Data-Driven Historical Evaluation and Prediction of Fuel Consumption in Inland Vessels Employing Autonomous Lane Assist

Giacomo Carachino

Delft University of Technology

TU Delft

Data-Driven Historical Evaluation and Prediction of Fuel Consumption in Inland Vessels Employing Autonomous Lane Assist

by

Giacomo Carachino

to obtain the degree of Master of Science

at the Delft University of Technology,
Faculty of Mechanical Engineering,
track of Multi-Machine Engineering

to be defended publicly on Wednesday May 13, 2026 at 15:00.

Student number:	6068316
Responsible Supervisor:	Frederik Schulte
Supervisor:	Dr.ir. Mahnam Saeednia
Presentation date:	13 May 2026
Faculty:	Faculty of Mechanical Engineering, Delft
Master track:	Multi-Machine Engineering

Cover: Container Inland ship, from https://www.portofrotterdam.com/en/logistics/connections/intermodal-transportation/inland-shipping

TU Delft

Data-Driven Historical Evaluation and Prediction of Fuel Consumption in Inland Vessels Employing Autonomous Lane Assist

Giacomo Carachino

May 5, 2026

Report number: 2026.MME.9187

Contents

1 Introduction

Inland shipping is a critical mode of freight transport in Europe, offering a cost-effective and sustainable alternative to road and rail. However, as sustainability targets grow stricter and operational costs rise, optimizing the energy efficiency of inland vessels becomes increasingly important. Autonomous and assisted navigation technologies offer promising avenues for improving efficiency and reducing emissions. Among these innovations, Shipping Technology's Autonomous Lane Assist (ST-ALA) system stands out as a concrete implementation of semi-autonomous sailing, actively guiding rudder movements to follow optimal trajectories. **This thesis aims to quantify the fuel savings that can be attributed to the use of ST-ALA.**

Inland waterway transport in Europe is rapidly evolving towards greater automation and improved fuel efficiency. This shift is driven by several converging factors.

Environmental and climate targets are a major motivator, the inland shipping sector has committed to steep emission reductions (for example 35 to 50% CO_2 reduction by 2035 in the Netherlands) with an ultimate goal of zero emissions by 2050, in line with EU climate neutrality objectives. Meeting these green goals requires cutting fuel consumption and adopting cleaner, smarter technologies.

At the same time, the industry faces a shortage of skilled boatmen and crew, as the European inland fleet struggles to replace an aging workforce. Automation is seen as a key solution to labor shortages, allowing fewer or remote operators without compromising safety.

Finally, the broader digital transformation of transport is enabling these changes – by harnessing real-time data, connectivity, and AI, inland shipping can optimize voyages, avoid delays, and reduce unnecessary fuel burn.

Shipping Technology, founded in 2018, is a Dutch startup leading innovation in inland shipping through its digital hardware and software suite. Its core product, the ST-BRAIN, is already installed on over 250 vessels and collects high-frequency nautical data to enable real-time monitoring, assisted navigation, and autonomous control. The ST-ALA module is the first operational step toward full autonomy, automatically managing the rudder to follow an optimal route computed using a data driven approach. The underlying hypothesis is that such automation results in smoother steering, less rudder movement, and consequently lower fuel consumption. However, due to the high variability between trips, ranging from vessel characteristics, cargo load, water conditions, and human behavior, this claim must be rigorously validated.

Further discussion on the impact of this research is presented in Section 6

2 Problem Definition

Quantifying the fuel savings resulting from ST-ALA is challenging because of the complexity and variability of real-world sailing conditions. Factors such as draft, cargo weight, water levels, flow velocity, and captain behavior all influence fuel consumption. Many of these parameters are difficult to control or measure directly. This study proposes to develop a robust data-driven approach that can isolate the effect of ST-ALA from other sources of variation.

Furthermore this study aims at developing a data-driven approach (through ML modeling) to predict fuel consumption of inland vessels using only data available pre-departure. This predictive model could in the future be used for various fuel consumption optimization purposes.

To overcome this, the project proposes to model fuel consumption per river segment (or graph edge), using historical data collected at a 1 Hz frequency. By training machine learning models for specific river sections, and comparing similar trajectories with and without ST-ALA activation, the aim is to estimate the efficiency gain in a statistically meaningful way.

3 Research Objectives

The primary objective of this research is:

To develop a method to predict the fuel consumption of river journeys by inland vessels and use this model to quantify the fuel savings attributable to ST-ALA.

This thesis addresses its core objective through a **two-part investigation**:

1. **Historical Analysis** — Evaluate fuel savings due to the ST-ALA system using historical voyage data.

   This refers to the need of estimating **causal inference** from the binary variable ST-ALA on/off on the continuous variable "fuel consumption". Specifically the objective is to make a **counterfactual prediction** of fuel consumption with ST-ALA turned off versus turned on.

2. **Predictive Modeling** — Develop a predictive model for voyage-level fuel consumption based solely on data available at the time of departure.

To achieve this, several sub-objectives must be fulfilled:

- Identify the parameters that most influence fuel consumption.
- Select and preprocess relevant data from the ST-BRAIN database.
- Develop and validate predictive models of fuel consumption per river segment.
- Isolate the influence of ST-ALA activation within these models.
- Benchmark the model's predictions against existing frameworks (GLEC [17]).

4 Research Question

Main research question:

How much fuel is saved when an inland vessel sails using the Autonomous Lane Assist system compared to conventional steering methods?

Sub-questions:

1. Which variables most significantly affect fuel consumption in inland shipping?
2. How can historical navigation and fuel data be structured to compare between ST-ALA and non-ST-ALA trips?
3. Under what conditions does the ST-ALA cause the most fuel savings?
4. What is the most appropriate modeling approach to predict fuel consumption per river segment?
5. How can the fuel consumption prediction models be used to predict CO2 emissions?

5 Research Approach

Fuel consumption is dependent on a number of factors, including engine and rudder utilization, ship and engine type, current or speed through water, water levels, keel clearance, draft (immersion), and weather conditions.

Some of these parameters are directly measured by the ST-BRAIN system, while others are not. By predicting fuel consumption per river section, these unmeasured variables can be **implicitly accounted for through data aggregation**. This aggregation step is crucial, as it allows the models to indirectly learn river-specific characteristics,such as curvature, current strength, or navigational complexity,that are not explicitly recorded in the dataset but are reflected in aggregated traversal statistics. Only a few prior studies in the literature have adopted such segmentation-based aggregation, which is one of the key methodological innovations of this work.

Shipping Technology has divided inland waterways into graph-based sections. Routes can be planned from A to B, and the river is divided into "edges" (graph segments). These edges serve as the natural units for data aggregation and model training. A proper definition of "edge"in this context and the available dataset are described in Section 7.

This research follows an **exploratory and iterative approach**. Rather than committing to a fixed modeling pipeline, the process is structured as a cycle of experimentation, evaluation, and refinement. Each iteration contributes both to answering the research questions and to improving the understanding of the fuel consumption dynamics of inland vessels.

5.1 Historical Evaluation

The objective "Evaluate fuel savings due to the ST-ALA system using historical voyage data" (Section 3) refers to estimating the **causal effect** of the binary variable `ALA_on/off` on the continuous outcome `fuel_consumption`. The fundamental goal is to compute a **counterfactual prediction**: how much fuel a vessel would have consumed had ALA been active versus inactive during the same traversal.

To achieve this, two complementary causal inference approaches were used: **G-computation** and **Inverse Probability Weighting (IPW)**. Both are grounded in the potential outcomes framework and are well suited for observational data with rich confounding structure.

G-computation. G-computation [8, 5], originally proposed by Robins (1986), estimates causal effects by modeling the outcome conditional on treatment and covariates, and then simulating counterfactual outcomes under alternative treatment assignments. This method is particularly appropriate in this study because:

- it allows the use of flexible, non-parametric ML models to approximate the conditional expectation $E[Y \mid A, L]$;
- it preserves a transparent causal interpretation through the g-formula;

- the treatment (ALA activation) is effectively static within each traversal;
- the dataset contains a large and diverse set of measured confounders (operational, hydrodynamic, and environmental).

In practice this will be implemented by:

- Fitting a predictive model of fuel consumption on historical traversals using all relevant covariates, including vessel characteristics, hydrodynamics, loading, environmental conditions, and ALA status.
- Evaluating the model twice on every traversal: once with `ALA = ON` and once with `ALA = OFF`. This produces two counterfactual fuel-consumption predictions per row.
- The difference between these two counterfactual outcomes is interpreted as the estimated individual-level treatment effect, which is then averaged to obtain global and device-level ATEs.

Inverse Probability Weighting (IPW). As a robustness check, IPW was applied in parallel. A propensity score model was first trained to estimate

$$e(L) = P(A = 1 \mid L),$$

representing the probability of ALA activation given the observed covariates. Each observation was then weighted by the inverse probability of receiving its observed treatment. Under correct specification of the treatment model, IPW provides an unbiased estimate of the ATE even if the outcome model is misspecified. This allowed triangulation of causal estimates and helped verify that the inferred ALA effect was not an artifact of a single modeling pipeline.

Conditional Average Treatment Effect (CATE). In addition to global ATE estimation, this study also estimated **CATEs** to identify *under what operational conditions ALA yields the largest fuel savings*. Continuous variables (`avg_engine_rpm`, `avg_engine_load`, `med_depth`, `total_mass`) were discretized into percentile-based bins, and categorical features (`environment`, `ship_type`) were considered directly. This partitioned the feature space into several hundred regions, each representing a distinct operational regime. Within each region, savings were averaged to estimate the CATE.

Summary. The combination of G-computation (outcome modeling), IPW (treatment modeling), and CATE (effect heterogeneity) provides a robust and comprehensive evaluation of the causal impact of ST-ALA on fuel consumption. A detailed implementation of the g-computation pipeline, model specification, diagnostics, and CATE construction is provided in Section 8.1.5.

5.2 Pre-departure Prediction

A separate predictive task focuses on estimating fuel consumption *before* the voyage begins, using only information available at departure, meaning planned route, vessel characteristics, loading condition, and expected water level. Two model families were used for this purpose.

First, a **CatBoost decision-tree model**, consistent with the one employed for the g-computation outcome model, was trained to provide a strong tabular baseline with robust handling of categorical and nonlinear interactions.

Second, a **Deep Backpropagation Neural Network (DBPNN)** was implemented following the architecture presented by Yuan et al. (2020) [21], who showed that multilayer feedforward neural networks can effectively capture nonlinear hydrodynamic relationships affecting inland-vessel fuel consumption. This DBPNN serves as a complementary benchmark representing the class of neural models commonly used in the maritime prediction literature.

Together, these two approaches allow comparison between a state-of-the-art gradient-boosted tree model and a neural-network baseline tailored to ship-fuel prediction.

5.3 Benchmarking

To contextualize the results and evaluate robustness, two benchmarking strategies are used:

- **Means Analysis:** A naive benchmark comparing average fuel consumption per km for ALA-on versus ALA-off traversals. This approach mimics the informal evaluations typically performed in practice,running a few voyages with and without ALA and comparing averages. Such an approach ignores the large variance and most importantly the confounding introduced by environmental and operational factors, illustrating how simple comparisons can lead to misleading conclusions. This serves as a contrast to the causal model.
- **Linear Regression Benchmark:** For the predictive models, a simple linear regression is used as a baseline due to its interpretability and common use in applied modeling. It provides a transparent reference point against which more flexible models (e.g., gradient-boosted trees) can be evaluated.

Continuous Refinement and Model Monitoring Model performance is continuously monitored through:

- **Residual analysis:** Residual patterns are examined to detect potential model misspecification. For example, if residuals systematically vary with `environment` (upstream, downstream or tidal) or other features, it suggests unmodeled interactions or missing variables.
- **Feature importance analysis:** The contribution of individual features to model predictions is tracked to ensure interpretability and model stability. For linear models,

importance is inferred from the magnitude of model coefficients. For flexible machine learning models (e.g., Decision tree models), **SHAP (SHapley Additive exPlanations)** values are used to quantify each feature's marginal effect on the predicted fuel consumption.

Whenever systematic residual patterns or unstable feature importances are detected, new features are engineered, subsets of data are re-evaluated, or model types are reconsidered. This iterative refinement continues until both the predictive and causal models exhibit satisfactory performance and robustness.

6 Literature Review

A brief Literature Review was conducted to better understand how this research fits in the broader landscape of fuel consumption estimation in shipping.

6.1 Causal Inference estimation methods

The aim of this subsection is to inform the methodological choices made to estimate **causal inference**.

Causal inference is defined as the estimation of the effect that a given treatment or intervention would have on an outcome on a given population, had it been applied, compared to the same outcome had it not been applied.

In the context of this research, the relation to the terms *treatment*, *outcome* and *population* is straightforward:

- *treatment* corresponds to the activation (*administration*) of the Autonomous Lane Assist (ALA) system;
- *population* corresponds to the entirety of the fleet for which Shipping Technology has operational data;
- *treatment effect* being estimated is the causal effect of ALA activation on fuel consumption.

Let:

- A denote the treatment variable (for example, $A = 1$ when the treatment is applied and $A = 0$ when it is not),
- Y denote the outcome of interest

The notation Y^a represents the *potential outcome* that would be observed if the treatment A were set to value a. Since for each observational unit only one of these two potential outcomes can be observed(in the case of binary treatment like this one, more information in

section 7), causal inference focuses on comparing the expectations of these counterfactual outcomes across the population. Formally, the average treatment effect (ATE) is defined as:

$$ATE = E[Y^{a=1}] - E[Y^{a=0}]$$

where $E[\cdot]$ denotes the expected value (population mean).
This formulation expresses the difference in the expected outcome if all units in the population were exposed to the treatment ($A = 1$) versus if all were unexposed ($A = 0$) [8].

Among the methodological frameworks available for estimating causal effects, [5] presents and empirically compares three main families of estimators, collectively known as the *g-methods*: **g-computation**, **inverse-probability weighting (IPW)**, and **targeted maximum likelihood estimation (TMLE)**. These methods were originally developed by Robins and colleagues to address complex confounding structures and feedback mechanisms that invalidate standard regression approaches.

G-computation models the joint distribution of the observed data and uses it to simulate potential outcomes under alternative exposure (treatment) scenarios. It provides consistent estimates when the model for the outcome is correctly specified or estimated with sufficiently flexible machine-learning algorithms.
Inverse-probability weighting instead constructs a pseudo-population in which the treatment is independent of confounders, by weighting each observation by the inverse of its estimated probability of receiving the treatment actually observed.
Targeted maximum likelihood estimation (TMLE) combines both strategies, providing a doubly robust estimator that remains consistent if either the treatment or outcome model is correctly specified.

The comparative analysis in [5] demonstrates that g-computation performs particularly well when the outcome model is well specified or estimated using flexible learning techniques, conditions that can be reasonably achieved in the present study due to the abundance of high-frequency data available from the ST-BRAIN system.

Following the formal definition of causal effect given by [8], the objective is thus to estimate the expected difference in fuel consumption if all vessels had sailed with ALA active versus if all had sailed without it.

Given its interpretability, theoretical soundness, and strong empirical performance under flexible model specifications, **g-computation is adopted as the method of choice** for causal inference in this work. A detailed description of its implementation is presented in Chapter 8.1.5.

6.2 Statistical Inference for the Estimated ATE

After estimating the average treatment effect (ATE) via g-computation, it is necessary to quantify the statistical significance of the estimated fuel-consumption reduction. Because the data exhibit strong clustering at the vessel level, and because the g-computation estimator is a plug-in estimator built on a machine-learned outcome model, classical confidence intervals

would be inappropriate. This subsection reviews the modern inference procedures used to obtain statistically valid confidence intervals and p-values. Full theoretical development is delegated to the original authors, and only the elements relevant to the present study are summarized.

Plug-in ATE Estimator Given unit-level estimated savings s_i, obtained from g-computation, the plug-in estimator of the ATE is simply the sample average:

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} s_i.$$

This estimator is consistent under standard causal-identification assumptions and serves as the baseline input for all subsequent inference procedures.

Influence-Function Based Standard Errors Hansen and Overgaard (2024) [7] provide a complete treatment of variance estimation for g-computation estimators. They show that $\widehat{\tau}$ admits an asymptotic linear representation, which implies that its variance can be estimated using the empirical influence function

$$\widehat{\varphi}_i = s_i - \widehat{\tau}.$$

The corresponding variance estimator is then

$$\widehat{\mathrm{Var}}_{\mathrm{IF}}(\widehat{\tau}) = \frac{1}{n^2} \sum_{i=1}^{n} \widehat{\varphi}_i^2.$$

This is the recommended state-of-the-art method for uncertainty quantification of g-computation estimates. Its implementation in this thesis follows the authors' guidance directly.

Non-Parametric Bootstrap Although Hansen and Overgaard [7] note that bootstrap procedures can be computationally intensive or unstable when too few replicates are used, they remain widely used for model-agnostic inference. In the non-parametric bootstrap, one repeatedly resamples indices with replacement and recomputes the ATE:

$$\widehat{\tau}^{(b)} = \frac{1}{n} \sum_{i \in \mathcal{I}^{(b)}} s_i, \qquad b = 1, \ldots, B.$$

A bootstrap variance estimator then follows as

$$\widehat{\mathrm{Var}}_{\mathrm{boot}}(\widehat{\tau}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\tau}^{(b)} - \overline{\tau} \right)^2.$$

Here, the bootstrap is used as a robustness check on the stability of the influence-function estimates.

Cluster-Robust Inference Because many traversals originate from the same vessel, observations within vessels cannot be treated as independent. Standard errors that ignore clustering would be severely underestimated. Cluster-robust methods are the standard remedy. As summarized by MacKinnon, Nielsen, and Webb (2023) [14], cluster-robust variance estimation aggregates the influence-function contributions within each cluster:

$$\Phi_g = \sum_{i \in \mathcal{I}_g} \widehat{\varphi}_i,$$

leading to the cluster-robust estimator

$$\widehat{\mathrm{Var}}_{\mathrm{CR}}(\widehat{\tau}) = \frac{1}{n^2} \sum_{g=1}^{G} \Phi_g^2.$$

This estimator remains valid under arbitrary within-vessel correlation, which matches the structure of the dataset in this thesis.

Wild Cluster Bootstrap When the number of clusters G is modest common in transportation datasets—cluster-robust standard errors alone may be unreliable. MacKinnon et al. [14] recommend the *wild cluster bootstrap* (WCR) as the preferred method for reliable small-G inference. For each bootstrap replicate, cluster-level perturbation weights $w_g^{(b)} \in \{-1, +1\}$ are drawn, and the perturbed ATE estimator is computed as

$$\widehat{\tau}^{(b)} = \frac{1}{n} \sum_{g=1}^{G} \sum_{i \in \mathcal{I}_g} w_g^{(b)} \, s_i.$$

6.3 Impact of Automated Navigation Systems on Fuel Consumption

Automated navigation and rudder control systems have been increasingly adopted in both seagoing and inland vessels to enhance steering precision and improve fuel efficiency. The underlying mechanism is straightforward: every rudder movement generates hydrodynamic resistance that the propulsion system must overcome. Minimizing unnecessary rudder angles therefore directly reduces total drag and, consequently, fuel consumption.

According to the *American Bureau of Shipping Energy Efficiency Advisory* [1], although resistance due to steering typically accounts for only a small fraction of total hull resistance, avoiding excessive rudder deflections can still reduce overall fuel consumption by up to 1%. This estimate, while modest, is significant for large commercial vessels where propulsion power is substantial.

Beyond conventional rudder optimization, a number of energy-saving devices have been developed to improve hydrodynamic efficiency. For instance, the *Gate Rudder® System* introduces a novel twin-rudder configuration designed to generate additional thrust similar to an accelerating ducted propulsor [18]. While this device functions primarily as an

energy-saving appendage rather than an autopilot system, its field trials demonstrated approximately 14% fuel savings during sea trials of a 2,400 GT container ship in Japan, together with enhanced manoeuvrability and reduced vibration.

More closely related to automated steering, manufacturers of inland autopilot systems such as *Tresco Engineering* report substantial gains from "smart steering." They claim that their "Track Pilot" causes fuel reductions of up to 15% [19]. In a representative case published on their website, an Antwerp–Duisburg voyage of roughly 400 km achieved fuel savings of about 800 L,corresponding to roughly €1,000,simply by engaging the TrackPilot autopilot [19].

Complementary evidence comes from *Anschütz GmbH*, a manufacturer of advanced autopilot systems for seagoing and inland vessels. In field tests involving five twin-rudder tankers operating in North America, Anschütz reported fuel savings in every scenario, with reductions of up to 4.7% and an average just below 2% compared to manual steering [2]. The report, published on their website, attributes these gains to the adaptive *ECO and Course Control* modes of its NautoPilot series. They claim that the results were validated through computational fluid dynamics simulations and were consistent with extensive operational experience, reinforcing the causal link between smoother autopilot steering and lower fuel burn.

Despite the growing body of evidence from industry and pilot projects and a really clear causal link between unnecessary rudder utilization and fuel consumption, there remains a clear research gap: no prior study has systematically and causally quantified the effect of an **Autonomous Lane Assist (ALA)** system on fuel consumption using a data-driven, counterfactual framework. The present research directly addresses this gap by employing causal inference methods,specifically g-computation,to estimate the expected fuel consumption under alternative scenarios with and without ALA activation, providing the first statistically rigorous evaluation of its true efficiency impact.

6.4 Fuel Consumption Prediction in the Maritime context

Table 1 summarizes a selection of recent studies on **Fuel Consumption Prediction** in maritime contexts, with particular attention to studies relative to inland shipping. The table is structured to highlight four key dimensions:

- **Methods** – to identify prevailing modeling approaches.
- **Operating Domain**
- **Aim**
- **Prediction Timeframe** – as this determines practical applicability; for example, long-term predictions enable planning, while short-term estimates are often limited to post-estimation.

Entries highlighted in yellow represent elements that align most closely with the scope of this thesis.

Identified Research Gap. While previous work provides valuable insights, there are several key limitations with respect to the specific aims of this study:

- No study focuses on the Rhine river system.
- Existing models rely on time-series sensor data, while this implementation uses *per-edge aggregated traversal data*.
- No study explicitly learns *edge-specific patterns* along planned routes.
- Most works labeled as "prediction" actually operate as *post-estimation* tools. They use known future variables (e.g., wind, speed, rudder angle), which are not available at the time of decision-making. This limits their use for pre-voyage fuel optimization.

In contrast, the present study aims to develop a predictive model capable of estimating fuel consumption across entire voyages (often spanning dozens of hours or days), **based only on data available at the time of departure**.

Ref.	Methods	Operating Domain	Aim	Prediction Timeframe
[22]	RNNs,LSTM,RSSA	Inland – Yangtze River	Optimize engine speed for fuel consumption	Real-time(next minute)
[21]	BPNN	Inland – Yangtze River	Optimize engine speed for fuel consumption	Real-time(next minute)
[13]	LSTM, BP	Inland – Yangtze River	Operational Auxiliary role	Next half hour
[10]	ANN	Sea	Operational Auxiliary role	Real-time
[9]	MTL	Sea	Operational Auxiliary role	Real-time(next 5 minutes)
[4]	MLR, SVR, ANN	Port - port calls	better Emission Inventories	Entire port call
[6]	Parametric White Box,Black Box(LAR,RF,RLS) Grey Box	Sea	Trim Optimization	Real-time(next 15 seconds)
[12]	MLR, ANN	Sea	Operational Auxiliary role	Real-time(next minute)
[15]	MPR, ANNs, XGBoost	Sea - Oil tankers	Operational Auxiliary role	Entire Voyage(up to 24 days)
[3]	Optimization Under Uncertainty	Inland around a lock	Minimize fuel consumption	Entire crossing of water segment
[23]	LR,SVR,ANN	Sea	Reduce fuel consumption	Real-time(30-110s)
[11]	FNNs	Sea	data pre-processing effect	Real-time(5 minutes)

Table 1: Summary of literature on fuel-consumption prediction

7 Dataset Description

7.1 Original Dataset

As mentioned in the Introduction 1 , the data is obtained from Shipping Technology's core product, the ST-BRAIN, installed on over 250 vessels.
It collect high frequency nautical data of many types, using third party sensors installed on board to enable real-time monitoring, assisted navigation, and autonomous control.

7.1.1 Edge definition

Before introducing the individual variables that are present in this dataset and their meaning, it is important to clarify how the "edges"(directed river sections) introduced in Section 1 are defined.

As described in the Literature Review 6, a key characteristic of this study is the fact that the data is aggregated per directed river section(edge), this choice was made on the assumption that this would render implicit a few river characteristic, such as curvature, width, and even usual traffic that are not directly measured, but influence fuel consumption.
The employed ML models might infer these characteristics by weighing the feature "edge" (referring to a oriented river segment) differently and using it to influence fuel consumption directly.

