007.4425056

Nha, V. T. N., Djahel, S., & Murphy, J. (2012). A comparative study of vehicles' routing algorithms for route planning in smart cities. *2012 First International Workshop on Vehicular Traffic Management for Smart Cities (VTM)*, 1–6. https://doi.org/10.1109/VTM.2012.6398701

Plotting Pitfalls — Datashader. (2024, July 4). https://datashader.org/user_guide/Plotting_Pitfalls.html

Rijkswaterstaat. (n.d.). *IVS Next*. Retrieved September 17, 2024, from https://www.rijkswaterstaat.nl/water/scheepvaart/scheepvaartverkeersbegeleiding/ivs-next

Rijkswaterstaat. (2024a). *Handleiding: Invoeren en aanmelden van een reis | BICS-website*. https://www.bics.nl/nl/node/47

Rijkswaterstaat. (2024b). *PROJECT BEVAARBAARHEID IJSSEL*.

Rijkswaterstaat Environment. (n.d.). *Volatile Organic Compounds (VOC) - Rijkswaterstaat Environment*. Retrieved September 13, 2024, from https://rwsenvironment.eu/subjects/air/volatile-organic/

RIVM. (n.d.). *Luchtverontreinigende stoffen deeltjes- en gasvormig | RIVM*. Retrieved November 26, 2024, from https://www.rivm.nl/ggd-richtlijn-medische-milieukunde-luchtkwaliteit-en-gezondheid/blootstelling-aan-luchtverontreiniging/stoffen-deeltjes-en-gasvormig

Schultes, D. (2008). *Route planning in road networks*. Karlsruhe, Univ., Diss., 2008.

Sheather, S. J. (2004). Density Estimation. *Statistical Science*, *19*(4), 588–597. http://www.jstor.org/stable/4144429

Sigrid, L., & Leisner, M. (2021). *Proposed Volatile Organic Compounds (VOC) Control Measures Under MARPOL Annex VI*.

Silverman, B. W. (1998). Density estimation for statistics and data analysis. In *Density Estimation: For Statistics and Data Analysis* (1st ed.). Routledge. https://doi.org/10.1201/9781315140919/DENSITY-ESTIMATION-STATISTICS-DATA-ANALYSIS-BERNARD-SILVERMAN/ACCESSIBILITY-INFORMATION

Singh, P., Singh, N., Singh, K. K., & Singh, A. (2021). Chapter 5 - Diagnosing of disease using machine learning. In K. K. Singh, M. Elhoseny, A. Singh, & A. A. Elngar (Eds.), *Machine Learning and the Internet of Medical Things in Healthcare* (pp. 89–111). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-821229-5.00003-3

Stuurgroep Varend Ontgassen. (2023). *Roadmap varend ontgassen*. https://www.rijksoverheid.nl/documenten/rapporten/2023/12/04/bijlage-2-roadmap-varend-ontgassen-periode-2023-2024

Swaen, B., & George, T. (2024, September 5). *What Is a Conceptual Framework? | Tips & Examples*. https://www.scribbr.com/methodology/conceptual-framework/

The World Bank Group. (2021). *Netherlands - Climatology | Climate Change Knowledge Portal*. https://climateknowledgeportal.worldbank.org/country/netherlands/climate-data-historical

United Nations. (2011). *European Agreement concerning the International Carriage of Dangerous Goods by Inland Waterways (ADN)*.

VNCI, Vemobin, Vereniging van Nederlandse tankopslagbedrijven, & Koninklijke Binnenvaart Nederland. (2024). *Approaching ban on barge degassing*.

Walt, S. van der, & Smith, N. (n.d.). *matplotlib colormaps*. Retrieved December 9, 2024, from https://bids.github.io/colormap/

Wu, Q., & Shan, J. J. (2000). The application of genetic algorithm in GIS network analysis. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *33*, 1184 – 1191. https://www.scopus.com/inward/record.uri?eid=2-s2.0-71049152069&partnerID=40&md5=933fcb168ef162ec3546a8a29028d1a5

Yadav, R. K., Kishor, G., Himanshu, & Kashyap, K. (2020). Comparative Analysis of Route Planning Algorithms on Road Networks. *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 401–406. https://doi.org/10.1109/ICCES48766.2020.9137965