An **edge** is defined as a **directed river segment**.
The process followed to divide the river into directed river segments is as follows:

- **Landmarks** are defined as fixed coordinates along the river, each identified by a unique ID (e.g., *Landmark 141*). They serve as reference points from which surrounding regions are defined. Shown in red in Figure 1.
- Around each landmark, a **Voronoi polygon** is constructed. This polygon represents the neighborhood region closest to that landmark compared to all others (e.g., *Polygon 141*). In this way, the river is divided into local areas centered on each landmark.Shown in green in Figure 1.
- A **Node** represents a directional crossing between two neighboring polygons. It indicates the transition from one Voronoi region to another (e.g., *Node (141,142)* represents a movement from Polygon 141 to Polygon 142).Shown in blue in Figure 1.
- An **Edge** is defined as a traversal through a polygon. It connects two consecutive nodes: one entering the polygon and one exiting it. For example, *Edge ((141,142), (142,143))* represents the traversal across Polygon 142, entering from Node (141,142) and exiting through Node (142,143)).Shown in orange in Figure 1.
- The collection of all such edges forms a directed graph representation of the river network, where:

– vertices correspond to nodes (polygon crossings),
– edges correspond to directed river segments,
– and each traversal through the river can be described as a sequence of directed edges.

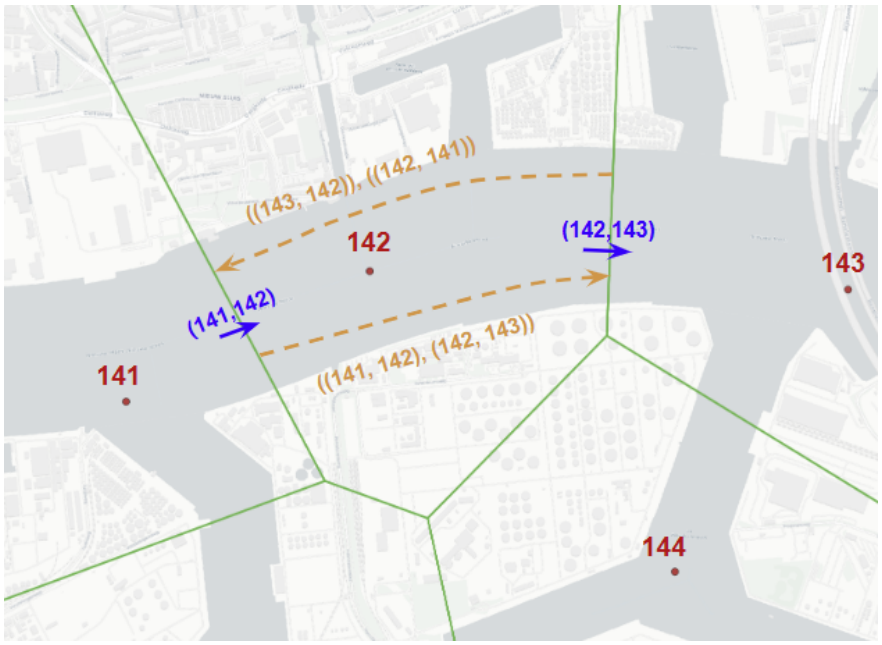


Figure 1: Division of the river into edges: directed river segments

7.1.2 Variables

The core dataset consists of edge-level traversals on the Rhine network, with one record per traversal. After cleaning, it contains 1.012.192 traversals from 132 distinct vessels between 2018 and June 2025.

Variables fall into the following groups:

- **Vessel identifiers and characteristics:** device ID (ship proxy), ship type (cargo, tanker, container, ...), build year, registered length and beam, number of engines.
- **Trajectory and environment:** directed river segment (`edge`), previous/next edge, distance, environment category (UPSTREAM, DOWNSTREAM, TIDAL), start/end timestamps, stop indicator, approximate course-over-ground.
- **Hydrostatic and loading state:** bow and stern draft, median draft and trim, depth under keel, median cargo weight, estimated total displaced mass.
- **Engine and operational variables:** average engine RPM, load, fuel rate, speed over ground, total traversal time.
- **Automation and steering metrics:** ALA activation time and sessions, binary ALA indicator (`ala_on`), heading-change metrics (`hc_mean_abs_deg`, `hc_std_abs_deg`, etc.).

Derived variables used as main outcomes or covariates (e.g. total fuel consumption, fuel consumption per kilometre and per tonne-kilometre, total mass, relative heading change) are described alongside the models that use them. A full variable dictionary, including aggregation rules, units, and data types for every column, is provided in Appendix A.

7.2 Dataset Cleaning

The uncleaned dataset contains more than 10 million rows, and all sorts of shipping operations are captured.

This Dataset Cleaning step has **two main objectives**.

Firstly to remove sensor errors and outliers, a necessary step for Dataset interpretation, analysis, and training ML models.

Secondly to select for a "comparable sailing condition", this is necessary to ensure "positivity" within the dataset, a requirement for the causal inference methods chosen, as described in Section 8.1.2.
One objective of this study is to estimate the effect of the Autonomous Lane Assist on fuel consumption under.
The ALA can only be used (at the moment) in "normal" sailing operations, and it only makes sense to compare its results to other situations in which, at least in principle, it could be used.

To put it more formally, for all relevant variable both ala_on = 1 and ala_on = 0 should occur with non zero probability, meaning there should be appearances of both in the dataset. This characteristic is also known as "Positivity", and is a requirement for G-computation to work as intended.

Without defining a "comparable" sailing condition and limiting the data cleaning to removing outliers and sensor errors, many traversals with the impossibility of ALA activation would remain.
The counterfactual analysis is predicated on the fact of toggling the ala_on variable on or off for all traversals, so all traversals need to be such that ALA could in principle be activated, under those conditions.

For these reasons, a "comparable" traversal is defined as a traversal that aims at moving over the specific edge smoothly, and without sudden acceleration or deceleration.
In practice this means:

- removing traversals that are U-turns
- removing traversals that self intersect
- removing traversals that contain stop
- removing traversals that just came out of a stop,as they may be heavily accelerating
- removing traversals that are going to stop soon, since they might have already started breaking, especially while going downstream.
- restricting the geographical area to one where there are any examples of traversal that employ and do not employ the ALA.

Therefore, for the reasons above, the following cleaning steps were performed:

- **Ship-level consistency:** removed ships that report an inconsistent number of engines (`n_engines` varying across traversals).
- **GPS and geometric errors:** GPS errors, especially near bridges, caused linestrings that show an extreme heading change between every section of the linestring, so traversals with a `max_turn_angle_deg` greater than 20° were removed, which typically indicates corrupted or noisy GPS data. Traversals that self-intersect were also removed using the `.is_simple` method from the `shapely.wkt` module, which flags geometries crossing themselves.
- **Treatment consistency:** kept only traversals where `ala_active_percentage` equals 0 or 100. This ensures that ALA is either fully active or fully inactive during a traversal, removing time-dependent treatment changes and allowing a static G-computation setup.

- **Draft definition consistency:** removed ships whose `med_draft` values consistently ranged between 0.15 and 0.6 m. These ships appear to have sensors functioning but with an offset definition of draft, making their measurements non-comparable.
- **Non-negative filtering:** removed rows with non-positive values in any of the following columns: `computed_speed`, `avg_speed`, `fuel_consumption`, `fuel_consumption_lh`, `avg_engine_load`, `total_time`, `distance`, `med_weight`, `med_draft`. water levels were excluded from this filter since their scales are arbitrary.
- **Missing values:** dropped rows containing NaNs in critical variables such as `fuel_consumption`, `fuel_consumption_lh`, `total_time`, `distance`, `avg_speed`, `computed_speed`, `avg_engine_rpm`, `avg_engine_load`, `ala_active_percentage`, `environment`, `prev_trav_id`, `next_trav_id`, and `med_weight`.
- **Operational condition filtering:**
  - Removed U-turns, defined as edges of the form `((a, b), (b, a))`.
  - Removed traversals that contain a stop.
  - Removed traversals immediately after a stop (likely accelerating).
  - Removed traversals immediately before a stop (likely decelerating).
- **Exclude barge pushers:** removed ships longer than 135 m, since these often push or carry barges for which mass and hydrodynamic effects are unknown.
- **Geographical restriction:** only considered edges between Dordrecht and Koblenz. North of Dordrecht, tidal influence and port operations introduce inconsistent sailing conditions. South of Koblenz, numerous locks cause abrupt and irregular sailing behavior difficult to model uniformly. The selected region still covers the vast majority of all traversals.
- **Outlier filtering:** removed outliers based on physically plausible ranges (either within $\pm 3\sigma$ or validated by experts):
  - `computed_speed`, `avg_speed`: 0–15 knots
  - `total_time`: 125–1000 seconds
  - `avg_engine_rpm`: 300–1920 rpm
  - `fuel_consumption_lh`: 10–348 liters/hour
  - `med_draft`: 0.6–4.4 meters
  - `med_trim`: –1.15–2 meters
  - `med_weight`: 0–4.800 tons
  - `total_mass`: 0–6.000 tons
  - `med_depth`: 0–8 m
  - `waterlevel_koblenz`: 0–700 (in arbitrary units

- **Edge frequency:** kept only edges appearing at least 50 times to ensure sufficient data for per-edge model training. This condition was never broken after selecting for the Dordrecht - Koblenz region.
- **Ship IDs:** Kept only ships that have the ST-ALA system installed and have used it at least once.
- **Treatment presence:** kept only edges where both ALA-on and ALA-off traversals occur, with at least 15 instances of ALA-on (100% activation). This guarantees that both conditions appear with non-zero probability ("positivity"), allowing valid counterfactual estimation. This condition was never broken after selecting for the Dordrecht - Koblenz region.

total time A major concern regarded the variable `total_time`.
As mentioned above, it is defined as the difference between `end_time` and `start_time` for each traversal.

The problem lays with the fact that they are defined, respectively, as the last and first point of the trajectory linestring that is part of that edge(meaning physically inside of it), and the linestring is obtained by sampling the GPS position of the vessel every 15 seconds.

This means that there is a portion of linestring at the beginning and end of each traversal that is not considered in the calculation of `total_time`, and since a traversal can last even as little as 200 seconds, an error of 15 seconds on `total_time` could mean an error of 7.5%, and since `total_time` appears in the calculation of `fuel_consumption_km`, which is the outcome for the causal inference methods and the target for the prediction ones, the error could propagate into the model.

Luckily from the definition of `total_time` it follows that the error should be unbiased by ala status, and have a constant expected error of -15 seconds. Meaning that `total_time` is on average 15 seconds shorter than the real time that the traversal took.

To check this, an alternative time was computed, using `distance` and `avg_speed`.

$$\text{computed\_time} = \frac{\text{distance}}{\text{avg\_speed}}$$

Then the difference between `computed_time` and `total_time` was obtained and its distribution was plotted, the result is in Figure 2

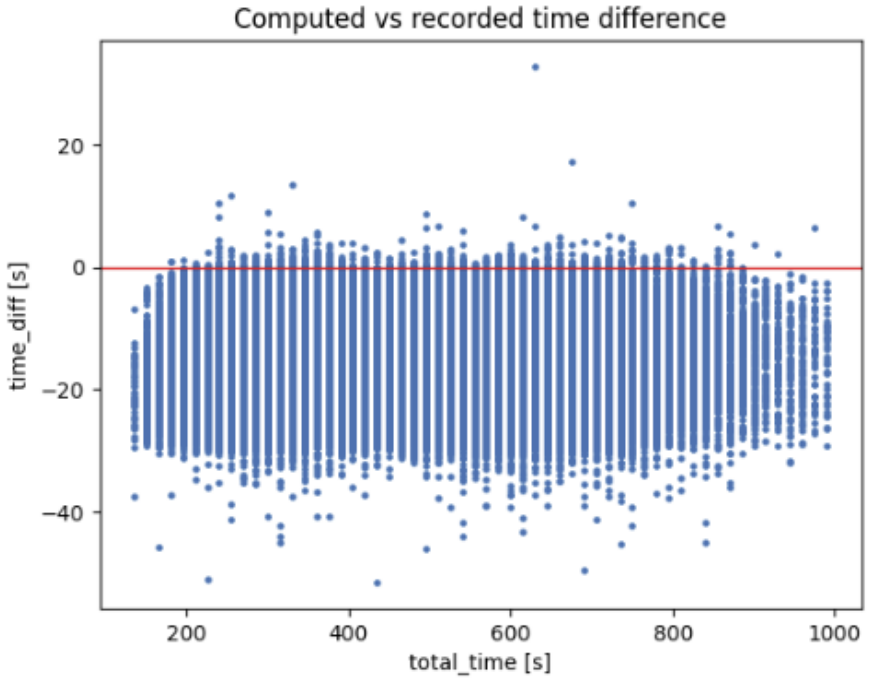


Figure 2: Distribution of the estimated error on `total_time`

The plot confirms that the error on `total_time` is independent of its magnitude, and -15s on average. `computed_time` was therefore used going forward instead of the original value.

The dataset as obtained above is the dataset used in Section 8.2 for Prediction of fuel consumption per river section.
The causal inference methods, on the other hand, need further data processing, as described in Section 8.1.2.

8 Methodology

8.1 Historical Evaluation Modeling

As described in Section 6.1, **G-Computation** is the method chosen to answer the main research question.
How much fuel is saved when an inland vessel sails using the Autonomous Lane Assist system compared to conventional steering methods?

A complete description of the Methodological Framework is provided in Section 8.1.5.

8.1.1 Modeling Goals and Assumptions

The purpose of the modeling framework developed in this study is to estimate the causal effect of the `Autonomous Lane Assist (ALA)` system on fuel consumption using the g-computation approach introduced in Section 8.1.5. The models are designed not merely for prediction but for *counterfactual estimation*, where the objective is to infer what the total fuel consumption of a traversal would have been under an alternative treatment condition (ALA on or off).

The key modeling goal is to approximate the conditional expectation functions $E[Y \mid A, L]$ and $E[M \mid A, L]$ accurately enough to enable valid computation of counterfactual outcomes through the g-formula and its iterated variant. To achieve this, flexible, non-parametric models based on gradient-boosted decision trees were adopted, allowing nonlinear and interaction effects to be captured without restrictive assumptions on functional form.

A critical assumption underlying the validity of the g-computation in this setting is that the **data aggregation methods** described in Section 7 retain sufficient temporal and spatial resolution for counterfactual inference.

The traversal-level dataset aggregates all measured variables over river segments and computes summary statistics such as mean speed, engine load, and total fuel consumption. This aggregation assumes that the essential causal structure linking ALA activation to fuel consumption,through rudder activity, hydrodynamic resistance, and traversal duration,is preserved.
Specifically, some variables are sampled at a frequency of one point every 15 seconds along each `linestring`, with `start_time` and `end_time` recorded for each traversal. This sampling introduces a known approximation error in `total_time`, which is a key mediator in the proposed causal pathway from ALA usage to total fuel consumption.

This is one of the main methodological concerns, since bias in the measurement of total traversal time could directly affect the estimated indirect effects through speed and duration.

Since `total_time` ranges between 200s and 800s, an error of 30s(2 x sampling time) on

this variable, could cause an error between 15% and 3,75%.
This alone has the same order of magnitude as the effect on fuel consumption as referenced in Section 6.3, and would possibly render this study inconclusive. However, this issue was addressed in the dataset cleaning step (Section 7.2): an alternative traversal time was computed from recorded distance and average speed, and verified to be unbiased with respect to ALA status. This `computed_time` variable is used as the outcome throughout the analysis in place of the raw `total_time`.

Beyond data fidelity, standard causal assumptions apply, as outlined in [8, 5]. These include:

- **Consistency:** Each traversal's recorded fuel consumption corresponds to the potential outcome under its observed ALA activation state.
- **Positivity:** For all relevant combinations of confounders L, both ALA on and off states occur with non-zero probability.
- **Conditional Exchangeability:** Given the observed covariates L, ALA activation is independent of unmeasured factors affecting fuel consumption.

Finally, the modeling assumes that residual bias from unmeasured confounding,such as local wind, current fluctuations, or crew behavior,is small relative to the variance explained by the observed variables, and that the aggregation level preserves the essential variability of ship-environment interactions.

Under these assumptions, the model estimates obtained through the g-computation and its iterated form can be interpreted as consistent estimators of the causal effect of ALA activation on fuel consumption across river traversals.

8.1.2 Requirements

For g-computation and IPW to yield valid causal estimates, three key assumptions must hold: **consistency**, **positivity**, and **exchangeability**. These conditions ensure that the estimated contrasts between potential outcomes correspond to the true causal effects, rather than being distorted by measurement inconsistencies, data imbalance, or hidden confounding. Their formal definitions and implications are discussed in [8] (Chapters 3 and 19) and [5]. Below, each is presented in the context of this study and checked against the characteristics of the dataset.

Consistency. Consistency requires that the observed outcome for each unit equals the potential outcome corresponding to the treatment actually received:

$$Y = Y^A$$

This means that the variable labeled `fuel_consumption_km`(the outcome value) must truly represent the realized outcome under the treatment state experienced by each traversal. In practice, this assumption is satisfied when:

1. The treatment variable (`ALA`) is well-defined and unambiguous for each traversal.
2. The measurement quality and method of the outcome does not depend on the treatment assignment.
3. ALA status is unique for each traversal

In the dataset, this is ensured by including only traversals with `ala_active_percentage` equal to either 0% or 100%, so that the Autonomous Lane Assist system is never switched on or off mid-traversal. The treatment is therefore binary and stable across the unit of analysis.

The outcome is defined as:

$$\text{fuel\_consumption\_km} = \text{avg\_engine\_fuel\_consumption} \times \text{n\_engines} \times \frac{\text{total\_time}}{3600}$$

The same physical definition and measurement units apply for both ALA and non-ALA traversals, satisfying the uniformity condition.
However, one potential concern arises: since ALA can affect `total_time`, this variable indirectly enters the definition of fuel consumption. While this does not violate consistency, it implies that total time acts as a mediator in the causal chain, and care must be taken to account for it in the modeling process rather than adjusting for it as a confounder. Overall, given the data structure, the consistency assumption is considered satisfied.

Positivity Positivity requires that, for every relevant combination of confounders L, both treatment levels occur with non-zero probability [8, 5]:

$$0 < P(A = a \mid L = l) < 1.$$

In this setting, this means that for every "comparable sailing condition",defined by vessel characteristics, loading condition, water level, and operational environment,the dataset must contain both ALA-activated and non-activated traversals. Without such overlap, the g-computation procedure would be forced to extrapolate beyond the support of the observed data, resulting in unstable and unreliable counterfactual estimates.

To enforce this requirement, a dataset was first constructed where, for each edge, only traversals that the ST-ALA would realistically attempt were retained (see Section 7.2), using distance-based filtering.
A propensity score was then estimated using an XGBoost decision-tree model, with input features `ship_type`, `total_mass`, `avg_engine_rpm`, `avg_engine_load`, `environment`, `med_depth`, `waterlevel_koblenz`, `edge`, and `device_id`, and treatment variable `ala_on`.
The propensity score is defined as the conditional probability of receiving treatment given covariates:

$$e(L) = P(A = 1 \mid L).$$

The SHAP analysis (Fig. 3) shows that `device_id` is the dominant driver of the propensity model, indicating that the specific vessel is the main determinant of whether ST-ALA was used, some ships(captains) simply use the autonomous lane assist system more than others.
Beyond this, low water levels, high RPM, and high engine load increase the probability of ALA activation, whereas the individual river segment (`edge`) has only a minor influence.

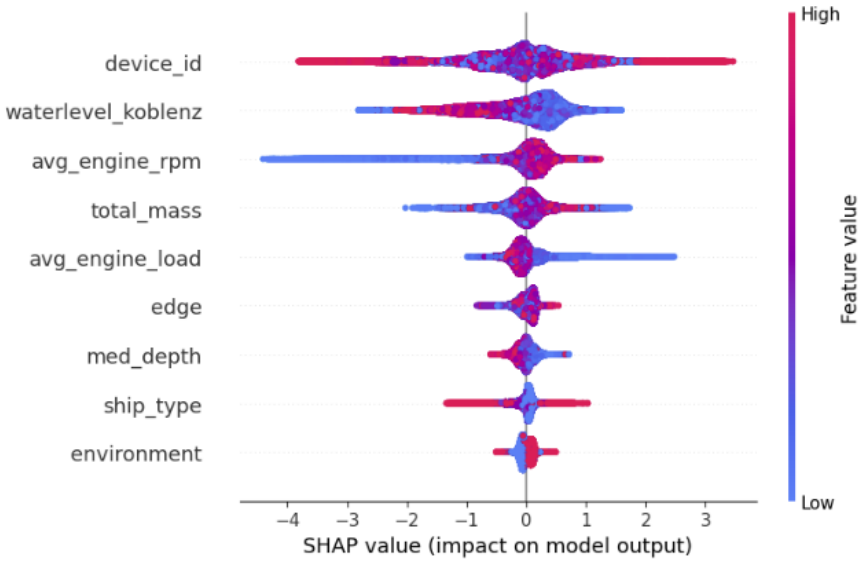


Figure 3: SHAP Analysis of the Propensity Score Model, `device_id` has the biggest effect on ST-ALA usage

Across vessels, the minimum average propensity score was 0.43, implying that all ships operate in conditions where ST-ALA could in principle be used, even though empirical usage rates vary (as low as 12% for some devices).
At the edge level, the minimum observed ALA usage was 23%, and the lowest predicted propensity was 0.27, again indicating that no edge falls into a never-treated or impossible-treatment region.

The propensity score distributions by ALA status (Fig. 4) nevertheless reveal areas of no overlap, meaning in the dataset there are conditions under which ST-ALA would not be used, as shown in Figure 4

To ensure strict practical positivity and avoid extrapolation, the dataset was trimmed to the central region of the propensity-score distribution (0.1–0.9), as suggested in the literature.
A more conservative trimming (0.2–0.8) yielded nearly identical results, the improved overlap reduced sample size without influencing the results. The final trimmed dataset therefore satisfies the positivity requirement sufficiently to support credible and stable counterfactual estimation.

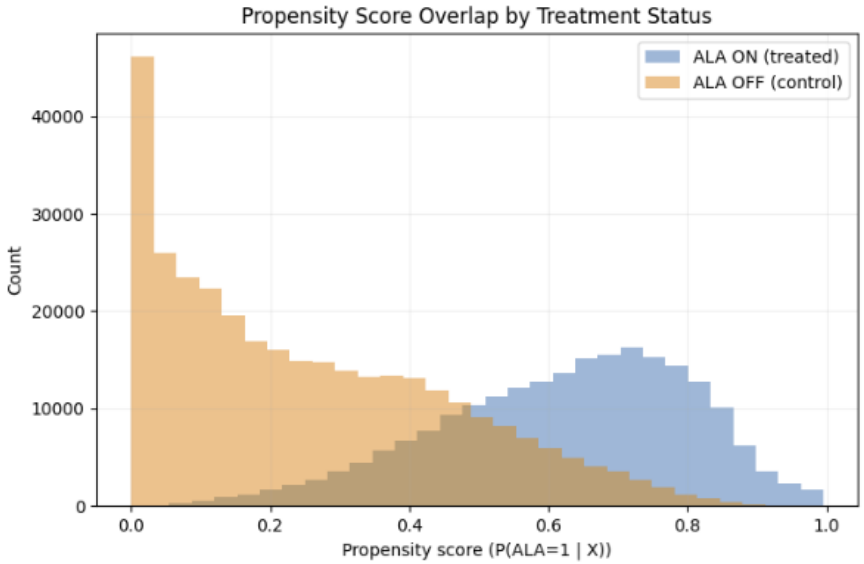


Figure 4: Propensity Score distribution by ST-ALA status(on or off), trimming is necessary as there are areas of non-overlap at the extremes

Exchangeability. Exchangeability (also referred to as conditional independence) requires that, after conditioning on the observed confounders L, treatment assignment is independent of the potential outcomes:

$$Y^a \perp A \mid L$$

This implies that, within levels of L, any remaining differences in fuel consumption between ALA and non-ALA traversals are attributable to the treatment itself rather than to unobserved causes [8, 5].

In this dataset, complete exchangeability cannot be guaranteed,unmeasured variables such as local weather (wind, rain, traffic) may influence both ALA activation and fuel efficiency. However, most of these effects are expected to be weak or indirectly captured through the available environment variables (e.g., `environment`, `edge`, `waterlevel`, and `tidal` conditions). Therefore, the remaining unobserved confounding is assumed to be small relative to the main effects. This is a **subject-matter assumption**, justified mechanistically rather than statistically.

Formally, the following is assumed:

$$E[Y^a \mid A = a, L] = E[Y \mid A = a, L]$$

and that residual bias from omitted variables does not significantly alter the sign or magnitude of the estimated ALA effect. Supporting this claim, the high predictive accuracy of the model trained on all observed variables (CV $R^2 = 0.996 \pm 0.016$, CV RMSE $= 0.486$) indicates that most of the variation in fuel consumption is already explained by the available confounders. This suggests that any remaining confounders not included in L have limited influence on the outcome.

Residual confounding may persist, but given the breadth of included operational and environmental variables, and the high explanatory power of the model, the exchangeability assumption is treated as approximately satisfied for the purposes of causal estimation.

8.1.3 Means Analysis

A means analysis was used to benchmark the counterfactual prediction of the g-computation.

A means analysis consists of:

- selecting a subset of the Dataset
- subdividing the observation based on the treatment (ala_on = True vs ala_on = False)
- averaging the value of the fuel consumption over treatment consistent subsets
- comparing the results

The exact variable chosen as the treatment is `fuel_consumption_km` defined as

$$\text{fuel\_consumption\_km} = \left(\frac{\text{fuel\_consumption}}{\text{distance}}\right) \times 1000$$

This is the fuel consumption normalized by distance.. This normalization allows for comparability of fuel consumption between edges that have different intrinsic lengths.

A problem with the use of this choice is the fact that the variable `distance`, which is the length of the linestring per traversal, is a possible mediator between ALA activation and fuel consumption (in absolute terms[Liters]), since more unnecessary rudder movement might mean more deviation from the desired path and a longer traversal.
This issue is explored further in subsequent sections.

In the absence of a confounder robust counterfactual estimation method, normalization by distance seems the most straightforward and sensible way of rendering the results comparable between edges that have different lengths.

The results of the mean analysis on all data is as summarized in Table 2.

Table 2: Overall mean fuel consumption and savings percentage

Key	**Mean (ALA on)**	**Mean (ALA off)**	**n obs (ALA on)**	**n obs (ALA off)**	**Savings %**
overall	9.977145	9.348046	49975	152612	-6.72974

The results as on the data divided per environment are as summarized in Table 3.
The results as on the data divided per ship type are as summarized in Table 4.

Table 3: Mean fuel consumption and savings per environment

Environment	**Mean (ALA on)**	**Mean (ALA off)**	**n obs (ALA on)**	**n obs (ALA off)**	**Savings %**
DOWNSTREAM	4.20209	3.99414	23178	78969	-5.206388
TIDAL	8.018028	7.871808	2317	5873	-1.868948
UPSTREAM	15.630389	15.714618	24480	67770	0.535988

Table 4: Mean fuel consumption and savings per ship type

Ship type	**Mean (ALA on)**	**Mean (ALA off)**	**n obs (ALA on)**	**n obs (ALA off)**	**Savings %**
cargo	9.588716	11.41825	4950	9102	16.022889
container	10.722474	10.70879	12873	23580	-0.136298
tanker	9.738532	8.923566	32152	119930	-9.132736

Finally Table 5 subdivides the data both by environment and ship type.

Table 5: Mean fuel consumption and savings per environment and ship type

Environment	**Ship type**	**Mean (ALA off)**	**Mean(ALA on)**	**n obs (ALA off)**	**n obs (ALA on)**	**Savings %**
DOWNSTREAM	cargo	4.577325	4.144178	5072	2546	9.462876
DOWNSTREAM	container	5.45925	5.47932	12817	6177	-0.36764
DOWNSTREAM	tanker	3.638275	3.666497	61080	14455	-0.77568
TIDAL	cargo	7.660511	6.614548	204	184	13.65395
TIDAL	container	9.20528	8.670743	1068	1064	5.80683
TIDAL	tanker	7.571646	7.611888	4681	1069	-0.53147
UPSTREAM	cargo	20.687395	16.079275	3826	2220	22.25703
UPSTREAM	container	17.812207	16.860613	9695	5632	5.34237
UPSTREAM	tanker	14.989038	15.153775	54249	16628	-1.09904

This approach, blindly averaging over traversals, ignored all possible confounders, defined as all variables that affect both the likelihood of the ALA system being active and fuel consumption and is therefore unfit to answer the research question/

Nevertheless this approach is a good benchmark since it mimics what may be a naive intuition about the nature of the savings that is caused by drawing conclusions from anecdotes.
Someone, or a client, might take various observations about fuel consumption from various voyages and blindly compare the fuel consumption averaged while using or not using the autonomous lane assist.

This means analysis showed that using this approach, randomness can really dominate, and any result or its opposite are possible, therefore any type of claim can be made by simply cherry picking a subset of the data.

In fact it is striking how big of a variability there is between rows of Table 5, cargo ships going upstream for example save 22% of fuel while sailing with the Autonomous Lane Assist active, while tanker ships going upstream loose 1%, and all while tanker ships in general suffer a 9% loss in fuel efficiency while sailing with ALA active.

There are many trend reversals based on category subdivisions, and even within categories, there can be a trend reversal after further subdividing.
For example looking at the downstream environment in Table 4, a saving of -5% is shown, but when subdividing the downstream environment by ship type savings of 9% -0.3% and -0.7% are displayed.

Or the ship category "tanker" displays a -9% saving overall but -0.7%, -0,5% and -1% when further subdivided by environment.

The root cause of this can be identified in Sympson's Paradox:
Simpson's paradox is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined. This result is often encountered in social-science and medical-science statistics, and is particularly problematic when frequency data are unduly given causal interpretations. The paradox can be resolved when confounding variables and causal relations are appropriately addressed in the statistical modeling.[20]

In fact a simple, first level investigation into the possible causes of such a big saving for cargo ships going upstream, shows that when the ALA is active, these ships appear to be sailing at lower rpms, which is one of the main drivers of fuel consumption, as shown in Figure 2 below.

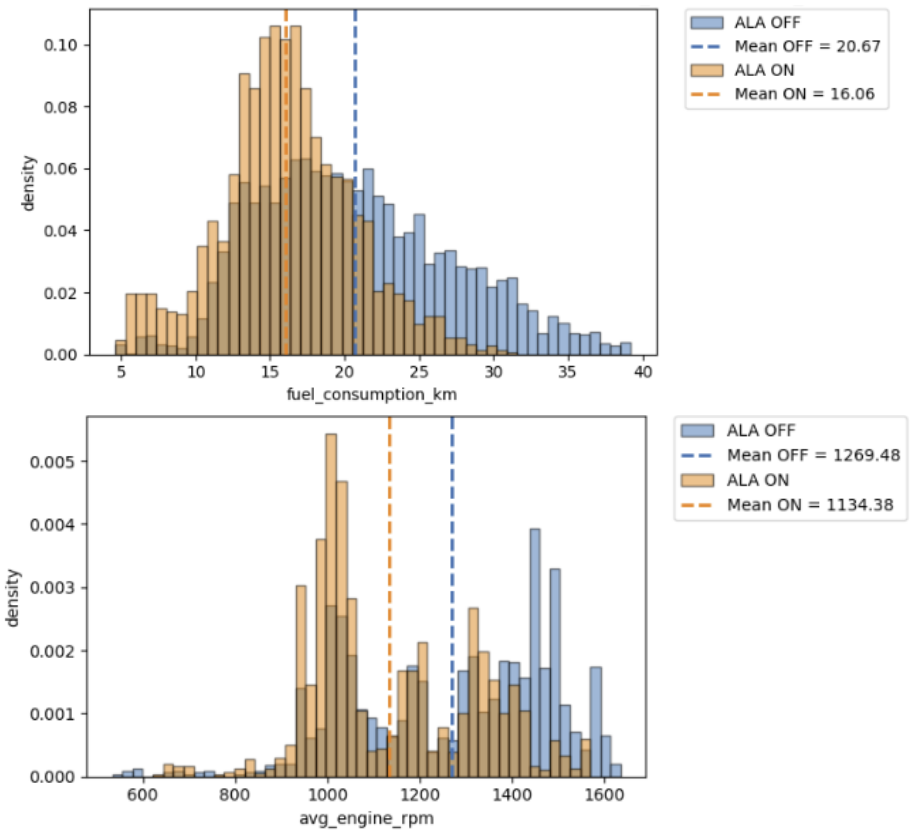


Figure 5: Distribution and mean of fuel consumption per km and engine rpm for cargo ships going upstream

8.1.4 Inverse Probability Weighting (IPW)

Inverse Probability Weighting (IPW) was implemented as a complementary causal estimation strategy alongside the primary g-computation approach.
Both methods were applied to the same cleaned dataset, after removing erroneous traversals, filtering outliers, and enforcing overlap through propensity-score trimming.
The goal of IPW in this context is not to provide the main estimate of the ALA effect, but to offer an orthogonal benchmark: if the estimated effect has the correct order of magnitude under two very different identification strategies, confidence in the robustness of the inferred effect increases.

Principle of IPW. IPW reweights the observed data such that the distribution of confounders becomes independent of treatment assignment. Given the propensity score

$$e(L) = P(A = 1 \mid L),$$

each observation is assigned a weight equal to the inverse probability of having received its observed treatment:

$$w_i = \begin{cases} \dfrac{1}{e(L_i)} & \text{if } A_i = 1, \\ \dfrac{1}{1 - e(L_i)} & \text{if } A_i = 0. \end{cases}$$

To limit the influence of extreme propensities, stabilized weights were used:

$$w_i^{\text{stab}} = \begin{cases} \dfrac{P(A=1)}{e(L_i)} & \text{if } A_i = 1, \\ \dfrac{P(A=0)}{1 - e(L_i)} & \text{if } A_i = 0. \end{cases}$$

After computing the stabilized weights, their empirical distribution was inspected (minimum, median, 90th and 95th percentiles, maximum) to verify that no extreme weights dominated the estimation.
This step confirms the practical positivity induced by the trimming applied earlier in the pipeline.

IPW estimation of the ATE. Under IPW, the mean potential outcomes are estimated as weighted averages of the observed outcomes. In plain terms, the procedure is:

1. Compute the weighted mean fuel consumption for traversals with ALA activated, using the stabilized IPW weights.
2. Compute the analogous weighted mean for traversals with ALA off.
3. Take the relative difference between these two weighted means to obtain the ATE in percentage terms.

Formally, letting Y denote fuel consumption per km and w_i the stabilized weight:

$$\hat{\mu}_1 = \frac{\sum_{i:A_i=1} w_i Y_i}{\sum_{i:A_i=1} w_i}, \qquad \hat{\mu}_0 = \frac{\sum_{i:A_i=0} w_i Y_i}{\sum_{i:A_i=0} w_i},$$

and the relative ATE is computed as

$$\widehat{\text{ATE}}_{\text{IPW}} = \frac{\hat{\mu}_0 - \hat{\mu}_1}{\hat{\mu}_0} \times 100.$$

Applying this procedure to the cleaned dataset produced:

$$\widehat{\text{ATE}}_{\text{IPW}} = -0.35\%,$$

indicating slightly higher fuel consumption when ALA is active. Although the sign is flipped compared to the g-computation results the order of magnitude is consistent with it, a near zero effect is observed.

Role of IPW relative to g-computation. In this study, g-computation is the primary estimator because it leverages a rich, flexible outcome model capable of capturing complex nonlinear relationships between covariates and fuel consumption. IPW, by contrast, relies entirely on the specification of the treatment model and does not model the outcome.

In other words, IPW is not expected to outperform g-computation here. Its function is diagnostic: if both approaches, despite their methodological differences, point toward comparable effects, it strengthens the credibility of the causal conclusions. That is precisely what is observed in this analysis.

8.1.5 G-Computation

The **G-computation formula**, first introduced by Robins and later formalized in the epidemiological and statistical literature [8, 5], is a fundamental method for estimating causal effects from observational data. It belongs to the family of *g-methods*. These methods provide consistent estimates of contrasts between potential outcomes (e.g., differences or ratios) under a less restrictive set of assumptions than standard regression techniques, particularly when confounding and feedback between treatment and covariates are present [5].

In this study, the goal of the g-computation is to estimate the causal effect of the `ST-ALA` system (Autonomous Lane Assist) on `fuel consumption per traversal`.
Since each data point corresponds to a single aggregated river traversal, the treatment (`ALA on/off`) and outcome (`fuel consumption`) are defined at the same temporal level. Furthermore while performing cleaning of the original Dataset, traversal where ALA status was changed during the traversal were eliminated, under this condition the treatment `ALA on/off` is not time depended within a traversal. Therefore, the static (non–time-varying) version of the g-formula can be directly applied.

The static g-computation proceeds through the following conceptual steps [8, 5]:

1. Define a clear **causal model**, typically through a Directed Acyclic Graph (DAG).
2. Identify the **treatment** variable A, the **outcome** variable Y, and the relevant **confounders** L.
3. Verify the three core **identification assumptions**: consistency, positivity, and exchangeability.
4. Estimate the conditional expectation $\widehat{E}[Y \mid A, L]$ using an appropriate model.
5. Predict potential outcomes under each treatment level by setting $A = a$ for all observations.
6. Average over the confounder distribution to obtain the marginal expected outcomes $E[Y^a]$ and their contrasts.

ATE aggregation strategies. Because traversals are not independent and ships navigate heterogeneous river sections, three aggregation methods were used to summarize the unit-level savings into interpretable Average Treatment Effects (ATEs):

1. **Global ATE (row-weighted).** The simplest approach averages the per-row relative savings across all traversals:

$$\widehat{\mathrm{ATE}}_{\mathrm{global}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{savings}_i.$$

   This treats each traversal equally, regardless of edge characteristics or ship identity. It provides a high-level summary but may overweight vessels or edges that appear frequently in the dataset.

2. **Edge-weighted ATE.** To reduce overrepresentation of heavily-traversed edges, savings are first averaged within each unique river section:

$$\bar{s}_e = \frac{1}{n_e} \sum_{i \in e} \text{savings}_i,$$

followed by averaging across all edges:

$$\widehat{\text{ATE}}_{\text{edge}} = \frac{1}{E} \sum_{e=1}^{E} \bar{s}_e.$$

This treats each edge equally and better reflects the effect of ALA across the geographical network rather than traffic frequency.

3. **Ship-weighted ATE.** Analogously, to prevent ships with many recorded traversals from dominating the estimate, savings are first averaged within each vessel:

$$\bar{s}_d = \frac{1}{n_d} \sum_{i \in d} \text{savings}_i,$$

and then averaged across ships:

$$\widehat{\text{ATE}}_{\text{ship}} = \frac{1}{D} \sum_{d=1}^{D} \bar{s}_d.$$

This yields an estimate of the "typical vessel-level effect," preventing bias from a few high-frequency or atypical ships.

Together, these three aggregation perspectives provide a more complete understanding of the ALA effect. The global ATE reflects the empirical distribution of traversals, the edge-weighted ATE reflects spatial balance across the river network, and the ship-weighted ATE reflects balance across vessels with different operational profiles.

Directed Acyclic Graph (DAG). A **DAG** represents the assumed causal structure between treatment, outcome, and covariates. It encodes directional dependencies through arrows and explicitly rules out feedback cycles. Proper DAG specification is crucial because it determines which variables are true confounders (common causes of A and Y) and which should be excluded from adjustment to avoid bias [8]. The DAG guiding this work, described in Section 8.1.6, illustrates the causal pathway through which the ST-ALA system affects fuel consumption, mediated by factors such as rudder activity, trajectory curvature, and traversal duration.

Treatment and Confounders. In this framework, the **treatment** A is the binary activation state of the Autonomous Lane Assist system ($A = 1$ for active, $A = 0$ for inactive), and the **outcome** Y is the measured fuel consumption per traversal. A **confounder** L is a variable that influences both the treatment assignment and the outcome, thereby inducing a spurious association if unadjusted. Formally, a variable L is a confounder if it is a common cause of both A and Y and is not on the causal pathway from A to Y [8]. Properly identifying L is one of the most important steps in causal inference, since model specification,and thus the validity of the estimated effect,depends entirely on whether the causal structure is correctly defined.

Identification Requirements. To identify the causal effect using observational data, three conditions must hold:

- **Consistency:** the observed outcome for an individual equals the potential outcome corresponding to the treatment actually received, $Y = Y^A$.
- **Positivity:** for all levels of the confounders L, both treatment states must occur with non-zero probability, $0 < P(A = a \mid L) < 1$.
- **Exchangeability (No unmeasured confounding):** conditional on L, treatment assignment is independent of potential outcomes, i.e. $Y^a \perp A \mid L$.

These assumptions were discussed in more detail in Section 8.1.2.

The G-Formula. Under these assumptions, the average potential outcome under intervention $A = a$ can be expressed as:

$$E[Y^a] = \int E[Y \mid A = a, L = l]\, f(L)\, dl \tag{1}$$

In practice, this means fitting a model to estimate $E[Y \mid A, L]$, predicting outcomes under each treatment level a, and then averaging across the observed distribution of L. In this study, $E[Y \mid A, L]$ is estimated using flexible machine learning models trained on the historical dataset, enabling accurate approximation of the conditional expectation.

Model Specification. In the causal inference context, a model is said to be **correctly specified** if its conditional expectations are unbiased estimators of the true expectations of the outcome given the covariates. Formally, for any model $E[Y \mid A, L]$:

$$E[Y - \widehat{E}[Y \mid A, L] \mid A, L] = 0$$

This condition defines correct specification as a causal estimand, not merely as a predictive model. A model may be misspecified for prediction but still yield valid causal estimates if the conditional mean is unbiased with respect to the true data-generating process. In this project, machine learning models (e.g., CatBoost) are used as non-parametric estimators of $E[Y \mid A, L]$, following the principle that flexible models can approximate the true functional form without restrictive parametric assumptions [8, 5].

In summary, g-computation in this thesis provides a systematic framework to estimate the counterfactual fuel consumption that would have been observed had the ST-ALA system been active (or inactive) across all traversals. Its validity depends on:

- the correctness of the causal structure
- satisfaction of identification assumptions
- adequate model specification

8.1.6 DAG and Mediator Pathway

A **Directed Acyclic Graph (DAG)** is a graphical representation of the assumed causal relationships between variables in a system [8]. Each node in the graph represents a variable, and each directed edge (arrow) represents a causal influence from one variable to another. The DAG encodes the causal assumptions that underpin the entire g-computation procedure, as it determines which variables are confounders to be adjusted for, which are mediators to be modeled, and which should be excluded. Constructing it correctly is therefore critical, since any misspecification of causal relationships can invalidate the estimation of the causal effect.

In this study, the DAG is constructed to represent the causal mechanism linking the activation of the Autonomous Lane Assist (`ALA`) to `fuel consumption per traversal`. The graph includes both **original** and **computed** variables, as defined in Section 7.

The term **acyclic** indicates that the graph cannot contain feedback loops,no variable can influence itself directly or indirectly through a cycle. This is a fundamental requirement of causal modeling: the relationships are unidirectional and represent causal ordering in time or logical dependence. Cycles would imply mutual causation, which cannot be handled within the structural causal model framework and would make identification of causal effects impossible.

The proposed DAG is shown in Figure 9. Each arrow represents a hypothesized direct causal influence between variables, as derived from both physical reasoning and expert domain knowledge of ship operations.

The **treatment** is `ALA_on`

The **outcome** is `fuel_consumption_km` [L/km].

The chosen outcome variable for the g-computation is defined as:

$$\text{fuel\_consumption\_km} = \left(\frac{\text{fuel\_consumption}}{\text{distance}} \right) \times 1000,$$

which normalizes fuel consumption to a per-kilometre basis. Since river sections (*edges*) differ in length, this transformation ensures comparability across traversals and maintains

consistency with the outcome used in the means analysis and IPW benchmarks.

In the counterfactual analysis, the model predicts total fuel consumption per unit distance for every traversal under the two treatment assignments (A=0 and A=1). Relative savings are then computed at the traversal level.

The **confounders** are:

- **Ship Characteristics** including `device_id`, `ship_type`, `build_year`, `n_engines`, `dimensions_lengt` and`dimensions_width`
- **Constant Edge Characteristics** including `edge` and `environment`
- **Time Dependent Edge Characteristics** including `med_depth` and `waterlevel_koblenz`
- **Operational Sailing Parameters** including `med_weight`, `total_mass`, `med_draft` and `med_trim`
- **Captain's choice** including `avg_engine_load`, `avg_engine_rpm` and `distance`

All of these influence fuel consumption by defining the Hydrodynamic Resistance and parameters defining the engine thrust, but also the Captain's choice, having an effect on whether ALA is activated or not.

Particular attention is necessary towards the choice of using `distance` as a **proxy for trajectory choice** and it's role in the DAG.
One might argue that the Autonomous Lane Assist controls the trajectory of the ship, by definition, and not the captain. An important observation is that the ALA proposes a possible trajectory, using a data driven approach, but the captain can shift it left or right in set increments.

At the moment, ALA proposes a trajectory based on how captains historically sailed in that river section, but it does not try to optimize it for fuel consumption.
The savings come from a better adherence to the proposed trajectory, and not from the choice of a better one.

Regarding the choice of using `distance` as a **proxy for trajectory choice**, it is important to note a couple of facts about inland sailing:

- the speed of the water is higher in the middle of the river, due to boundary conditions slowing it down near the riverbank, a speed gradient is present in between.
- It is common practice (and also verifiable from the data) that ships sailing downstream stay close to the center, where water speed is maximum in magnitude, which is desirable while it aligns in direction with the sailing.
- Ships that sail upstream usually stay close to the edge due to the lower water speed there, from the data it is possible to observe a bimodal distribution of trajectories, ships sailing upstream often choose to stay close to one of the riverbanks, where water speed is lower.

- While sailing upstream there exists a tradeoff between sailing close the outer or inner shore. Following the inner trajectory, the distance traveled will be minimum, also the speed of the current will be minimum, but the inertia of the water during the curve will push it towards the outside, making the inner trajectory also more shallow, this can be a particularly important variable for big ships transporting a lot of cargo. Following the outer trajectory means finding a higher current speed (even if still lower than being in the middle) and having to sail a longer distance, but the depth of the river is more likely to be appropriate in all conditions.

Furthermore frequently changing side may be technically challenging, or dangerous due to traffic.

Figure 3 shows what an inner and outer trajectory often look like in the same river segment.

Figure 6: BLUE: Inner trajectory while sailing upstream, RED: outer trajectory while sailing upstream

It is possible to leverage the fact that outer trajectories are longer than inner ones and use `distance`, which is defined as the length of the linestring as a proxy for trajectory.

In fact, plotting the distance distributions of linestrings going over the same edge as in Figure 6, it is possible to observe the bimodal distribution shown in Figure 7

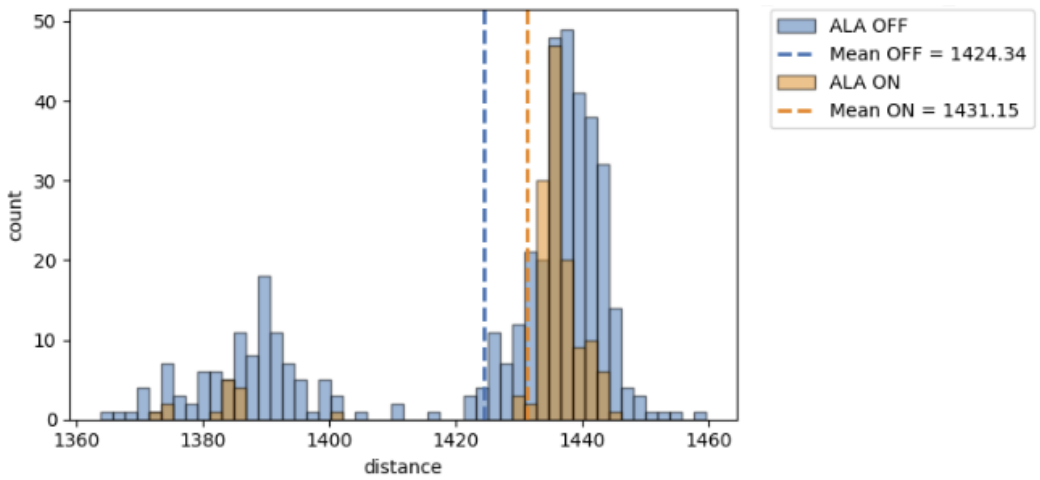


Figure 7: Distribution of the length of the linestring (distance) for traversals

It is clear form these plots that the variable `distance` is correlated with the trajectory choice and is a proxy for it.
Nevertheless `distance`, being the length of the linestring of the traversal, may also be influenced by the use of the Autonomous Lane Assist, since a higher/less efficient rudder activity, also results in undesired change of heading, as verified from the linestrings using the `hc_mean_rel` and capturing the mean of the change of heading as compared to the mean heading change computed over all the traversals over that edge. Figure 8 shows how the traversal that use the ALA display a relative heading change lower than the traversal that do not use the ALA.

Therefore `ALA_on` has an effect on `distance` and through it on `fuel_consumption` . This effect is not the main driver of the relation between `ALA_on` and `fuel_consumption`, but is still present.
The modeling approach shown in this DAG, ignores this.

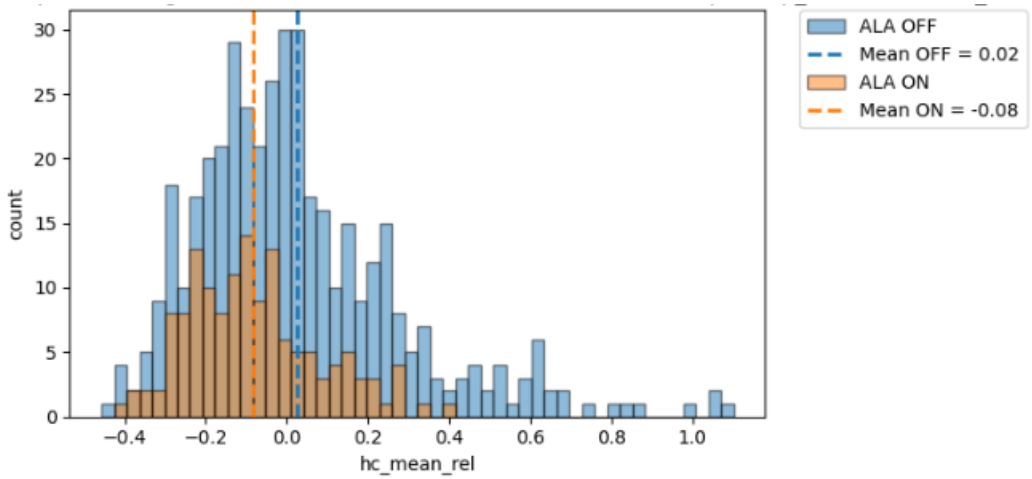


Figure 8: Distribution of relative heading change over an edge

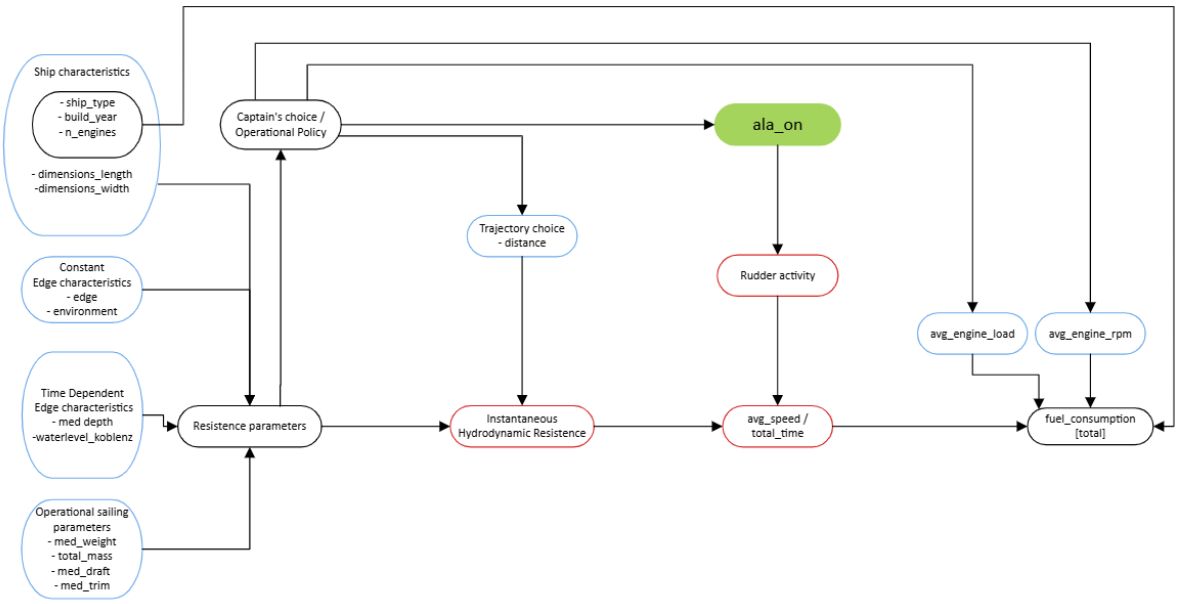


Figure 9: Directed Acyclic Graph describing the proposed causal relation between variables. In Blue are confounders(L), in Red are Mediators (M)

Mediators and the Mediator Pathway. A **mediator** is a variable that lies on the causal pathway between treatment and outcome. Mediators transmit part of the causal effect of the treatment on the outcome, allowing one to decompose total effects into direct and indirect components [8]. In this study, understanding and correctly defining the mediator pathway is particularly important, as it reflects the physical mechanism through which the ALA can influence fuel consumption.

Captains control the ship's propulsion via the **revolutions per minute (RPM)** of the engine, adjusted manually through the throttle. The onboard controller then automatically injects the appropriate amount of fuel to maintain the desired RPM, effectively adjusting the **engine load**. In normal sailing conditions, engine load remains nearly constant while RPM is stable. The scatter plot in Figure 10 shows how the variables `avg_engine_rpm` and `avg_engine_load` relate for a representative ship from the dataset. The relationship is almost one-to-one, with small deviations explained by the aggregation process used to obtain traversal-level averages.

The combination of RPM and load fully defines the **instantaneous fuel consumption** (in liters per hour). Indeed, captains often describe operating conditions in these terms, for example "doing 100 L/h" or "120 L/h," acknowledging that fixing the RPM effectively fixes the hourly fuel consumption under given conditions. Consequently, the final two arrows pointing into `total fuel consumption` in the DAG originate from `avg_engine_rpm` and `avg_engine_load`. Together, these determine the rate of fuel consumption, while the variable `total time` provides the final component: multiplying instantaneous consumption (L/h) by duration (h) yields total fuel consumed (L).

Captain's choice Another important observation is on the role of the block named "Captain's choice" , in this implementation, it is not a variable itself, but simply a bundle term fro all the variables pointing into it.
`Captain's choice` flows into `ALA_on`, `avg_engine_load`, `avg_engine_rpm` and `distance`(a proxy for trajectory choice) as the captain influences all of these directly.

As an input, instead, the `Captain's choice` block receives ship characteristics, constant and time dependent edge and environment characteristics and operational sailing conditions, symbolizing the fact that captains make decisions based on all these factors.

Proposed Mediating Mechanism. Based on first-principles reasoning, if the ALA affects fuel consumption, it must do so through a reduction in rudder activity. By optimizing rudder control, the ALA reduces lateral energy losses due to hydrodynamic resistance, effectively allowing the ship to travel faster at the same engine RPM and load. In other words, with smoother steering, less energy is wasted on corrective maneuvers, leading to shorter traversal times for the same propulsion effort. This mechanism links ALA activation to `fuel consumption` through its effect on `speed` (or equivalently `total time`) and hydrodynamic efficiency.

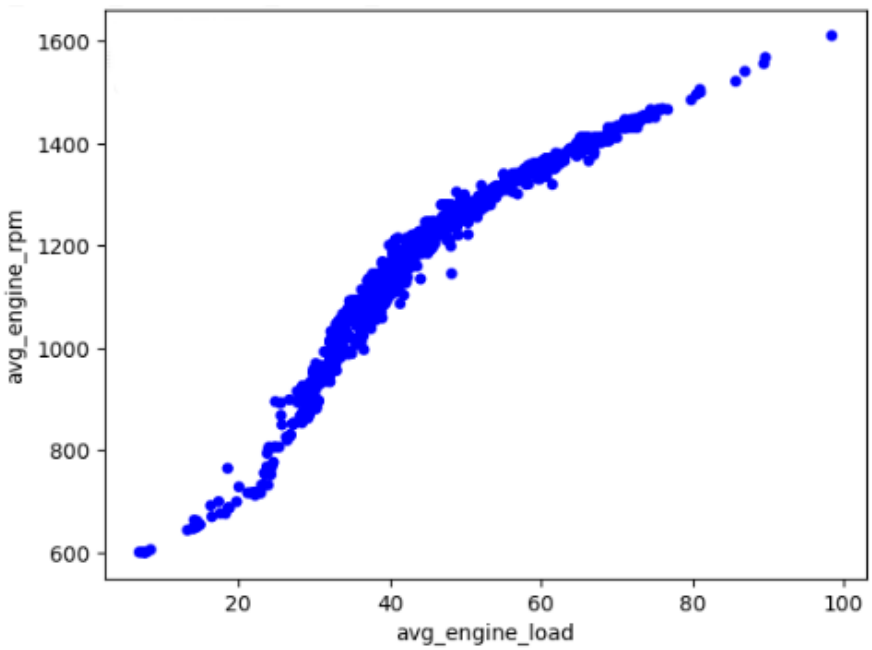


Figure 10: Scatter plot of `avg_engine_rpm` versus `avg_engine_load`, every point represents an observation from traversal data.

This reasoning is consistent with both physical intuition and the practical experience of ship captains, who often remark that excessive rudder movement "breaks the ship" or "pulls the ship back." It also aligns with industry understanding and previous studies reviewed in Section 6, which agree that smoother trajectories and reduced rudder angles are the primary mechanisms through which the Autonomous Lane Assist system achieves measurable fuel savings.

8.1.7 Model Specification

The predictive step of the g-computation procedure requires a flexible supervised learning model capable of estimating the conditional expectation

$$E[Y \mid A, L],$$

while capturing nonlinear interactions between ship characteristics, voyage conditions, and environmental covariates. In the present work, a **CatBoost Gradient Boosted Decision Tree** model was adopted.

CatBoost is an ensemble method based on gradient boosting over oblivious decision trees. Its main distinguishing feature is its *native handling of categorical variables*, achieved through an ordered target-based encoding scheme that avoids target leakage and reduces variance. This is particularly advantageous in this dataset, where several key predictors are categorical with high cardinality (`device_id`, `ship_type`, `environment`, `edge`) and exhibit complex

interactions with continuous variables such as engine load, draft, and water level.

CatBoost incrementally fits trees to the negative gradients of the loss function. For regression, the model minimizes the squared error loss

$$\mathcal{L} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

with predictions updated at boosting iteration t according to

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \, f_t(x_i),$$

where η is the learning rate and f_t is the t-th regression tree.

Model choice and data preparation. CatBoost was selected for three main reasons:

1. **Native categorical processing**. Unlike XGBoost or LightGBM, CatBoost does not require one-hot encoding or ordinal encoding. The algorithm performs an ordered statistical encoding internally, which preserves information content and stabilizes training when categorical cardinality is high.

2. **Robustness to missing values**. CatBoost handles missing entries internally by learning optimal default leaf assignments during tree construction. In this dataset, missingness was present but not substantial, for example:

   - `med_weight`: 5.58% missing
   - `med_trim`: 0.13% missing
   - `build_year`: 9.79% missing

   No explicit imputation was therefore required prior to model fitting.

3. **Performance on structured/tabular data**. The model consistently outperformed alternative approaches considered during exploratory modeling, offering lower residual variance and more stable cross-validation error across devices and edges.

The data cleaning and filtering steps applied before training are identical to those described in Section 7.3. After preprocessing, both categorical and continuous predictors were fed directly into CatBoost without additional encoding.

Interpretability and computational considerations. Oblivious trees used by CatBoost apply the same splitting condition at every level of the tree, which improves training speed and yields models that are easier to interpret in terms of global feature importance and monotonic patterns. This architecture is computationally efficient and well suited for a large per-edge dataset.

Future work may reintroduce sequence-based models to capture temporal correlations between consecutive river segments, but for the present g-computation procedure,focused on predicting counterfactual fuel consumption for isolated traversals,the CatBoost specification offers an appropriate balance between accuracy, robustness, and complexity.

Features and target. The model was trained using a comprehensive set of predictors capturing vessel characteristics, hydrodynamics, loading state, environmental conditions, trajectory information, and engine behaviour. Specifically, the feature vector included:

- **Vessel characteristics:** `device_id`, `n_engines`, `build_year`, `ship_type`, `dimensions_width`, `dimensions_length`.
- **Hydrodynamics:** `med_trim`, `med_depth`, `waterlevel_koblenz`.
- **Loading & mass:** `total_mass`, `med_weight`, `med_draft`.
- **Environmental conditions:** `environment`(upstream, downstream or tidal), `edge`.
- **ALA activation:** `ala_on`.
- **Trajectory information:** `distance`.
- **Engine behaviour:** `avg_engine_rpm`.

This feature set collectively describes the vessel type, its loading condition, the hydrodynamic state of the waterway, external environmental influences, the specific river segment traversed, and the instantaneous operational regime of the engine.

The treatment variable `ala_on` indicates whether the Autonomous Lane Assist was active.
The **target variable** considered is total fuel consumption in the traversal per unit distance(`fuel_consumption_km`), consistently with the targets used in the means analysis and Inverse Probability weighting(IPW) benchmarks.

Model training and evaluation. Model performance was evaluated using grouped five-fold cross-validation, split randomly. For each fold, a CatBoost model was trained independently and evaluated with several regression metrics.

The model exhibited extremely high stability and accuracy across folds:

$$R^2_{CV} = 0.996 \pm 0.016, \qquad \text{RMSE}_{CV} = 0.486, \qquad \text{RMPE}_{CV} = 0.045.$$

The distribution of percentage errors also confirmed excellent consistency:

$$\text{Median absolute \% error} = 0.022, \qquad \text{95th percentile absolute \% error} = 0.080.$$

During cross-validation, early stopping selected an optimal number of boosting iterations for each fold, with an average of approximately 1999 iterations. These fold-specific best iteration values were stored, and the final model used for g-computation was trained using the **median** of these iteration counts to obtain a stable choice of model complexity.

Before fitting the final model, a focused hyperparameter search was performed.
The selected configuration balanced model flexibility and overfitting control: a tree depth

of 7, a learning rate of 0.05, and a moderate regularization strength. Subsampling of both observations and features was also applied to improve robustness. These hyperparameters were chosen because they consistently produced low validation error while remaining stable across folds, and a low for a reasonable training time.

Finally, the model was retrained on the full dataset using these settings. Its role in this thesis is not to achieve the highest possible predictive accuracy, but to serve as a reliable estimator of the conditional expectation $E[Y \mid A, L]$. In g-computation, the primary requirement is unbiased and well-calibrated counterfactual prediction, not predictive optimisation.

Residual diagnostics and treatment-specific bias. An essential requirement for using a predictive model within the g-computation framework is that the residual error must not systematically depend on the treatment variable (`ala_on`). If the model were to under-predict or over-predict fuel consumption differently for ALA-on versus ALA-off traversals, the resulting counterfactual contrasts would be biased, directly contaminating the estimated ATE.

To verify this, the distribution of percentage errors was analysed separately for the two treatment states. Empirical error distributions were plotted in an overlapping graph in Figure 11 . The results show that the residuals are centred symmetrically around zero for both conditions, with nearly identical dispersion.

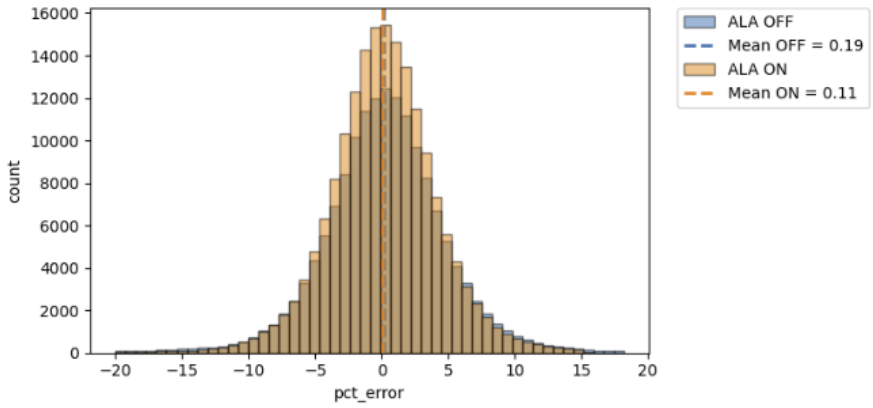


Figure 11: Distribution of percentage error across observations with `ALA_on` and `ALA_off`. The distributions almost perfectly overlap.

Quantitatively, the mean absolute percentage error is approximately 3.13% for ALA-off and 2.75% for ALA-on, with corresponding standard deviations of 3.64% and 2.72%. These differences are negligible relative to the overall error scale and well within the noise expected from operational variability.

This confirms that the model does not exhibit treatment-dependent bias: prediction errors are statistically similar regardless of whether ALA was active or not. Consequently, the fitted CatBoost model satisfies the key requirement of residual neutrality and is therefore

appropriate for use as the conditional expectation estimator $E[Y \mid A, L]$ in the g-computation procedure.

Feature importance and SHAP analysis. To assess whether the model captured relationships compatible with physical intuition, a SHAP (SHapley Additive exPlanations) analysis was performed on the final CatBoost model, shown in Figure 12. The global SHAP summary plot illustrates, for each feature, both the magnitude of its impact on the prediction and the direction in which high or low feature values shift fuel consumption.

The dominant phisical drivers of the model's output are the engine-related variables: `avg_engine_load` and `avg_engine_rpm`. High values of these features consistently push predictions upward, reflecting the fundamental mechanical relationship between propulsion effort and fuel use. The `environment` and `edge` variables also show strong and wide-spread SHAP contributions, confirming that river-flow regime (upstream, downstream, tidal) and local segment-specific hydrodynamic conditions exert major influence on fuel consumption.

Ship- and trip-level characteristics such as `total_mass`, `ship_type`, `build_year`, and `device_id` have moderate but meaningful effects, consistent with the idea that vessel design, loading state, and behavioural differences across ships introduce systematic variability. Dimensional and hydrodynamic indicators (`dimensions_width`, `dimensions_length`, `med_depth`, `med_weight`, `waterlevel_koblenz`, `med_trim`, `med_draft`) exhibit smaller but coherent contributions, indicating that the model incorporates them without overamplifying noise.

Importantly, the treatment variable `ala_on` appears at the bottom of the importance ranking and produces only very small SHAP deviations. This is expected: the ALA effect is subtle compared to the dominant physical drivers of fuel consumption. Its low SHAP influence reinforces the necessity of g-computation, as such a small causal effect cannot be reliably extracted from direct model-based interpretation alone. The SHAP patterns therefore confirm that the model learned physically consistent relationships while not embedding spurious or exaggerated treatment effects, supporting its suitability for counterfactual inference.

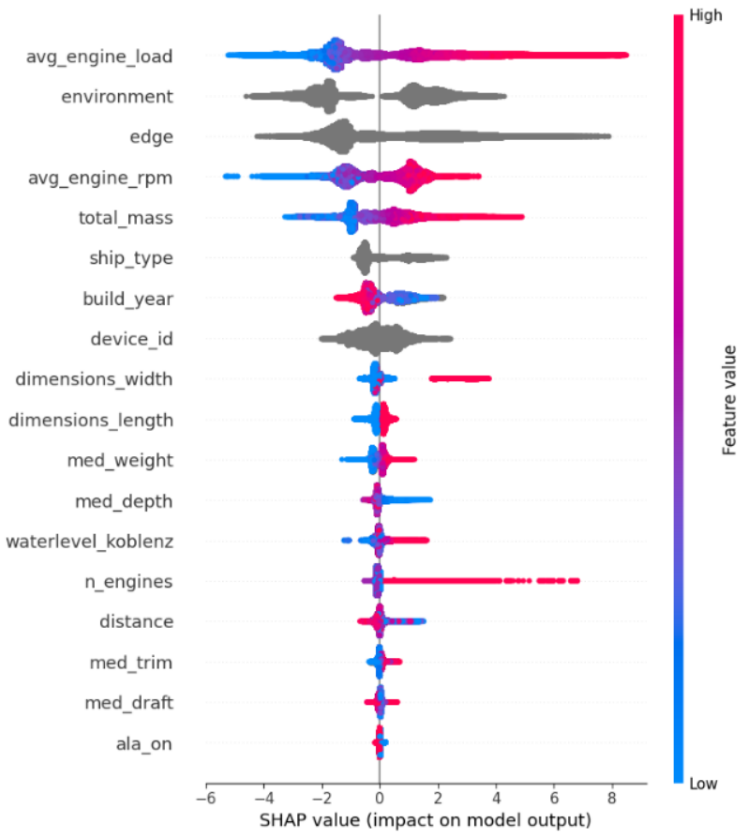


Figure 12: SHAP analysis performed on the Cat Boosted decision tree model used for g-computation

Collinearity considerations. No feature reduction or decorrelation step was applied prior to model training, as CatBoost is inherently robust to multicollinearity.

In this dataset, several variables,such as `med_draft`, `med_weight`, and `total_mass`,describe closely related aspects of the same physical quantity: the vessel's loading condition and hydrodynamic immersion. Their high correlation is a natural property of the system rather than a data quality issue.

While traditional linear models may suffer from coefficient instability in the presence of colinear predictors, tree-based models like CatBoost distribute importance across correlated features without compromising predictive accuracy.

In the SHAP analysis, these variables should therefore be interpreted collectively as one conceptual unit rather than as independent contributors. The model may attribute more influence to one or another depending on the specific partitioning of the feature space, but together they represent the same underlying physical mechanism linking vessel load to fuel consumption.

Environment-specific performance and model specification. The strong SHAP dominance of `environment` suggested potential heterogeneity in residual behavior across different sailing regimes.

To assess this, residuals were analyzed per environment. The global residual statistics (mean residuals close to zero: -0.20 downstream, -0.22 tidal, -0.11 upstream) and balanced distributions across conditions confirmed that no major systematic bias was present. However, RMSE values (1.67 downstream, 3.10 tidal, 4.22 upstream) indicated varying predictability due to physical variability,particularly higher uncertainty in upstream conditions, where resistance and maneuvering effects have a more unpredictable effect.

Figure 13 shows the residuals per environment.

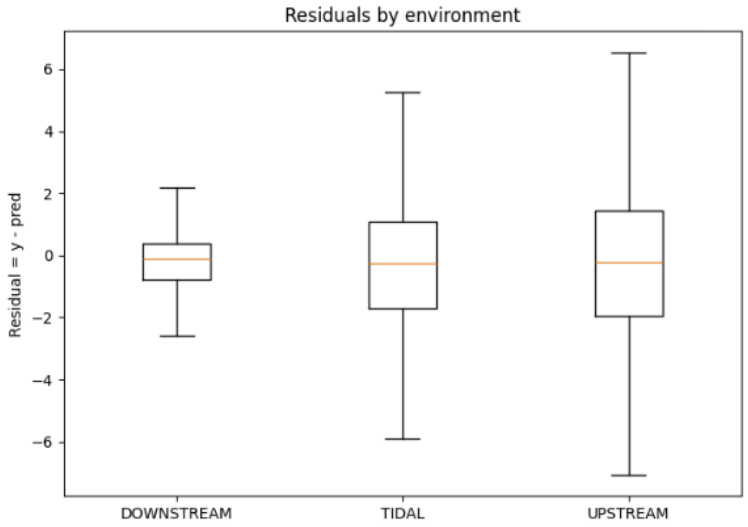


Figure 13: Residuals per environment in the general Cat Boosted decision tree model.

These results motivated the training of separate models for each environment. The per-environment models slightly reduced residual variance and improved interpretability, confirming that environment-specific models can better represent the unique hydrodynamic regimes of each sailing direction. This step further validated that the global model was correctly specified but benefited from stratified refinement.

Interpretation Correct specification in this context means that the conditional expectation $\widehat{E}[Y \mid A, L]$ estimated by the model is unbiased for the true $E[Y \mid A, L]$, i.e.,

$$E[Y - \widehat{E}[Y \mid A, L] \mid A, L] = 0.$$

Although this cannot be proven formally, several indicators support this assumption: the high and stable R^2, near-zero mean residuals, physically coherent SHAP feature ranking, and consistent behavior across environments. Together, these elements indicate that the

CatBoosted decision tree model is both well-specified and interpretable, providing a reliable estimator of $E[Y \mid A, L]$ suitable for use within the g-computation framework.

8.1.8 Methodology Validation

Before interpreting the causal estimates obtained through the g–computation pipeline, it is essential to verify that the procedure is capable of (i) reporting no effect when none exists and (ii) recovering a known effect when one is present.
This validation step is particularly important in the present context, because the estimated ALA effects on the actual dataset are small in magnitude.
When estimated effects lie close to the noise floor of the model, methodological validation provides evidence that the pipeline is not *structurally biased* toward producing spurious savings, and that it would indeed detect an effect if one existed.

To this end, three controlled validation experiments were conducted. Each experiment tests a different property of the pipeline: (1) ability to report no effect in a dataset with no signal; (2) ability to recover an artificially injected effect in a dataset with no other structure; and (3) ability to recover the injected effect in the original dataset, where all genuine relationships between covariates and outcome remain intact. In all experiments, the same CatBoost model and the same g–computation procedure used in the main analysis were applied unchanged.

1. Dataset with no treatment effect (randomized target). A synthetic dataset with *no causal structure* was generated by randomly shuffling the outcome variable `fuel_consumption_km` across all observations. This removes *all* relationships between covariates and the outcome, including any dependence on `ala_on`. In this setting, a well-specified estimator must (i) fail to learn predictive structure and (ii) yield an average treatment effect (ATE) statistically indistinguishable from zero.

The CatBoost outcome model performed exactly as expected, with

$$\text{CV } R^2 = 0.000 \pm 0.000, \qquad \text{CV RMSE} = 8.352, \qquad \text{CV RMPE} = 1.964,$$

indicating absence of learnable structure. The g–computation ATE was

$$\widehat{\text{ATE}}_{\text{global}} = 0.09\% \pm 0.35\%,$$

fully consistent with a null effect. This confirms that the estimator does not hallucinate savings in the absence of signal.

A SHAP analysis of this model shows no meaningful feature contributions; specifically, `ala_on` has no predictive influence, as desired.

2. Injected 5% effect into the randomized-target dataset. To verify that the pipeline *detects* an effect when one exists, a 5% fuelsaving effect was injected into the same randomized dataset by multiplying fuel consumption by 0.95 for all rows with `ala_on = 1`. This creates a dataset where: (i) no variable except `ala_on` contains information about the outcome, and (ii) the magnitude of the true ATE is exactly 5%.

As expected, the outcome model again showed no ability to learn patterns (since none exist other than the injected effect):

$$\text{CV } R^2 = 0.001 \pm 0.000, \qquad \text{CV RMSE} = 7.219, \qquad \text{CV RMPE} = 1.471.$$

The corresponding SHAP analysis (Fig. 14) shows that `ala_on` is the *only* feature with predictive contribution, validating that the model correctly identifies the only available signal.

The recovered ATE was:

$$\widehat{\text{ATE}}_{\text{global}} = 4.87\% \pm 0.62\%,$$

which is extremely close to the ground-truth injected effect of 5%. This confirms that the g-computation pipeline is sensitive to treatment effects of the magnitude relevant for this study.

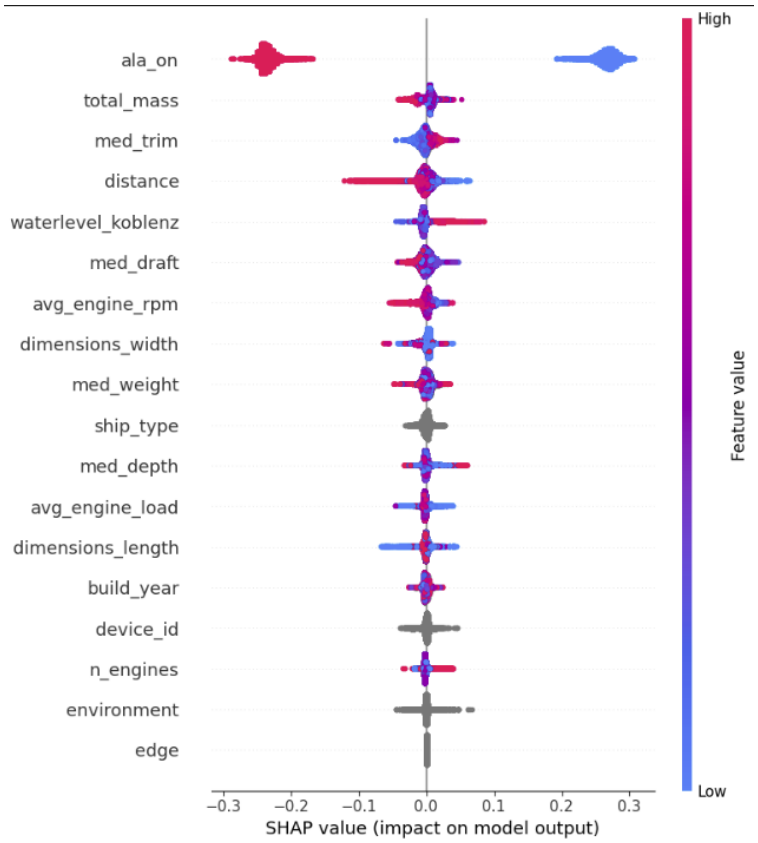


Figure 14: SHAP analysis on the CatBoost model trained on randomized outcome dataset with a 5% injected effect. `ala_on` is the only feature that has any real relation to the outcome, and the CatBoost reflects it correctly.

3. Injected 5% effect into the original dataset. Finally, a 5% fuelsaving effect was injected into the *real* dataset used for the main g-computation analysis, preserving all existing

relationships between fuel consumption and the covariates. For all rows with `ala_on = 1`, the fuel consumption was again multiplied by 0.95.

In this setting, the model retains its full predictive power:

$$\text{CV } R^2 = 0.996 \pm 0.000, \qquad \text{CV RMSE} = 0.476, \qquad \text{CV RMPE} = 0.045,$$

with an absolute percentage error distribution compatible with the baseline model (SD 0.034, P50 = 0.022, P95 = 0.081). The SHAP analysis (Fig. 15) now shows that `ala_on` becomes a substantially more important predictor, correctly rising above variables such as vessel dimensions, water level, and keel clearance.

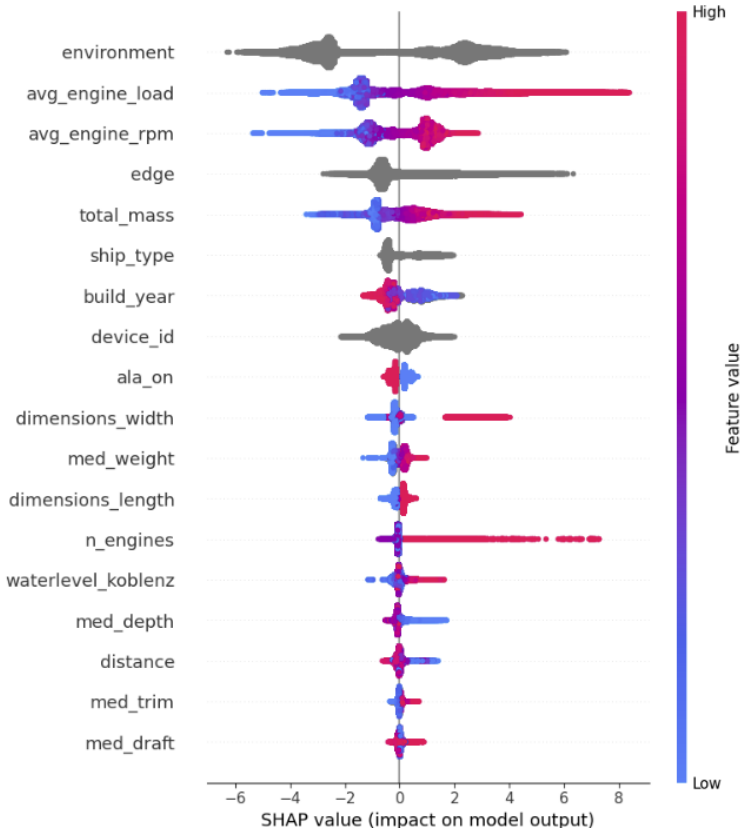


Figure 15: SHAP analysis on the CatBoost model trained on the real dataset with a 5% injected effect. The effect of `ala_on` is correctly shown within the context of the other important features.

The estimated ATE was:

$$\widehat{\text{ATE}}_{\text{global}} = 5.26\% \pm 1.12\%,$$

matching the true injected effect.

Summary and implications. Across all three experiments, the g-computation pipeline behaves as a correct causal estimator should:

- It reports no effect when none exists (Experiment 1).
- It accurately recovers an injected effect in a dataset with no other structure (Experiment 2).
- It accurately recovers the same effect in the real dataset with all genuine physical relationships preserved (Experiment 3).

Importantly, the SHAP analyses consistently confirm that:

1. when no causal signal exists, `ala_on` has no predictive influence;
2. when the only existing signal is the injected treatment effect, `ala_on` becomes the sole important feature;
3. in the real-data injection experiment, `ala_on` becomes substantially more influential but remains correctly contextualized among other physically meaningful predictors.

These results demonstrate that the CatBoost outcome model underlying the g-computation framework is capable of (i) detecting the ALA signal when present, (ii) ranking it correctly relative to other explanatory variables, and (iii) returning near-zero effects when no true effect exists. Therefore, the g-computation pipeline used in the main analysis can be reasonably trusted to produce valid causal estimates for the actual ALA dataset.

8.2 Predictive Modeling

In the context of this study, **prediction** refers to estimating the expected total fuel consumption of a voyage *before it begins*, using only information that is available at the time of departure. The goal is to provide an estimate of fuel use that reflects realistic operating conditions while being computed solely from pre-departure variables.

Among the four research dimensions guiding the literature review (causal inference, predictive modeling, evaluation methods, and automation impact), this predictive modeling task is the area most directly informed by existing literature.
Many prior studies in maritime fuel prediction address similar operational challenges, focusing on short- and long-term consumption forecasts using voyage and environmental parameters. These works served as a key reference point for the development and evaluation of the predictive models.

For this study, three predictive models are employed:

- a **linear regression** model, used as a transparent and interpretable baseline;
- a **CatBoost decision-tree model**, of the same model family adopted in the G-computation framework for causal inference;
- a **Deep Backpropagation Neural Network (DBPNN)**, selected based on its prevalence and demonstrated performance in the fuel-consumption literature (see Section 5.2).

All models are trained on the same dataset described in Section 7. Apart from the standard cleaning steps applied uniformly across this thesis,removal of outliers, inconsistent entries, and erroneous measurements,no additional preprocessing is performed.
In particular, *no positivity-related adjustments* are required in this context, as the predictive task does not involve counterfactual estimation or treatment-model constraints.
The predictive models are therefore trained on the cleaned operational dataset as it naturally occurs, without rebalancing or enforcing any treatment-coverage requirements.

8.2.1 Modeling Goals and Assumptions

The objective of the predictive modeling component is to develop a *pre-departure* tool capable of estimating the expected fuel consumption of an inland-waterway voyage before it begins.
The envisioned practical use case is operational planning: captains, fleet managers, and voyage planners should be able to input the intended route and operating parameters and obtain a reliable estimate of the fuel required for the trip. Consequently, the modeling goal is not purely statistical accuracy, but decision usefulness.
In this context, predictive performance is evaluated not only by numerical error but by whether the model can operate using only information available at departure time, whether it is interpretable, and whether it is deployable in operational workflows.

To satisfy these requirements, several assumptions define the structure and scope of the predictive models used in this thesis.

Target variable. All predictive models are trained to estimate *total fuel consumption per traversal* (liters), not fuel consumption per unit distance. This aligns with the intended deployment scenario: voyage-level predictions are formed by summing per-edge fuel consumption predictions along the planned route.

Feature set and pre-departure availability. The feature set is identical across all models and includes only variables that are assumed to be known at the time of departure. These are grouped as follows:

- **Ship characteristics:** `device_id`, `n_engines`, `build_year`, `ship_type`, `dimensions_width`, `dimensions_length`.
- **Loading and hydrostatic information (pre-departure estimates):** `total_mass`, `med_weight`, and the forecasted `waterlevel_koblenz`.
- **Environmental context:** `environment` (upstream, downstream, or tidal) and `edge`(river section).
- **Trajectory definition:** and `distance`, both determined from the user's selected start and end locations.
- **Control and automation variables:** `ala_on` and the user-specified `avg_engine_rpm`.

In practice, ship characteristics, route geometry, and the environment category are fixed for a planned voyage. The user supplies only three inputs: the origin–destination pair (from which the sequence of edges is automatically determined), the intended average engine RPM for the trip, and the cargo weight to be transported.

A practical assumption is therefore that vessels using the model have these quantities measured, recorded, or otherwise retrievable before departure.

RPM as a scenario input. A key modeling assumption is that `avg_engine_rpm` is an *exogenous scenario variable*: users choose a target RPM that they intend to maintain throughout the voyage. This assumption is operationally realistic for inland vessels.
Captains do not directly command engine load or propulsion power; instead, they input a desired RPM setpoint, and an internal control system automatically adjusts fuel injection and load to achieve the commanded RPM.
Thus, in real operation, desired RPM is the primary actionable control variable, and treating it as a user-specified input rather than an emergent consequence of other variables is appropriate.

Implicitly, the model assumes:

- the vessel is physically capable of maintaining the selected RPM under the encountered conditions;
- the RPM setpoint is stable enough across edges that an average value is meaningful for prediction.

Additivity of per-edge predictions. Voyage-level consumption is estimated by summing predicted fuel consumption across the edges of the planned route. This requires the assumption that per-edge consumption is sufficiently independent that

$$\widehat{F}_{\text{voyage}} = \sum_{e \in \text{route}} \widehat{F}_e$$

approximates the true voyage-level consumption. This assumption reflects the operational segmentation already used by Shipping Technology's navigation system and is consistent with how fuel consumption accumulates along a river trajectory.

Forecasting and measurement assumptions. Because some pre-departure inputs (e.g., `waterlevel_koblenz`) are not exactly known at planning time, the model assumes that forecasts or proxy values are sufficiently accurate to support useful predictions.
Moreover, the validity of the predictive model depends on reliable measurement of ship characteristics, loading estimates, and route planning.

Distributional stability. The predictive framework assumes that the statistical relationship learned from historical data generalizes to deployment conditions.
Formally, if P_{train} and P_{deploy} denote the joint distributions of features and fuel consumption during training and deployment, the model assumes that

$$P_{\text{train}}(Y \mid X) \approx P_{\text{deploy}}(Y \mid X),$$

so that patterns learned from historical traversals remain valid for future voyages. This assumption is standard but necessary to justify the use of historical data for pre-departure prediction.

Overall, these modeling goals and assumptions reflect the intended operational use case: the focus is on constructing a predictive model that is practically deployable, interpretable, and based only on information available before the voyage begins, even at the cost of marginally lower statistical accuracy compared to models that rely on full in-voyage sensor traces.

8.2.2 GLEC Framework

The Global Logistics Emissions Council (GLEC) Framework is the globally recognized standard for calculating greenhouse gas (GHG) emissions across all logistics modes.

For inland waterway transport (IWT), the GLEC Framework provides a set of *default emission intensity factors* expressed in grams of CO_2e per tonne-kilometre (g/tkm).
These intensities represent the expected GHG emissions per unit of freight moved and allow standardised, comparable benchmarking across vessel types, sizes, and operational profiles.

The factors used in this study are derived from the official GLEC IWT report [16], which aggregates real-world operational and fuel-consumption data from European inland cargo vessels, tankers, container ships, and convoy configurations. These values represent Well-to-Propeller (WTP) emission intensities, using a diesel greenhouse gas factor of:

$$\mathrm{EF}_{\mathrm{diesel}}^{\mathrm{WTP}} = 3240\,\mathrm{g\,CO_2e\,L^{-1}}$$

as specified in [16]. This same factor is used in this research to convert measured and predicted fuel consumption into comparable WTP GHG emissions.

GLEC Intensity Categories for Inland Waterways The GLEC Inland Waterway Transport methodology distinguishes vessel classes primarily according to vessel type and nominal vessel length. For this research, three relevant cargo categories are present in the dataset: *general cargo*, *tanker*, and *container* vessels.
For each of these categories, the GLEC Framework provides g/tkm intensities as shown in Tables 6, 7, and 8, extracted from [16].

Table 6: GLEC emission intensities for cargo (dry bulk / general cargo) vessels [16].

Vessel class	**GHG Intensity [g/tkm]**
Motor vessel ≤ 80 m	29.5
Motor vessel 85-86 m	20.7
Motor vessel 87-109 m	18.4
Motor vessel 110 m	18.4
Motor vessel 135 m	19.0
Coupled convoy (163-185 m)	17.0
Pushed convoy (+2 barges)	17.3
Pushed convoy (+4-5 barges)	9.7
Pushed convoy (+6 barges)	7.4

These intensities form the GLEC-based baseline against which the measured and predicted emissions of each traversal in this study are compared.

Table 7: GLEC emission intensities for tanker vessels [16].

Vessel class	**GHG Intensity [g/tkm]**
Tanker vessel 110 m	18.7
Tanker vessel 135 m	22.0

Table 8: GLEC emission intensities for container vessels [16].

Vessel class	**GHG Intensity [g/tkm]**
Container vessel 110 m	25.5
Container vessel 135 m	19.8
Container convoy (185 m)	19.7

Computation of GLEC Emissions per Traversal For each traversal in the dataset, the GLEC Framework prescribes computing theoretical emissions using:

$$\text{Emissions}_{\text{GLEC}} = \text{Intensity}_{\text{GLEC}} \times (\text{tonnes}) \times (\text{distance}\,[\text{km}])$$

where

$$\text{tonnes} = \frac{\text{med\_weight}\,[\text{kg}]}{1000}, \qquad \text{distance}\,[\text{km}] = \frac{\text{distance}\,[\text{m}]}{1000}.$$

Thus, for traversal i:

$$\text{Emissions}_{GLEC,i} = \text{glec\_intensity\_tkm}_i \times \left(\frac{\text{med\_weight}_i}{1000}\right) \times \left(\frac{\text{distance}_i}{1000}\right).$$

This yields the GLEC-prescribed emissions in grams of CO_2e for that traversal, representing the expected performance of a vessel of the same class under typical operational conditions.

Computation of Measured Emissions Measured GHG emissions are computed from recorded fuel consumption using the same WTP emission factor [16]:

$$\text{Emissions}_{\text{fuel}} = \text{fuel\_consumption}\,[\text{L}] \times 3240\ \frac{\text{g}\,CO_2e}{\text{L}}.$$

This produces directly comparable units $[g]$, enabling a consistent comparison with the GLEC baseline.

Comparison with Predicted and Measured Emissions The predictive models developed in this research output fuel consumption in liters per traversal. To evaluate their environmental implications, both the predicted fuel consumption and the actual measured fuel consumption are converted into WTP GHG emissions using the GLEC diesel factor [16]:

$$\text{Emissions}_{\text{pred}}, \text{i} = \hat{F}_i \times 3240 \text{ g}, \qquad \text{Emissions}_{\text{meas}}, \text{i} = F_i^{\text{meas}} \times 3240 \text{ g}.$$

These quantities represent, respectively, the model-estimated and the real operational emissions for traversal i. Both can be contrasted with the GLEC baseline emissions:

$$\text{Emissions}_{\text{GLEC}}, \text{i} = \text{glec\_intensity\_tkm}_i \times \left(\frac{\text{med\_weight}_i}{1000}\right) \times \left(\frac{\text{distance}_i}{1000}\right).$$

This enables a structured three-way comparison. First, the discrepancy between predicted and measured emissions indicates how accurately the fuel-consumption models capture real vessel behaviour:

$$\Delta_i^{\text{model}} = \text{Emissions}_{\text{pred}}, \text{i} - \text{Emissions}_{\text{meas}}, \text{i}.$$

Second, the deviation of the GLEC baseline from measured emissions shows how well the GLEC intensity classes approximate the real operational profile of each vessel:

$$\Delta_i^{\text{GLEC}} = \text{Emissions}_{\text{GLEC}}, \text{i} - \text{Emissions}_{\text{meas}}, \text{i}.$$

Finally, comparing the model-predicted emissions to GLEC emissions,

$$\Delta_i^{\text{pred-GLEC}} = \text{Emissions}_{\text{pred}}, \text{i} - \text{Emissions}_{\text{GLEC}}, \text{i},$$

provides insight into whether the machine-learning models align more closely with actual vessel performance or with the standardized GLEC expectations. This framework also allows evaluating how the Autonomous Lane Assist (ALA) system influences operational emissions relative to both real-world behaviour and the GLEC benchmark.

8.2.3 Linear Regression

To establish a transparent and interpretable benchmark, a multiple linear regression model was trained to predict fuel consumption using the same set of features later employed in the CatBoosted decision tree and the DBPNN models.

Linear regression was selected as a baseline because it provides an analytically simple reference for assessing the added value of nonlinear models, while still offering insight into the direct linear relationships between predictors and fuel consumption.

Data Preparation Unlike tree-based models, linear regression is sensitive to missing data, multicollinearity, and unscaled categorical variables. Therefore, additional data cleaning and preprocessing steps were necessary. The following predictors were included in the model:

> `dimensions_length`, `dimensions_width`, `med_trim`, `med_depth`, `waterlevel_koblenz`, `n_engines`, `build_year`, `ship_type`, `total_mass`, `med_weight`, `med_draft`, `environment`, `edge`, `ala_on`, `distance`, `avg_engine_rpm`, `avg_engine_load`.

Missing values were handled explicitly to avoid bias and ensure model stability:

- Rows with missing `waterlevel_koblenz` were completely removed, as this feature is critical for representing hydrodynamic resistance.
- Missing `med_depth` values were replaced by the median depth observed for the same river segment (`edge`) and a binary indicator variable `med_depth_missing` was added.
- Missing `build_year` values were replaced by the median build year across the dataset, with a corresponding indicator variable `build_year_missing`.

Categorical variables were treated according to their cardinality:

- High-cardinality features (`edge`, `prev_edge`, `next_edge`, `device_id`) were target-encoded with smoothing, reducing dimensionality while preserving the statistical relationship with the target.
- Low-cardinality features (`ship_type`, `environment`) were one-hot encoded. To prevent perfect multicollinearity (the "dummy variable trap"), one category per feature was dropped; for instance, `environment_TIDAL` was omitted, leaving the other two dummies (`UPSTREAM` and `DOWNSTREAM`) as active regressors.

After preprocessing, the dataset was split into training (80%) and testing (20%) subsets using random samplin.

Model Training and Evaluation The regression model was fit on the training set and evaluated using both in-sample and out-of-sample metrics. Root Mean Squared Error (RMSE) and the coefficient of determination (R^2) were computed to quantify overall model accuracy.

The out-of-sample results were:

$$\texttt{RMSE} = 3.999, \qquad R^2 = 0.9032.$$

To complement these scale-dependent metrics, percentage-based error measures were also computed. These provide a more interpretable assessment of relative predictive accuracy across vessels and operating conditions. The results were:

$$\texttt{RMPE} = 0.4471, \qquad \texttt{Median APE} = 0.1425, \qquad \texttt{95th percentile APE} = 0.9017.$$

The median absolute percentage error of 14.25% reflects reasonable central performance for such a simple model, but the much larger 95^{th} percentile error (90.17%) highlights substantial degradation in more challenging operating regimes.
These results reinforce that linear regression is not an adequate functional form for capturing the nonlinear dependencies governing fuel consumption; its usefulness is limited to providing a transparent baseline for comparison against more expressive models.

Coefficient Analysis. Because linear regression imposes a strictly additive and linear relationship between predictors and fuel consumption, the estimated coefficients reflect the best linear fit to a highly nonlinear system.
As a result, several coefficients appear counterintuitive or physically implausible. This does not indicate a real physical relationship but rather the limitation of using a linear model to approximate complex hydrodynamic and operational processes.

The largest negative coefficient is associated with `ship_type_passenger` (-3.65), which simply reflects average differences in consumption patterns across vessel categories.
More problematic is the strong negative coefficient for `med_draft` (-1.96). Physically, a deeper draft corresponds to higher displacement and typically *higher* resistance and fuel use. The negative sign is therefore not interpretable in hydrodynamic terms. It arises because linear regression can only attribute variation through straight-line effects; it cannot capture nonlinear interactions between draft, mass, speed, trim, and environmental conditions. In such cases, the model assigns compensatory coefficients that minimize global error even when the implied marginal effects are unrealistic.

Similarly, `med_trim` exhibits a large positive coefficient ($+1.96$). While extreme trim can indeed increase fuel consumption, the magnitude of this coefficient and its dominance relative to more physically meaningful variables confirm that the linear model is absorbing complex nonlinear behavior into a single linear parameter.
The same holds for the strong positive effect of `ship_type_container` ($+1.69$), as well as route- and vessel-specific dummies such as `edge_te` ($+1.22$) and `device_id_te` ($+0.71$). These terms capture systematic differences that the model cannot explain through the physical covariates alone.

Environmental effects instead comply with expected patterns: `environment_UPSTREAM` is positive ($+1.13$), which is expected `environment_DOWNSTREAM` is negative (-0.40), suggesting that downstream conditions lower consumption as compared to the TIDAL environment benchmark.

Engine-related variables follow a similar pattern. `avg_engine_load` shows a modest positive coefficient ($+0.17$), but `avg_engine_rpm` is almost zero ($+0.002$), and `n_engines` has a moderate effect ($+0.93$). These magnitudes do not correspond to realistic marginal influences on fuel consumption. Instead, they indicate that the model distributes explanatory power across correlated operational variables without capturing their true nonlinear dependence.

The coefficient for `ala_on` ($+0.13$) should likewise not be given any substantive interpretation. In this predictive setting, the coefficient reflects raw correlation rather than causal impact, and the sign is influenced by the contexts in which ALA is typically used.

Finally, variables such as `dimensions_width`, `dimensions_length`, `build_year`, `waterlevel_koblenz`, and all mass-related variables have coefficients near zero. This does *not* imply physical irrelevance; rather, the linear model fails to represent how these quantities influence fuel consumption through nonlinear and interacting mechanisms.

Overall, the coefficient structure highlights the fundamental limitation of linear regression for this task. A single global linear model cannot capture the nonlinear hydrodynamics, operational interactions, and vessel-specific effects inherent to inland navigation.
The resulting coefficients minimize prediction error but do not reflect interpretable or physically meaningful relationships. This motivates the use of more flexible models, such as CatBoost and DBPNN, which are better suited to learning the complex functional dependencies in the data.

8.2.4 CatBoost decision tree model

A gradient-boosted decision tree model based on the CatBoost framework was selected for this task.
The choice is motivated by its strong performance in the G-computation analysis, its ability to handle heterogeneous feature types, and its robustness to nonlinear interactions that are known to govern fuel consumption dynamics.
The model was trained on the same feature set and target variable described in Section 8.2.

Hyperparameter Selection A constrained hyperparameter search was performed due to computational limitations, exploring a small but targeted subset of depth, learning rate, regularization strength, subsampling rate, and feature sampling rate. The most influential hyperparameters are:

- **Tree depth (7):** controls model complexity and the ability to capture nonlinear relationships. Depth 7 provided a strong balance between expressiveness and overfitting risk.
- **Learning rate (0.05):** governs the incremental contribution of each boosting step. A relatively low value improved stability across folds.
- **L2 regularization (8.0):** penalizes large leaf values and mitigates overfitting in the presence of correlated operational features.
- **Subsample and feature sampling rates (0.8):** introduce randomness into row and feature selection, improving generalization and reducing variance.
- **Early stopping (patience 50):** prevents unnecessary boosting once validation performance plateaus.

The maximum number of boosting iterations was set to 3000, but early stopping determined the effective model complexity.

Cross-Validation Strategy. Model validation was conducted using five randomly selected folds. For each fold, training proceeded up to 3000 iterations with early stopping, and the iteration achieving the lowest validation RMSE was recorded. The final model was then trained on the full dataset using the *median* of the best iterations across folds, ensuring a model complexity representative of the cross-validated optimum.

Cross-Validation Results. The CatBoost model achieved excellent predictive performance:

$$\text{CV } R^2 = 0.995 \pm 0.000, \qquad \text{CV RMSE} = 0.945, \qquad \text{CV RMPE} = 0.054.$$

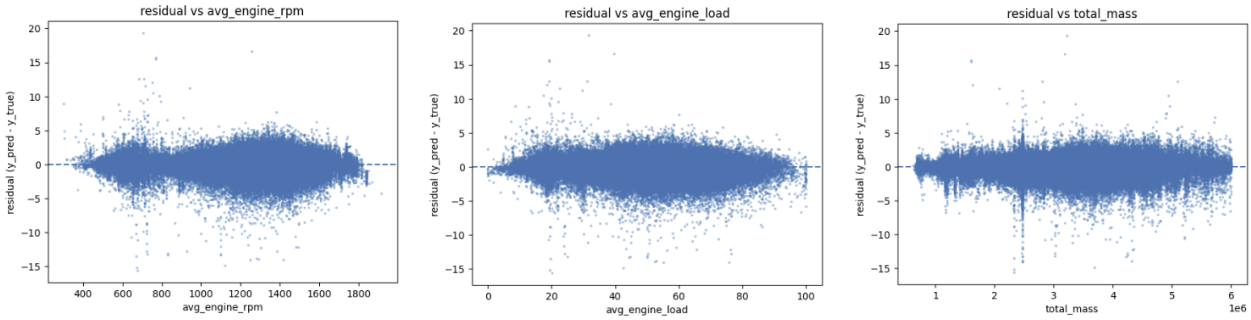
Percentage-based error metrics further confirm high accuracy:

$$\text{Median APE (P50)} = 0.026, \qquad \text{95th percentile APE} = 0.095,$$

with a standard deviation of absolute percentage errors of 0.040. The average best iteration across folds was 2999, with a small validation–training gap of 0.0383, indicating excellent generalization and minimal overfitting.

Residual Analysis. Residuals were inspected to assess model stability across key operational variables. Plots of residuals against `avg_engine_rpm`, `avg_engine_load`, and `total_mass` all show a uniform, structureless cloud across their respective ranges, confirming that the model does not systematically under- or over-predict in specific operating regimes.

Figures 16a-16c display these residual patterns side-by-side.



(a) Residuals vs. RPM (b) Residuals vs. Load (c) Residuals vs. Mass

Figure 16: Residual stability across key operational predictors.

Residuals were also evaluated across environmental categories. Mean and standard deviation values were:

$$\begin{aligned} \text{DOWNSTREAM:} &\quad \mu = -0.0089,\ \sigma = 0.487, \\ \text{TIDAL:} &\quad \mu = 0.0307,\ \sigma = 1.533, \\ \text{UPSTREAM:} &\quad \mu = -0.0932,\ \sigma = 1.311. \end{aligned}$$

These results indicate symmetric and low-mean residuals in all three environments, with higher variance in TIDAL and UPSTREAM regions corresponding to greater intrinsic variability in operating conditions. A dedicated plot of these distributions is provided in Figure 17.

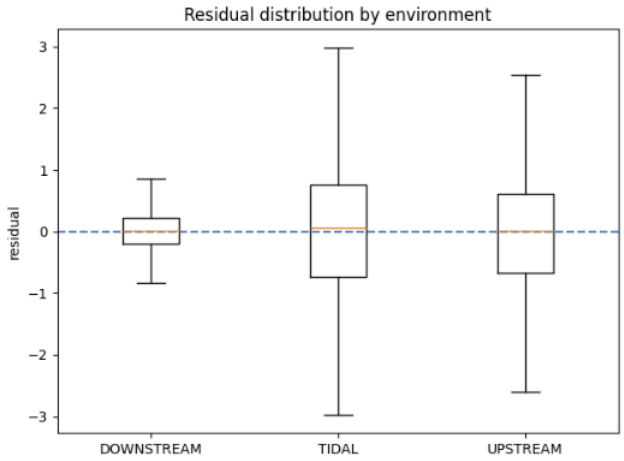


Figure 17: Residual distribution across environmental categories.

Finally, residuals were analyzed per vessel (`device_id`). The resulting plot (Figure 18) shows residuals with uniformly low means and no systematic patterns across ships, confirming that the model generalizes well across the heterogeneous fleet.

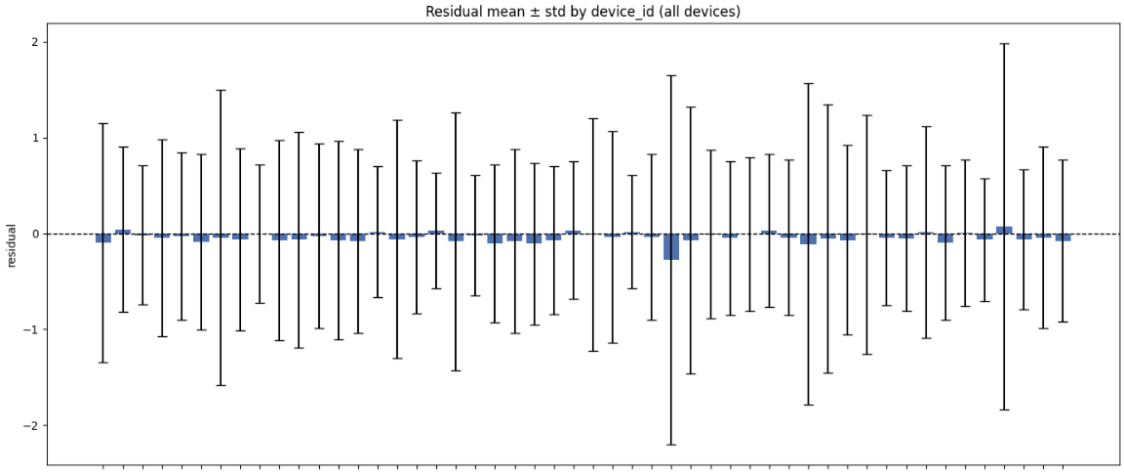


Figure 18: Residual distribution across vessels.

SHAP Feature Importance Analysis. To further interpret the CatBoost model and verify that it learned physically meaningful relationships, a SHAP (SHapley Additive exPlanations) analysis was performed.
The resulting summary plot (Figure 19) confirms that the model relies on the expected dominant drivers of fuel consumption. `avg_engine_rpm` is by far the strongest predictor, followed by `environment`, `edge`, and `total_mass`, in agreement with hydrodynamic intuition. The remaining structural variables (`ship_type`, `device_id`, dimensions, and build year) contribute meaningfully but with lower magnitude, indicating that the model captures vessel-specific behaviour without overfitting to identity effects.

Importantly, `ala_on` appears near the bottom of the ranking with a very small SHAP range, consistent with the small causal effect estimated via g-computation.

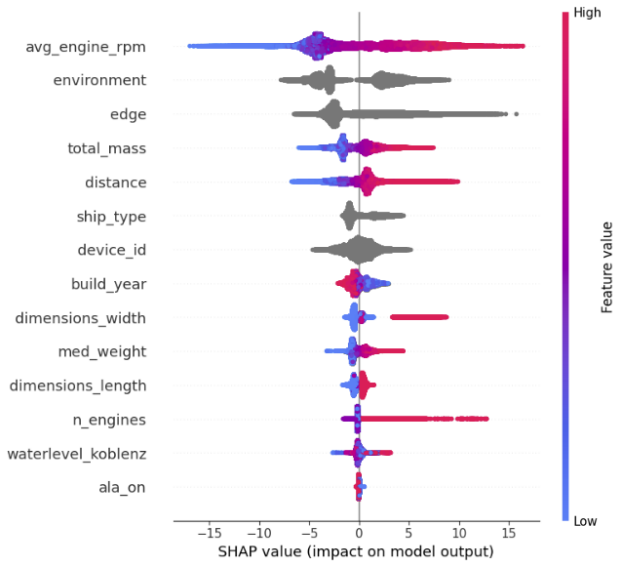


Figure 19: SHAP summary plot for the CatBoost prediction model. Higher absolute SHAP values indicate stronger influence on predicted fuel consumption.

Overall, the CatBoost model demonstrates excellent predictive accuracy and stability across operating regimes, environmental conditions, and vessel types, making it well-suited for pre-departure fuel consumption estimation.

Emission Prediction Performance. To assess not only fuel prediction accuracy but also its implications for environmental forecasting, the predicted fuel consumption values were converted into WTP CO_2e emissions using the GLEC diesel factor ($3240\,\mathrm{g\,CO_2e\,L^{-1}}$) and compared against measured emissions derived from recorded fuel consumption.
The resulting mean absolute percentage error (MAPE) of the CatBoost-based emission estimates is exceptionally low:

$$\mathrm{MAPE}_{\mathrm{predicted}} = 0.37\%.$$

This indicates that the decision-tree model reproduces traversal-level emissions with near-negligible average bias.

In contrast, the GLEC baseline emissions (computed using the vessel-type intensity classes from [16])exhibit a markedly poorer fit to real operational data:

$$\mathrm{MAPE}_{\mathrm{GLEC}} = 40.15\%.$$

This large deviation reflects the fact that GLEC intensity factors are designed as broad, population-level emission averages, intended for harmonised reporting rather than fine-grained vessel-specific estimation.

These findings are visually evident in Figure 20: the CatBoost emission errors (blue bars) are so small that they are nearly indistinguishable from the horizontal axis across all vessels, while the GLEC-based errors (orange bars) display substantial positive and negative deviations, often exceeding ±50% and in several cases surpassing 100%.
The contrast highlights the fundamental limitation of using standardized intensity factors for operational prediction: they cannot account for the heterogeneity in vessel design, loading conditions, propulsion characteristics, and hydrodynamic behaviour that strongly influence real emissions on a per-traversal basis.

Overall, the results demonstrate that the CatBoost model not only provides highly accurate fuel consumption predictions but also produces traversal-level emission estimates that are orders of magnitude more reliable than those obtained from the GLEC Framework.
This strongly supports the use of data-driven, vessel-specific models for operational and pre-departure emission forecasting, while GLEC remains more appropriate for aggregated, fleet-level accounting and reporting.

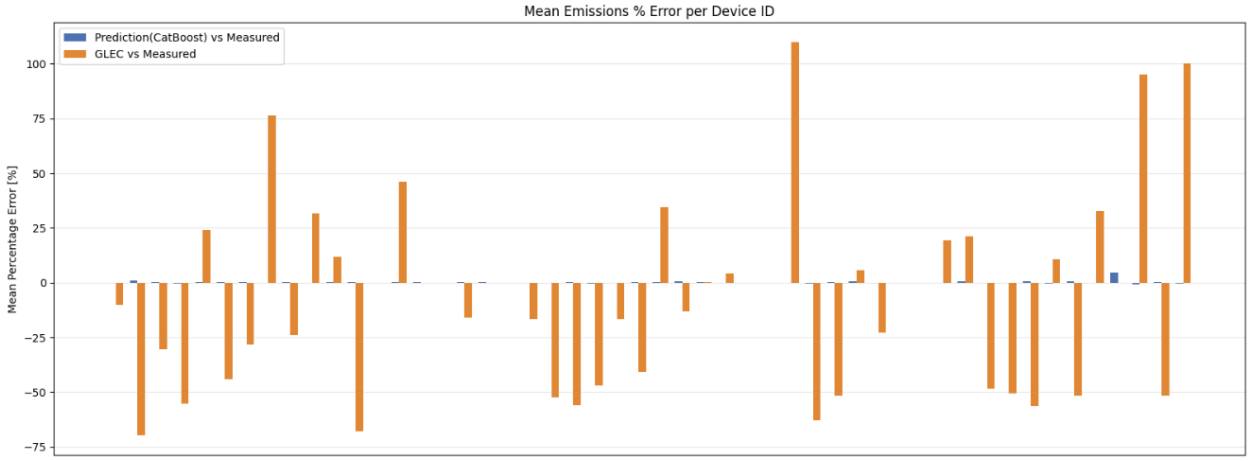


Figure 20: percentage error in emission prediction using the CatBoost prediction model or the GLEC framework

8.2.5 DBPNN model

A Deep Back-Propagation Neural Network (DBPNN) was implemented as a second predictive model for pre-departure fuel consumption estimation.

The model structure and training approach follow the methodology introduced in [21], where fully connected feed-forward neural networks are shown to be effective for nonlinear fuel–speed relationship modeling in marine transportation. DBPNNs approximate the fuel consumption function by learning a high-dimensional, nonlinear mapping from vessel characteristics, loading condition, and environmental inputs to fuel consumption. Unlike tree-based models, DBPNNs rely on stacked linear transformations and nonlinear activations, enabling smooth function approximation and control over model capacity through the number of layers and hidden units.

The same input features used in the CatBoost model were retained.
Categorical variables were one-hot encoded, numerical features were scaled to $[0, 1]$ using a MinMax transformation, and the target was similarly scaled before training.
The loss function was mean absolute error (MAE), as suggested in [21], and optimization was performed via Adam with a learning rate of 10^{-3}.
Training was conducted for at most 150 epochs, with early stopping (patience 10) based on validation MAE.

Architectures Tested. Following the structure of [21], several network depths and widths were evaluated to identify the best-performing configuration.
The following network architectures were tested:

- One hidden layer with 81 neurons ([81])
- One hidden layer with 128 neurons ([128])
- Two hidden layers with 81 and 8 neurons ([81, 8])
- Two hidden layers with 128 and 64 neurons ([128, 64])
- Three hidden layers with 128, 64, and 16 neurons ([128, 64, 16])

Table 9 summarises the validation performance across the tested architectures.
The two-layer network with 128 and 64 units achieved the best overall performance and was selected as the final model.

Final Model Results. Training the selected architecture yielded the following performance on the full training and test partitions (after inverse-scaling the predictions):

$$\begin{aligned} \text{Train RMSE} &= 0.9222, \quad \text{Train MAE} = 0.4754, \quad R^2_{\text{train}} = 0.99457, \\ \text{Test RMSE} &= 1.0180, \quad \text{Test MAE} = 0.5598, \quad R^2_{\text{test}} = 0.9934. \end{aligned}$$

Table 9: DBPNN architecture comparison. Metrics computed on the held-out test set.

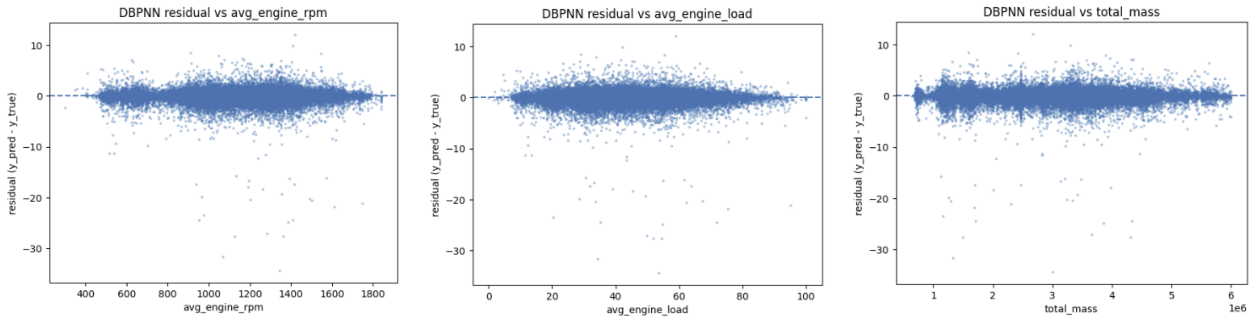
Architecture	RMSE	MAE	R^2	RMPE	SD Abs.%	P50 / P95
`[81]`	1.0946	0.5973	0.9923	0.0621	0.0496	0.0246 / 0.1071
`[128, 64]`	1.0180	0.5598	0.9934	0.0536	0.0419	0.0232 / 0.0929
`[128]`	1.0888	0.5970	0.9924	0.0622	0.0496	0.0248 / 0.1073
`[128, 64, 16]`	1.0265	0.5625	0.9932	0.0546	0.0430	0.0233 / 0.0942
`[81, 8]`	1.2027	0.6397	0.9907	0.0656	0.0521	0.0266 / 0.1122

Percentage-based errors further confirm high accuracy on the test set:

RMPE = 5.48%, SD Abs.% = 4.29%, Median APE (P50) = 2.35%, P95 APE = 9.58%.

These results are comparable in magnitude to the CatBoost model, though the latter retains a slight advantage in both RMSE and percentage error stability.

Residual Analysis. As with the CatBoost model, DBPNN residuals were examined to evaluate model stability across key operational features and to verify that the network does not systematically under- or over-predict within specific regions of the input space.
The residuals exhibit a narrow, structureless cloud across the full ranges of `avg_engine_rpm`, `avg_engine_load`, and `total_mass`.
This indicates that the neural network provides consistent predictions throughout the operating domain, without local bias.
The corresponding visualisations are included in Figures 21a-21c.



(a) Residuals vs. RPM (b) Residuals vs. Load (c) Residuals vs. Mass

Figure 21: DBPNN residual stability across key operational predictors.

Residuals were also aggregated by river environment (downstream, tidal, upstream). The mean residuals remain close to zero in all cases, with comparable standard deviations around unity. Table 10 reports the summary statistics.

Table 10: DBPNN residuals by environment.

Environment	Mean residual	Std. residual	n
DOWNSTREAM	−0.0458	1.0577	25600
TIDAL	0.0106	0.9745	2280
UPSTREAM	−0.0309	1.0478	29523

These values show no systematic drift across environments and confirm that the model generalises well across hydrodynamic regimes with differing resistances and velocity profiles. A dedicated visualisation of the residual distributions across environments is shown in Figure 22.

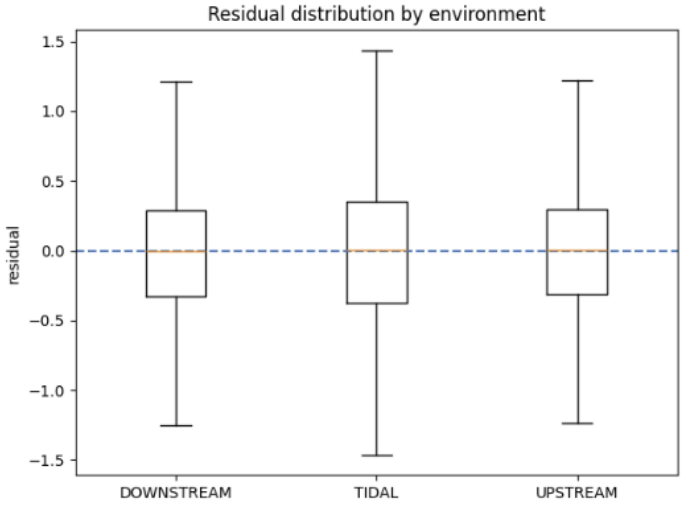


Figure 22: Residual distribution across environmental categories for the DBPNN model.

Finally, residuals were examined per vessel (`device_id`).
As with the CatBoost model, the DBPNN residuals display uniformly low means and consistent variance across ships, demonstrating that the model captures vessel-specific performance characteristics without overfitting to particular devices.
Figure 23 shows the distribution of residuals across vessels.

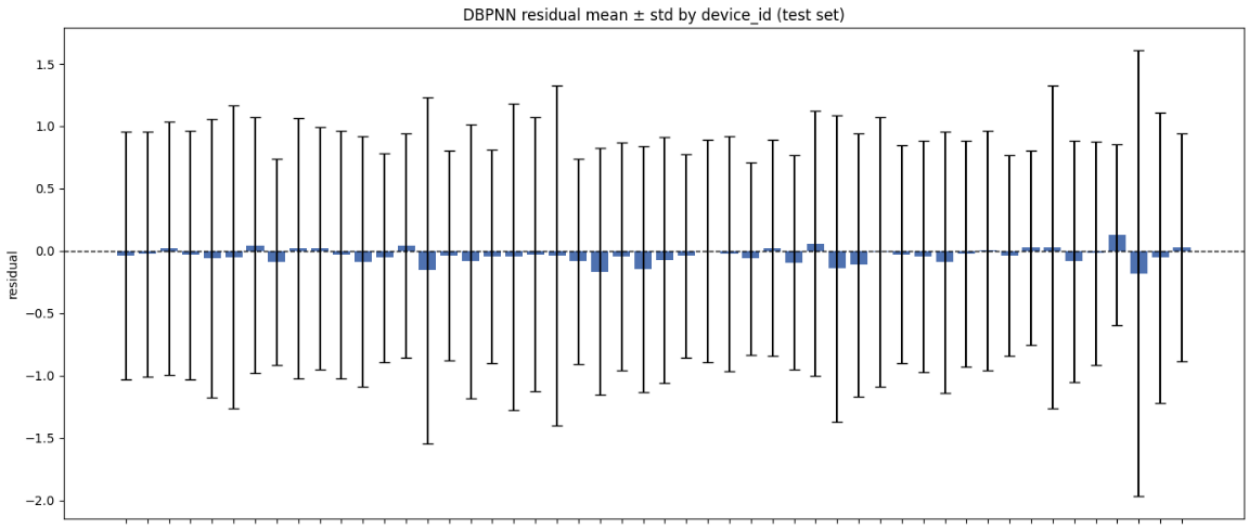


Figure 23: Residual distribution across vessels for the DBPNN model.

Emission Prediction Performance. To evaluate environmental prediction accuracy, the DBPNN fuel predictions were converted into WTP CO_2e emissions using the same GLEC diesel factor ($3240\,\mathrm{g\,CO_2e\,L^{-1}}$).
Emission errors were computed against the scaled test-set ground truth used during DBPNN evaluation, while GLEC baseline emissions were compared against the true measured fuel consumption.

The DBPNN achieves an exceptionally low emission MAPE:

$$\mathrm{MAPE}_{\mathrm{predicted}} = 0.27\%.$$

In contrast, the GLEC framework exhibits a substantially larger deviation:

$$\mathrm{MAPE}_{\mathrm{GLEC}} = 40.15\%,$$

As shown in Figure 24, the DBPNN emission errors (blue bars) are nearly indistinguishable from zero across vessels, whereas the GLEC-based errors (orange bars) display wide variability, often exceeding ±50% and in some cases surpassing 100%.
This again demonstrates that vessel-specific, data-driven models substantially outperform generic intensity-based methods for operational emission estimation.

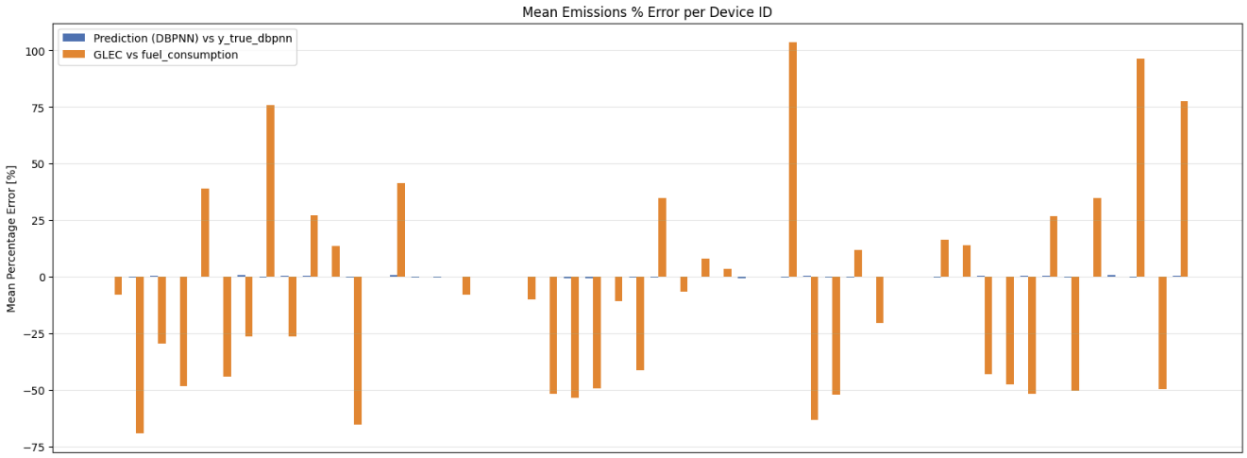


Figure 24: Percentage error in emission prediction using the DBPNN model and the GLEC framework.

Overall, the DBPNN provides a flexible and highly accurate functional approximation for fuel consumption and emissions, with performance close to that of the CatBoost model. Its smooth nonlinear structure makes it an appealing alternative in scenarios where differentiability or continuous sensitivity analysis is required (e.g., trajectory or RPM optimization), while tree-based models remain advantageous when interpretability or feature attribution is important.

9 Results

9.1 Causal Inference Results

G-computation is a model-based causal inference method that estimates the expected outcome under counterfactual conditions by using a predictive model to simulate "what would have happened" under alternative treatment values. In this study, it is used to estimate the causal effect of the Autonomous Lane Assist (ALA) on fuel consumption. The complete methodological description and model specification are presented in Section 8.1.5.

The approach consists of training a flexible predictive model for total fuel consumption, and then simulating counterfactual outcomes by toggling the ALA variable between active and inactive states for every observed traversal.
The difference between these two predictions represents the estimated causal effect of ALA for that specific traversal.

Aggregation of Counterfactual Results Predictions were computed for each traversal (row of the dataset) under both counterfactual scenarios (`ALA_ON` and `ALA_OFF`). The relative fuel savings for each traversal were then computed as:

$$\text{Savings}_i = \frac{\widehat{Y}_i^{\text{off}} - \widehat{Y}_i^{\text{on}}}{\widehat{Y}_i^{\text{off}}} \times 100.$$

Three KPIs were used to summarise these traversal-level savings, as described in Section 8.1.5: (1) the *global ATE*, averaging savings across all traversals; (2) the *edge-weighted ATE*, averaging savings within and then across river sections; (3) the *ship-weighted ATE*, averaging savings within and then across vessels.

The statistical significance of the results was evaluated using influence–function standard errors, Wald confidence intervals and hypothesis tests, as well as nonparametric bootstrap procedures. Further methodological details are provided in Section 6.2.

9.1.1 Global ATE

The traversal-weighted global ATE, the most direct estimate of the average fuel savings across the empirical distribution of observed trips, is:

$$\widehat{\text{ATE}}_{\text{global}} = 0.68\% \ \pm \ 1.52\%.$$

To assess statistical precision, inference was performed on the mean of the per-traversal savings:

- plug-in estimate: 0.6799%,

- influence-function standard error: 0.00262,
- 95% Wald confidence interval: (0.675%, 0.685%),
- two-sided Wald p-value: < 0.001.

Bootstrap inference with $B = 1000$ replications yielded nearly identical uncertainty estimates:

- bootstrap standard error: 0.00263,
- 95% percentile confidence interval: (0.675%, 0.685%).

Because traversals belonging to the same vessel are not independent, a cluster-robust variance estimator (clustering on `device_id`) was also computed. This produced a more conservative standard error:

$$\mathrm{SE}_{\mathrm{cluster}} = 0.301\%,$$

and a wild-cluster bootstrap p-value of < 0.001, confirming that the positive effect remains statistically significant even under strong intra-ship dependence.

9.1.2 Edge-weighted ATE

To avoid disproportionate influence from frequently-traversed segments, the effect was also aggregated at the river-section level. The resulting edge-weighted ATE is:

$$\widehat{\mathrm{ATE}}_{\mathrm{edge}} = 0.66\% \ \pm \ 0.31\%.$$

The distribution of edge-level effects (Figure 25) shows substantial heterogeneity across the Rhine network, but with consistently positive savings in all major environments:

Environment	Mean ATE [%]	Std. dev.	n_{edges}
UPSTREAM	0.577	0.387	213
DOWNSTREAM	0.702	0.182	213
TIDAL	1.014	0.175	20

Edges in tidal regions exhibit the largest average savings, whereas upstream sections show more variability, likely reflecting stronger interactions between steering behaviour, flow velocity, and vessel loading.

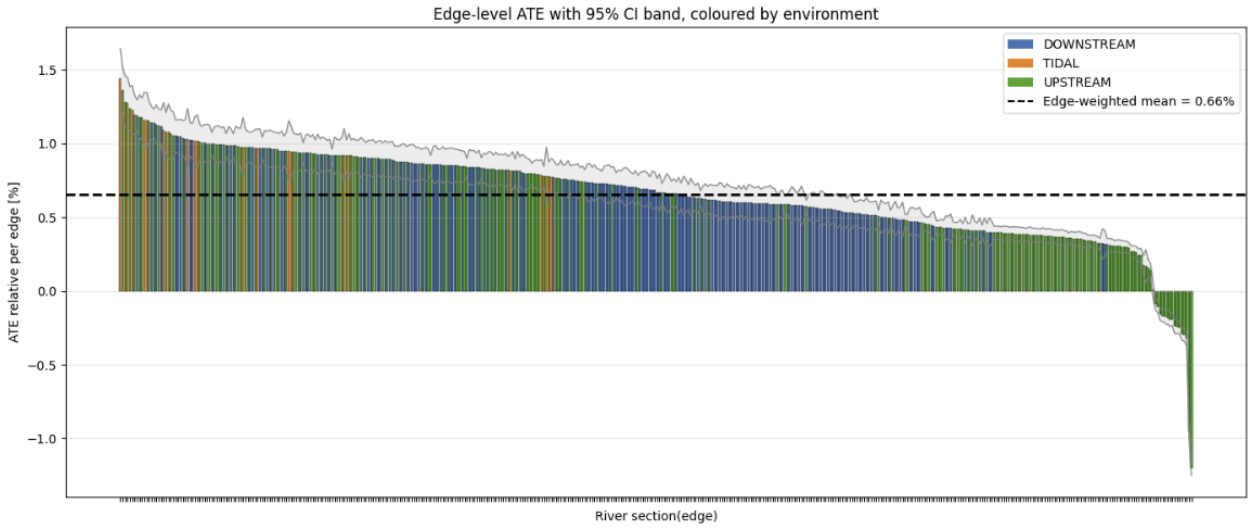


Figure 25: Fuel Consumption Saving Percentage per unique edge

9.1.3 Ship-weighted ATE

The device-/vessel-weighted ATE treats each vessel equally, preventing ships with many recordings from dominating the result. The estimate is:

$$\widehat{\mathrm{ATE}}_{\mathrm{ship}} = 0.77\% \ \pm \ 0.95\%.$$

The distribution of vessel-level effects (Figure 26) also indicates clear heterogeneity across ship types. Passenger and container vessels tend to show higher average savings, while certain cargo or tanker vessels exhibit effects closer to zero or mildly negative. This pattern is consistent with differences in maneuverability, hydrodynamic response, and steering behavior between class of vessels.

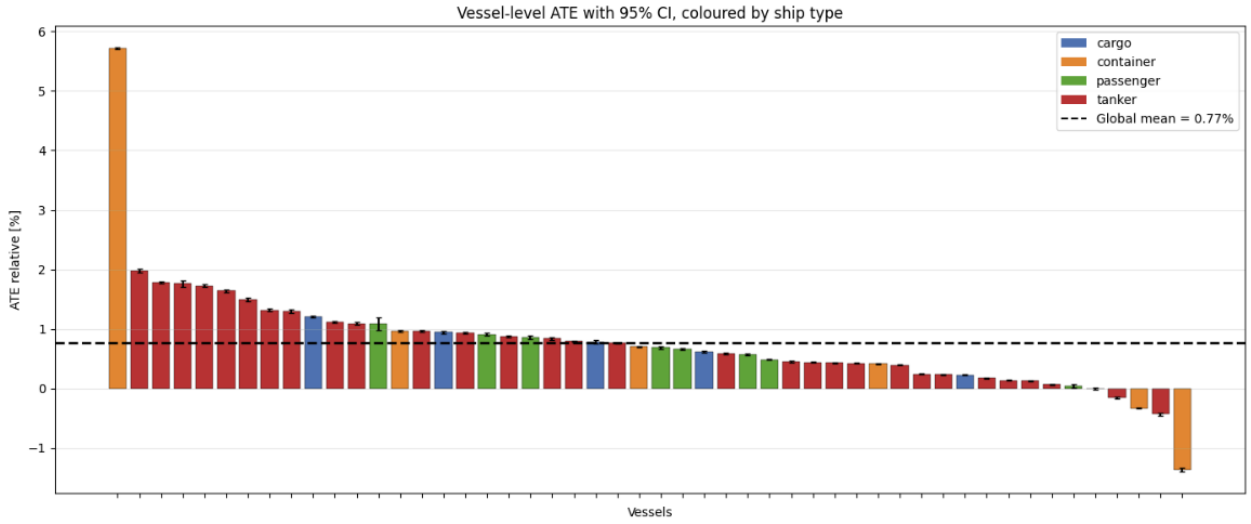


Figure 26: Fuel Consumption Saving Percentage per unique vessel

9.1.4 Summary of Estimated ALA Effect

Across all aggregation perspectives, ALA consistently reduces fuel consumption:

$$0.66\% \;\leq\; \widehat{\mathrm{ATE}} \;\leq\; 0.77\%.$$

The inference procedures (influence-function SEs, bootstrap SEs, cluster-robust SEs, and wild-cluster bootstrap tests) all confirm high statistical significance of the estimated savings. Although the absolute magnitude of the savings is modest, the effect is systematic, robust across vessels and river sections, and strongly supported by the data-driven counterfactual modelling framework.

9.2 Conditional Average Treatment Effect (CATE)

Beyond estimating a single global Average Treatment Effect, it is often informative to investigate how the effect of the treatment varies across different operational or environmental conditions. The **Conditional Average Treatment Effect (CATE)** quantifies the expected treatment effect within a specific region of the feature space:

$$\mathrm{CATE}(x \in \mathcal{R}) = E[Y(0) - Y(1) \mid X \in \mathcal{R}].$$

In this context, CATE allows us to understand *under which conditions ST-ALA achieves the largest fuel savings*, highlighting interaction patterns between vessel state, hydrodynamics, and environmental regime.

Construction of feature-space regions. To compute the CATE across distinct operating regimes, several continuous features were discretized into interpretable categories. The variables `avg_engine_rpm`, `avg_engine_load`, `med_depth`, and `total_mass` were binned into three percentile-based intervals (0-33rd, 33rd-66th, 66th-100th), producing low/mid/high operational strata for each variable. In addition, the categorical features `environment` (upstream, downstream, tidal) and `ship_type` were included.

This yielded a feature space partitioned into:

$$3^4 \times 3 \times 4 = 324\,\text{regions},$$

combining four continuous variables (three bins each), three environmental categories, and four ship types.

Estimation procedure. For each region, the treatment effect was obtained by averaging the relative savings:

$$\widehat{\tau}_{\mathcal{R}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \text{savings\_rel}_i,$$

where `savings_rel` is the per-traversal percentage savings estimated through g-computation.

The regions were then sorted by their estimated treatment effect to identify the conditions under which ALA provides the largest fuel reduction.

Dominance by individual ships. Because the dataset is large, all regions contain sufficient traversals; however, initial results showed that several top-ranked regions were overwhelmingly dominated by a single vessel, which itself exhibits unusually high savings. In some regions, this single device accounted for more than 99% of traversals, making the raw CATE uninformative: it merely reproduced the within-ship mean effect rather than isolating structural conditions.

Adjusted CATE to remove ship-level dominance. To avoid this degeneracy, a corrected CATE was computed by *centering each ship's savings around its own mean.* This adjustment removes constant ship-specific effects (e.g., hull geometry, engine behavior, captain style), allowing the CATE to capture treatment effect differences attributable to operational and environmental conditions rather than idiosyncrasies of one vessel.

Findings. After ship-centering, the regions with the highest additional savings revealed clear pattern:

- low engine RPM,
- low engine load,
- low total mass,
- downstream environment.

These top three regions contain 24, 5, and 12 unique ships respectively, with 3173, 748, and 2793 traversals,ample sample sizes for stable inference. The corresponding adjusted CATEs indicate additional savings of:

$$0.73\%, \quad 0.66\%, \quad 0.60\%$$

above each ship's baseline average savings.

A more extensive table is present in the Appendix B.

Interpretation. These results suggest that ST-ALA yields its greatest relative benefits under low-power, low-resistance operating regimes, particularly when traveling downstream and carrying lighter loads. Under such conditions, smoother steering and reduced rudder activity translate more effectively into marginal fuel savings, while high-load or high-resistance conditions appear to limit the relative advantage of assisted steering.

This CATE analysis therefore provides a structured and data-driven understanding of *where* ALA performs best and offers a meaningful complement to the global ATE used in the main evaluation.

9.3 Prediction Results

This section summarises the comparative performance of the three predictive models: the linear regression baseline, the CatBoost decision-tree model, and the DBPNN.
All metrics are reported on held-out test data and refer to the prediction of traversal-level fuel consumption in litres, unless otherwise stated.

Overall predictive accuracy. The linear regression model provides a reasonable but clearly limited baseline. On the test set it achieves

$$\text{RMSE} = 3.999, \qquad R^2 = 0.9032,$$

with percentage-based errors

$$\text{RMPE} = 44.7\%, \qquad \text{Median APE} = 14.3\%, \qquad \text{P95 APE} = 90.2\%.$$

While the median error is moderate, the very large 95th-percentile error indicates that the linear specification fails badly in more challenging operating regimes. Together with the counterintuitive coefficients discussed earlier, these results confirm that a single global linear model is not an adequate functional form for this task.

Both nonlinear models substantially improve on the linear baseline. The CatBoost decision-tree model achieves cross-validated performance of

$$\text{CV } R^2 = 0.995 \pm 0.000, \qquad \text{CV RMSE} = 0.945, \qquad \text{CV RMPE} = 5.4\%.$$

On the test data, its percentage error profile is very tight, with

$$\text{Median APE (P50)} = 2.6\%, \qquad \text{P95 APE} = 9.5\%, \qquad \text{SD Abs.\%} = 4.0\%.$$

Residual plots versus `avg_engine_rpm`, `avg_engine_load`, and `total_mass` (Figures 16a-16c) show a narrow, structureless cloud across the full range of each predictor, with no clear patterns by environment or vessel (Figures 17 and 18). This indicates excellent stability and generalisation across operating conditions and ships.

The DBPNN achieves similarly strong performance. For the selected $[128, 64]$ architecture, the final model yields

$$\text{Train RMSE} = 0.9222, \quad \text{Train MAE} = 0.4754, \quad R^2_{\text{train}} = 0.9946,$$

$$\text{Test RMSE} = 1.0180, \quad \text{Test MAE} = 0.5598, \quad R^2_{\text{test}} = 0.9934,$$

with percentage errors

$$\text{RMPE} = 5.48\%, \qquad \text{SD Abs.\%} = 4.29\%, \qquad \text{Median APE (P50)} = 2.35\%, \qquad \text{P95 APE} = 9.58\%.$$

Residuals again form a narrow cloud across RPM, load, and mass (Figures 21a-21c), and show low, symmetric means across environments and vessels (Table 10 and Figure 23). Overall, the DBPNN is only marginally less accurate than CatBoost, delivering a very similar error profile.

Table 11 provides a compact comparison of the three models.

Table 11: Comparison of predictive performance across models (test data).

Model	**RMSE**	R^2	**Median APE**	**P95 APE**
Linear regression	3.999	0.903	14.3%	90.2%
CatBoost	0.945	0.995	2.6%	9.5%
DBPNN	1.018	0.993	2.35%	9.58%

Emission estimation and comparison to GLEC. Converting fuel predictions into WTP CO_2e emissions using the GLEC diesel factor enables a direct comparison between model-based emissions, measured emissions, and GLEC baseline emissions. For the CatBoost model, the mean absolute percentage error between predicted and measured emissions is

$$\mathrm{MAPE}^{\mathrm{CatBoost}}_{\mathrm{predicted}} = 0.37\%,$$

while the GLEC baseline, computed from vessel-class intensity factors [16], exhibits

$$\mathrm{MAPE}_{\mathrm{GLEC}} \approx 40\%.$$

The DBPNN shows essentially identical behaviour, with

$$\mathrm{MAPE}^{\mathrm{DBPNN}}_{\mathrm{predicted}} = 0.27\%, \qquad \mathrm{MAPE}_{\mathrm{GLEC}} \approx 38.8\%.$$

In both cases, the model-based emission errors are so small that they are barely visible in the per-vessel error plots (Figures 20 and 24), whereas the GLEC-based errors display large positive and negative deviations, often above $\pm 50\%$.
This reflects the fundamental difference in purpose: GLEC intensities are designed as broad, fleet-average factors for harmonised reporting, not as vessel-specific operational predictors.

Summary The results demonstrate a clear hierarchy of performance.
The linear regression model provides a useful, interpretable baseline but fails to capture the nonlinear dependencies governing fuel consumption, leading to substantial errors in difficult regimes and physically implausible coefficients.
In contrast, both CatBoost and DBPNN deliver high-precision fuel and emission predictions, with $R^2 \geq 0.993$ and median absolute percentage errors around 2–3%.
CatBoost achieves the best overall accuracy and slightly more stable residuals, while the DBPNN offers comparable performance with a smooth, differentiable structure that can be exploited for optimisation.

Crucially, both data-driven models approximate traversal-level emissions orders of magnitude more accurately than the GLEC baseline, underscoring the value of vessel-specific predictive models for operational planning and environmental assessment.

10 Conclusion

10.1 Answers to Research Questions

This section synthesises the main empirical findings of the thesis and answers the research questions introduced in Section 1.

Main question *How much fuel is saved when an inland vessel sails using the Autonomous Lane Assist system compared to conventional steering methods?*

The causal analysis based on G-computation yields a small but statistically significant average reduction in fuel consumption when ST-ALA is activated.
Using the traversal-weighted estimator, the global average treatment effect is

$$\widehat{\mathrm{ATE}}_{\mathrm{global}} = 0.68\% \pm 1.52\%,$$

expressed as relative fuel savings per traversal.
However, traversals belonging to the same ship are strongly correlated, and vessels differ systematically in both design and operating style.
For this reason, the *ship-weighted* estimator—which assigns equal weight to each vessel rather than to each traversal—is arguably the most appropriate summary for fleet-level conclusions.
At ship level, the estimated effect is

$$\widehat{\mathrm{ATE}}_{\mathrm{ship}} = 0.77\% \pm 0.95\%.$$

Figure 26 shows the estimated savings for each vessel. There is substantial heterogeneity across ships: some appear to benefit more, some less, and a few vessels even show small negative point estimates. No ship type exhibits categorically better performance than the others; savings are driven more by vessel-specific characteristics and operating patterns than by coarse vessel class.

Despite the small magnitude of the average effect, the estimates are statistically significant under influence-function based standard errors and bootstrap confidence intervals.
The methodology validation experiments,using shuffled outcomes (no effect) and synthetic interventions (known effect),demonstrate that the G-computation pipeline is able to recover both null and non-null effects when they exist.
Taken together, the results support the conclusion that ST-ALA does deliver a real, albeit modest, reduction in fuel consumption at the traversal level.

Sub-question 1 *Which variables most significantly affect fuel consumption in inland shipping?*

Across both the causal and predictive modelling components, the variables that most strongly influence fuel consumption are consistent with physical intuition:

- **Average engine RPM** (`avg_engine_rpm`) is the dominant driver of fuel use. Higher RPM directly increases fuel injection and, therefore, consumption.

- **Sailing direction / environment** (`environment`) strongly affects resistance. Upstream sailing requires markedly more power and fuel than downstream, whereas tidal segments sit between the two.
- **River segment** (`edge`) captures local bathymetry, bends, speed limits, and traffic patterns. Edge-level effects are large and persistent.
- **Vessel mass** (`total_mass`, including cargo and ballast water) is a major determinant of hydrodynamic resistance and thus fuel consumption.

Secondary contributors include loading proxies (`med_weight`, `med_draft`), water level at Koblenz, vessel dimensions, and ship type.
SHAP analyses of both the G-computation outcome model and the predictive CatBoost model confirm that these variables jointly explain most of the variance in fuel consumption, while ST-ALA itself appears as a relatively small contributor.

Sub-question 2 *How can historical navigation and fuel data be structured to compare between ST-ALA and non-ST-ALA trips?*

To enable a valid causal comparison between ST-ALA and conventional steering, the historical data must be organised at the *traversal* (edge) level and enriched with all relevant pre-treatment covariates. In this thesis, each row corresponds to a single ship traversal of a specific river segment and contains:

- vessel characteristics (dimensions, number of engines, ship type, build year, device identifier);
- loading and hydrostatic information (total mass, median weight, draft and trim);
- environmental and route descriptors (environment category, edge identifier, distance, water level);
- operational settings and automation status (average engine RPM, engine load, ST-ALA on/off).

The dataset was then filtered and engineered to satisfy, as far as possible, the three key identifiability conditions for G-computation and IPW: *consistency*, *exchangeability*, and *positivity*.
Consistency is enforced by defining well-measured, non-overlapping traversals.
Exchangeability is addressed by conditioning on a rich set of covariates that jointly capture route, vessel, loading, and environment, so that within strata of these variables, ST-ALA activation can be regarded as as-good-as random. Positivity is checked by verifying that, for relevant combinations of confounders (e.g. vessel, environment, RPM ranges), both ST-ALA and non-ST-ALA traversals are present with non-zero probability.

In practice, structuring the data in this way,one row per traversal with comprehensive pre-treatment covariates and careful enforcement of positivity filters,is what makes it possible to obtain credible, model-based comparisons between ST-ALA and non-ST-ALA operations.

Sub-question 3 *Under what conditions does the ST-ALA cause the most fuel savings?*

The Conditional Average Treatment Effect analysis presented in Section 9.2 shows that ST-ALA is most effective under **low-power, low-resistance operating conditions**. Specifically, after removing vessel-specific baseline differences, the regions with the highest additional savings are characterised by low engine RPM, low engine load, low total mass, and a downstream sailing direction. The three highest-ranked operational regimes yield adjusted CATEs of

$$0.73\%, \quad 0.66\%, \quad 0.60\%$$

above each vessel's own average savings. At the environment level, tidal sections exhibit the largest mean edge-weighted savings (1.01%), followed by downstream (0.70%) and upstream segments (0.58%).

The pattern is physically intuitive: in low-resistance regimes the propulsive power demand is modest, so the marginal benefit of smoother rudder control (reducing hydrodynamic drag losses) represents a larger fraction of the total fuel budget. Under high-load or high-resistance conditions, engine output dominates and the relative contribution of steering-induced losses is smaller. These findings are consistent with the ship-level heterogeneity observed in the ATE results: vessels that habitually operate at lower engine loads tend to show disproportionately higher savings from ST-ALA activation.

Sub-question 4 *What is the most appropriate modeling approach to predict fuel consumption per river segment?*

The predictive modelling results in Section 8.2 show a clear hierarchy of model performance.
The linear regression baseline, while interpretable, achieves only

$$\text{RMSE} = 3.999, \quad R^2 = 0.903, \quad \text{Median APE} \approx 14\%.$$

Its coefficient structure contains several counterintuitive signs, and its error profile deteriorates sharply in challenging regimes, confirming that a single global linear model cannot adequately capture the nonlinear hydrodynamics of inland navigation.

In contrast, both nonlinear models perform extremely well. As summarised in Table 11, the CatBoost decision-tree model attains

$$R^2 \approx 0.995, \quad \text{RMSE} \approx 0.95, \quad \text{Median APE} \approx 2.6\%,$$

while the DBPNN yields

$$R^2 \approx 0.993, \quad \text{RMSE} \approx 1.02, \quad \text{Median APE} \approx 2.35\%.$$

Both have 95th-percentile absolute percentage errors below 10%, and residuals are well-behaved across RPM, load, mass, environment, and vessels.

Given these results, the most appropriate approach for per-segment fuel prediction is a flexible, data-driven nonlinear model.
CatBoost offers slightly better accuracy and excellent robustness, with the additional advantages of handling categorical variables natively and providing interpretable feature-attribution diagnostics.
The DBPNN achieves very similar performance while offering a smooth, differentiable mapping that may be advantageous for gradient-based optimisation of RPM or routing. For operational deployment, either of these two nonlinear models is vastly preferable to linear regression.

However, neither model is immediately ready for deployment in its current form. Several prerequisites must be satisfied before operational use: the models were trained on a specific fleet of Shipping Technology clients on the Rhine between Dordrecht and Koblenz, so generalisation to other vessels, river systems, or operating conditions cannot be assumed without revalidation. Additionally, the models rely on pre-departure inputs (in particular, forecasted water level and cargo weight) whose accuracy in practice has not been evaluated here. Deployment would therefore require integration with reliable forecasting services, a live data pipeline for vessel characteristics, and a validation study on prospective voyages to confirm that training-time performance holds in production.

Sub-question 5 *How can the fuel consumption prediction models be used to predict CO_2 emissions?*

Fuel consumption predictions can be converted into Well-to-Propeller CO_2 emissions by multiplying by the GLEC diesel emission factor [16]:

$$\text{Emissions}_{\text{pred}}, \text{i} = \hat{F}_i \times 3240 \text{ g CO}_2\text{e L}^{-1}.$$

This allows a direct three-way comparison between (i) predicted emissions from the models, (ii) emissions computed from measured fuel consumption, and (iii) GLEC baseline emissions obtained from vessel-class intensity factors.

For both CatBoost and DBPNN, the mean absolute percentage error between predicted and measured emissions is below 0.4% at traversal level, whereas the GLEC baseline exhibits MAPEs close to 40%. The per-vessel error plots show that model-based emissions are almost indistinguishable from the measured values, while GLEC-based emissions display large positive and negative deviations.

Therefore, once a reliable fuel-consumption model is available, converting its output to CO_2e using a standard factor provides highly accurate traversal-level emission estimates.
These data-driven models approximate operational emissions orders of magnitude more accurately than the GLEC default intensities, making them far better suited for ship-specific planning, benchmarking, and evaluation of technologies such as ST-ALA, while GLEC remains appropriate primarily for high-level, harmonised reporting.

10.2 Other Findings and Takeaways

Beyond the causal and predictive results presented in the previous sections, several additional insights emerged from exploratory analyses.

1. Strong Vessel-Specific Effects: the Dominant Role of `device_id`. One of the most striking findings of this thesis is the magnitude of vessel-specific heterogeneity.
The feature `device_id`,which encodes the individual vessel, and implicitly its mechanical configuration, steering-control system, company procedures, and captain behaviour,emerges consistently as a dominant predictor in both the propensity score model and the causal effect estimation.

This result implies:

- the *probability* of ST-ALA being activated is highly vessel-dependent, even under identical hydrodynamic and loading conditions;
- the *magnitude* of ST-ALA's fuel-saving effect is also highly vessel-dependent.

This suggests that the benefit of ST-ALA is not uniform across the fleet.
It is mediated by vessel-specific factors, such as steering actuator dynamics, autopilot interfaces, mechanical imperfections, and habitual captain steering styles.
For industry, this indicates that ALA deployment and tuning may need to be vessel-specific rather than uniform across the fleet

2. The "Outlier Vessel": A Case Study in High Fuel Savings. Figure 26 identified one vessel achieving average fuel savings of approximately 5.7%, substantially above the fleet-wide average ATE.
To investigate whether identifiable operational characteristics explain this result, the mean values of several key features for this vessel were compared against the rest of the fleet using percentile-normalized values.

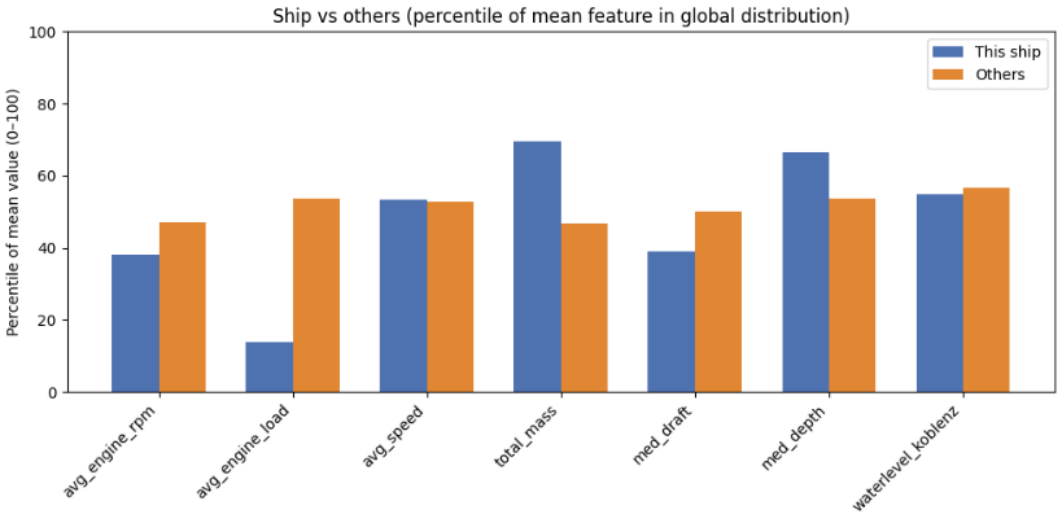


Figure 27: Percentile-normalized feature profile for the vessel with highest ALA savings, compared against the fleet average.

Two deviations stand out clearly:

- The vessel operates at a *substantially lower average engine load* than the fleet median.
- The vessel has a *higher total mass* relative to the rest of the fleet.

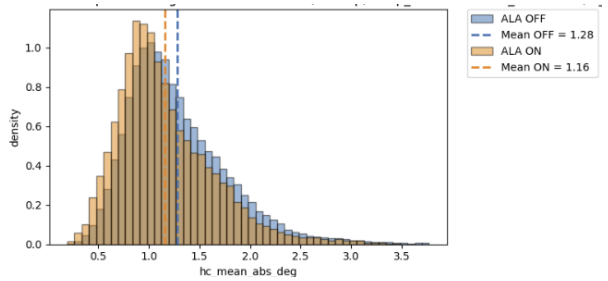
These findings are notable because the CATE analysis (Section 9.2) showed that ST-ALA achieves the highest savings in low-load operating regimes. Thus, the vessel's operational envelope aligns precisely with the region where ST-ALA is most effective.

Still, these factors alone do not fully explain the magnitude of the effect. Other vessel-specific mechanisms,such as control-system tuning, rudder response characteristics, or crew steering habits,may play a considerable role. This case study highlights the need for deeper investigation into vessel-by-vessel responsiveness to ALA.

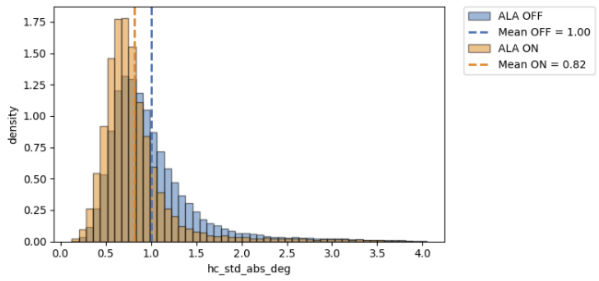
3. Influence of Internal Steering-Control Systems. ST-ALA does not directly actuate the rudder; instead, it outputs a desired steering rate (in degrees per minute), which is then translated into rudder commands by each vessel's own steering-control system.
Because ship control systems differ significantly in control-law design, feedback gains, and mechanical response, the same ST-ALA command may produce very different rudder dynamics across vessels. This may explain a large part of the heterogeneity in savings observed across vessels.

To explore whether ST-ALA leads to smoother trajectories in practice, the available traversal linestrings (sampled at 15-second intervals) were used to construct a simple proxy for rudder activity. For each traversal, the heading change between consecutive linestring segments was computed, from which both the mean absolute heading change and its standard deviation were derived.

Figure 28 compares the distributions of these metrics between ALA-ON and ALA-OFF observations(overlapped on top of each other). The results suggest that ALA-ON traversals exhibit, on average, slightly lower mean heading change and lower variability, indicating smoother trajectories and reduced rudder activity. However, the differences are small and must be interpreted with caution: 15-second sampling is far too coarse to capture high-frequency steering dynamics, and GPS error can be in the order of 5-15 meters, these metrics cannot be considered accurate depictions of true rudder usage by themselves.
Nevertheless, this exploratory analysis points toward a promising direction for future research based on higher-frequency data.



(a) Distribution of Mean absolute heading change with ST-ALA active or not



(b) Distribution of the Standard deviation of heading change with ST-ALA active or not

Figure 28: Distribution of heading-change metrics for ALA-ON and ALA-OFF traversals. These coarse indicators suggest smoother trajectories under ALA but are limited by the 15-second sampling resolution.

4. Behavioural and Operational Factors. Since captains tend to remain associated with the same vessel for extended periods, `device_id` implicitly captures behavioural effects as well.
Differences in steering style, trust in the automation, and operational strategy may all contribute to the observed heterogeneity.
This suggests a rich interaction between human behaviour and automation performance, a topic that has been minimally explored in inland navigation.

5. GLEC Framework effectiveness As described in 8.2.2 the GLEC Framework is the most widely adopted and internationally recognized standard for carbon emissions reporting in inland waterway transport.
It is based on category dependent "intensity factors" describing the emissions per ton of cargo per km sailed.
In this study fuel consumption was directly measured, and CO2 emissions could be calculated using a constant emission factor of $\mathrm{EF}^{\mathrm{WTP}}_{\mathrm{diesel}} = 3240\,\mathrm{g\,CO_2e\,L^{-1}}$.
As shown in Figure 24 estimating the Fuel consumption using the GLEC framework can cause an average error of up to 100% for some ships, the average absolute percentage error in the estimated CO2 emissions was 40% in the dataset used for this study.
This shows that the GLEC framework, although the most common method, can be seriously

flawed, and direct fuel consumption measurements should be used when available.

6. The economic implications of fuel savings As described in Section9.1 the estimated average savings per individual vessel caused by the use of ST-ALA are 0.77% $\pm$ 0.95%.

For an average vessel, assuming a cost of fuel of €0.80 per liter, in the case of a voyage from Rotterdam to Duisburg of about 230 km upstream, 25€ of savings can be expected, and up to 56€ within the confidence interval, while in the case of a voyage from Rotterdam to Basel of about 850 km upstream 92€ in savings can be expected on average and 206€ within confidence interval.

The effect is only marginal when compared with total fuel cost, but nonetheless present. Aggregation over long sailing distances and repeated voyages leads to meaningful cumulative reductions in fuel use, especially for vessels, or fleets with a high utilization rate.

7. The magnitude of the section specific contribution As described in Section 10.3 dividing the Rhine in river segments and aggregating the dataset by segment introduces some limitations, due to necessarily smoothing short-term dynamics(such as rudder activity, rpm and load changes).
A significant advantage of this approach is, on the other hand, the fact that it is easy to estimate the section specific effect of local river characteristics on fuel consumption.
To do this, the SHAP analysis of the prediction models described in Section 8.2 can be used, and a different model trained with the omission of the "edge" variable can be compared with the original ones.

The SHAP analysis shown in Figure 19 shows that the variable "edge" has the third biggest effect on fuel consumption, this is a significant contribution, as according to the model, the local effects have a bigger effect than the total mass of the vessel.

Using the same CatBoost model described in Section 8.2.4, but removing "edge" from the training features yields the following results, as compared to the original model

Table 12: Comparison of predictive performance across models

"edge"present	**RMSE**	R^2	**Median APE**	**P95 APE**
Yes	0.945	0.995	2.6%	9.5%
No	2.275	0.969	5.1%	18.1%

The results show that the model could extract local river characteristics, and understand the relations between them and fuel consumption.

RMSE and RMPE more than double when eliminating edge specific information, the variable "edge" allows for an extra 2,5% improved accuracy of the prediction model.
In the absence of the feature "edge" sailing direction(upstream or downstream) and waterlevel(as measured in koblenz) are the only information that the model has about the conditions of the water, the improvement in accuracy must therefore be due to section specific characteristics that are independent from sailing direction or water level.
These results indicate that local river characteristics can have a significant effect on fuel consumption, and taking them into account is necessary to build a precise fuel prediction model.

8. Results are independent of target feature Instead of using fuel consumption per km as target and outcome, one might argue that fuel consumption should rather be measured in liters per km sailed per ton of cargo transported, as prescribed by the GLEC framework. This measure is closer to the economic meaning of fuel consumption, as ultimately the mass of cargo and the distance to transport it over are the two main drivers of value provided to the customer(given a certain speed), and therefore a KPI used for price calculations as well. Furthermore using this measure might also be relevant since it normalizes on both distance and cargo mass, therefore making edges of different lengths and ships with difference cargo comparable.
A downside of this approach is the fact that only considering the mass of the cargo ignored the dead weight of the ship, ballast weight and other weight, that can have a big effect on fuel consumption.
For this reason the analysis was repeated using fuel consumption per ton (of both total mass and cargo) per km sailed as the outcome, results were extremely similar and were therefore not included explicitly in this report.
This is most likely because the flexible ML models trained using fuel consumption per km, had complete information about both total mass and cargo mass, and since these are confounders, the causal inference process was able to correct for the effect of those variables.

9. The potential effect of crossing the river while sailing As mentioned in 8.1.6, trajectory choice can have a significant effect on fuel consumption.
ST-ALA does not optimize trajectory, but adherence to a chosen one, the chosen track can be shifted by the captain, but frequent trajectory shifting rarely happens, as confirmed by the data(as many edges display similar behavior as shown in Figure 7)
.Most of the voyages done while ST-ALA is active stick to one side of the river while sailing upstream, and keep that side for the entirety of the voyage.
Nonetheless, while upstream, sailing closer to the inside riverbank can be advantageous since the speed of the water is lower and the distance to be covered is less.
Therefore, while sailing upstream when particularly curvy sections of the river are subsequent to each other, constantly changing sides can be a viable strategy for saving fuel.
This is never done by ST-ALA, but some captains tend to employ this strategy.
Sailing in such a way is, understandably, more dangerous than sailing straight, and not always allowed.
Specifically it is forbidden to cross the river in such a way north of Duisburg.

In the dataset this was mitigated by only selecting for trajectories that were taken by the ST-ALA, trajectory choice is therefore not a factor in the causal inference(it was controlled for). But if this were not to be the case, a step difference in behavior is evident around Duisburg while sailing upstream which is shown in image 29(on a dataset when a dataset without this correction was used).

Investigating the effect of trajectory choice on fuel consumption might be particularly relevant.

Figure 29: Fuel consumption savings with and without ST-ALA, obtained from a means analysis per edge on a dataset with no trajectory selection. The colors scale was set to show the step change in behavior in Duisburg and do not reflect the sign of the savings.

Taken together, these findings underscore that ST-ALA's effectiveness is not purely a function of environmental or hydrodynamic conditions, but is strongly tied to vessel- specific mechanical systems and human–automation interaction. This has important implications for both practical deployment and future research.

10.3 Limitations

The findings of this thesis must be interpreted in light of several methodological, data-related, and modelling limitations.
These limitations do not invalidate the results but delineate the scope within which the causal and predictive conclusions are reliable.
They reflect both the fundamental constraints of working with observational vessel-operation data and the simplifying assumptions required by the causal inference and predictive modelling frameworks adopted.

1. Limitations of the Data-Generating Process. Fuel consumption and navigation data are collected through heterogeneous onboard sensors whose calibration, firmware generation, and sampling frequencies vary across vessels and are not always reliable.
Systematic measurement error may therefore arise from sensor drift, latency misalignment between GPS and fuel-flow readings, and differences in hardware generations.
Since this research relies on *aggregated* per-edge quantities, small second-by-second inaccuracies can accumulate unevenly across long or heterogeneous segments.
Aggregation itself introduces attenuation bias: per-edge averages or medians smooth intra-edge variations in steering, acceleration, and hydrodynamic conditions, thereby compressing the signal-to-noise ratio of true fuel consumption dynamics.

Moreover, several hydrodynamically important variables are not measured at all, including precise water current velocity per edge, speed through water, wind and weather conditions, and engine temperature.
These omitted factors increase the risk of residual confounding in both the predictive and causal analyses.
Diesel pricing trends and different contractual fuel-consumption incentives across inland shipping companies and captains may play a role as well.

2. Limitations Specific to Causal Inference. The g-computation estimator depends on three key assumptions: consistency, exchangeability, and positivity. While consistency is reasonable in this context, exchangeability cannot be fully guaranteed in an observational setting.
Captains may activate ST-ALA in situations not fully captured by available covariates (e.g., fatigue, local traffic density, situational awareness, or subjective comfort), leaving open the possibility of unobserved confounding.

Interference is another concern: a captain who experiences improved steering efficiency with ALA may subsequently apply similar behaviour when steering manually, reducing the observable contrast between ALA-ON and ALA-OFF traversals and violating the no-interference component of SUTVA.
Furthermore, ALA adoption is temporally structured (post-2023), creating potential confounding from secular trends such as varying hydrological cycles, maintenance practices, or diesel price–driven behavioural adjustments.

3. Measurement Bias, Missingness, and Selection Effects. Ships equipped with ST-BRAIN hardware (i.e., Shipping Technology clients) constitute a non-random subset of

the European inland fleet.
This introduces selection bias and limits external validity: the estimated ALA effect applies to this fleet rather than to all inland cargo vessels.
Additionally, rows with missing or unreliable data were removed during cleaning; if missingness is not completely random, this may introduce availability bias.

A further limitation arises from *dataset cleaning and filtering decisions*. To construct a usable analytical cohort, hard thresholds were imposed on acceptable ranges for features such as draft, RPM, engine load, and water level, as well as geographical restrictions on where edges were considered valid.
Although justified to remove pathological or erroneous records, these cuts are partly arbitrary and may systematically exclude particular operating regimes, thereby introducing subtle biases into both predictive and causal estimates.

4. Model-Based Limitations. Predictive models require a pre-departure formulation of the problem, relying on average RPM, forecasted water level, medians of draft and weight, and other simplified aggregates.
They do not incorporate time-resolved engine dynamics, transient manoeuvres, or intra-edge hydrodynamic variability.
Both CatBoost and DBPNN assume stationarity of the data-generating process:

$$P_{\text{train}}(Y \mid X) \approx P_{\text{deploy}}(Y \mid X),$$

an assumption that may not hold during extreme hydrological conditions or under long-term behavioural changes in fleet operation.

Tree-based models are robust to nonlinearities but do not extrapolate reliably outside the training range, while DBPNN introduces additional dependence on scaling, initialization, and limited interpretability.
Hyperparameter searches were computationally constrained, so the selected architectures,though highly accurate,may not be globally optimal.

5. Limitations of the Estimated ALA Effect. The estimated average treatment effect is small in magnitude ($\sim 0.77\%$), though statistically significant. Such a small effect lies close to the intrinsic variability of fuel-flow sensors and hydrodynamic fluctuations, although the extensive validation performed confirms that the effect is genuine rather than an artefact of the estimation pipeline.
Nonetheless, the effect size is context-dependent: it may differ under other hydrological regimes, river geometries, operating constraints, or vessel types not represented in the dataset.
Dataset cleaning and cohort selection further restrict external validity, and the effect should therefore be interpreted as applying to conditions similar to those represented in the analysed data.

Overall, these limitations highlight the challenges of deriving causal and predictive insights from complex observational navigation data. They also clarify the boundaries within

which the conclusions of this thesis hold: the estimated ALA effect is valid for the vessels, river conditions, and operational settings represented in the dataset, and the predictive models are reliable within the empirical range of the observed covariates. Despite these constraints, the consistency of the results across multiple modelling approaches and extensive validation procedures supports the robustness of the main findings.

10.4 Future Work

Several avenues for future research emerge from this thesis. These directions would strengthen the causal findings, improve predictive accuracy, and deepen understanding of the mechanisms through which ST-ALA influences fuel consumption.

High-Frequency and Richer Sensor Data. The current analysis relies on 15-second sampling and per-edge aggregates. Future studies should incorporate high-frequency fuel-flow, RPM, rudder-angle data, and speed-through-water measurements. Such data would make it possible to directly analyse steering smoothness, rudder activity, and transient manoeuvres,the likely mechanisms behind ALA-induced fuel savings.

Vessel-Specific and Heterogeneous Treatment Effects. Given the strong vessel-level heterogeneity observed, future work should model vessel-specific responsiveness to ST-ALA using heterogeneous-effect estimators (e.g., causal forests). This would help identify which vessels benefit most and why, informing vessel-specific deployment and tuning strategies.

Mechanical vs. Behavioural Contributions. Controlled experiments or natural experiments (e.g., A/B testing on the same vessel with the same captain) would help separate mechanical benefits from behavioural adaptation, clarifying the causal pathways underlying fuel savings.

Integration of External Environmental Data. Adding hydrological and meteorological information,such as current velocity, wind, weather, and traffic density,could substantially improve both prediction accuracy and causal robustness.

Steering-Control System Modelling. Since ST-ALA interacts with vessel-specific steering-control laws, modelling rudder dynamics and actuator response could explain part of the observed heterogeneity and support the development of adaptive or vessel-specific ALA tuning.

Operational Deployment. Finally, the predictive models developed here could be integrated into decision-support tools,such as voyage-planning dashboards or RPM recommendation engines,to deliver direct operational value.

Overall, future work should combine richer data, more granular modelling, and broader fleet coverage to fully characterise and optimise the fuel-saving potential of autonomous steering systems in inland navigation.

10.5 Reflection on Methodology and Own Work

Conducting this thesis has been an intensive methodological, computational, and conceptual exercise, and it is worth reflecting on the process itself as well as on the strengths and limitations of the chosen approach.

A central enabling factor of this research was the availability of the large-scale dataset provided by Shipping Technology.
In contrast to much of the published academic literature on fuel-consumption prediction,where studies often rely on only a handful of voyages, a limited geographical area, or a single vessel,the dataset used here included millions of aggregated per-edge observations, spanning a diverse set of vessels, river environments, and operating conditions.
This level of coverage made it possible to estimate effects with much greater external realism and to control for local river-section characteristics, which are rarely available in comparable studies.
The dataset was therefore one of the major assets of this thesis, without which the causal analysis and high-fidelity predictive modeling would not have been feasible.

At the same time, the richness and heterogeneity of the data imposed its own challenges. The cleaning process was far more involved than initially expected: measurement inconsistencies across ships, heterogeneous sensor generations, incomplete entries, and operational anomalies required extensive preprocessing, filtering, and range validation.
Inconsistent definitions of measured variables were encountered across ships (for example, engine load or draft).
A large amount of work,important but not directly methodological,was needed simply to transform the raw data into a form suitable for causal inference and predictive modeling, or to ensure the comparability of the same features across different ships.
This was an instructive reminder that working with real-world operational data often involves substantial invisible labour, and that methodological elegance depends critically on disciplined data engineering.

A second area of reflection concerns the causal inference framework itself. Causal inference was not part of the formal coursework of this programme, and the need to independently acquire the relevant concepts (exchangeability, positivity, consistency, the g-formula, inverse probability weighting, and influence-function-based inference) represented a substantial additional challenge.
Guiding sources such as Hernán and Robins' *Causal Inference: What If* [8] were essential in building the foundation required to develop a defensible analytical pipeline.
Because there is essentially no prior work investigating the causal effect of autonomous lane-assist systems on fuel consumption in inland shipping, there was little precedent to follow.
The absence of similar studies was at times discouraging; having a comparable reference would likely have made certain methodological decisions more straightforward. Nevertheless, this necessitated an exploratory and critical mindset throughout the research, while still drawing inspiration from the literature where possible, for example from [21] when implementing the DBPNN prediction model in Section 8.2.5.

Due to time constraints, numerous simplifying assumptions were necessary,particularly regarding stationarity, unmeasured confounding, and the aggregation of high-frequency dynamics into per-edge summaries. Many limitations remain, as discussed in Section 10.3. These constraints should not be seen as oversights but as inherent to the scope of a master's thesis carried out on operational data with limited theoretical precedent.

Finally, an important part of the research process was embracing the exploratory and iterative approach described in Section 5. The savings results obtained from the causal analysis were smaller than expected based on sparse external references, and this discrepancy caused repeated re-evaluation of the models, assumptions, and diagnostics. Combined with the relative novelty of g-computation in this application domain and the heterogeneity of the dataset, this led to numerous moments of doubt about the correctness of the pipeline. However, these iterations (repeated model validation, construction of placebo datasets, and extensive diagnostic checks) ultimately made the conclusions more robust and deepened the understanding of the methodology.

Overall, this thesis was both a statistical and conceptual learning process. It clarified how challenging real-world causal inference can be, how essential rigorous diagnostics are, and how much of applied machine learning depends on careful data preparation.

References

[1] American Bureau of Shipping. Energy efficiency advisory: Improving energy performance through operational measures and technologies. Technical report, American Bureau of Shipping (ABS), 2022. Technical report, accessed October 2025.

[2] Anschütz GmbH. Anschütz autopilots save up to 5% fuel on twin-rudder vessels, 2023. Online company report, accessed October 2025.

[3] Moritz Buchem, Julian Arthur Pawel Golak, and Alexander Grigoriev. Vessel velocity decisions in inland waterway transportation under uncertainty. *European Journal of Operational Research*, 296(2):669–678, 2022.

[4] Philip Cammin, Jingjing Yu, and Stefan Voß. Tiered prediction models for port vessel emissions inventories. *Flexible Services and Manufacturing Journal*, 35(1):142–169, March 2023.

[5] Arthur Chatton, Florent Le Borgne, Clémence Leyrat, Florence Gillaizeau, Chloé Rousseau, Laetitia Barbin, David Laplaud, Maxime Léger, Bruno Giraudeau, and Yohann Foucher. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Scientific Reports*, 10:9219, 2020.

[6] Andrea Coraddu, Luca Oneto, Francesco Baldi, and Davide Anguita. Vessels fuel consumption forecast and trim optimisation: A data analytics perspective. *Ocean Engineering*, 130:351–370, January 2017.

[7] Stefan Nygaard Hansen and Morten Overgaard. Variance estimation for average treatment effects estimated by g-computation. *Metrika*, 88(4):419–443, 2024.

[8] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2024.

[9] Loukas Ilias, Panagiotis Kapsalis, Spiros Mouzakitis, and Dimitris Askounis. A Multitask Learning Framework for Predicting Ship Fuel Oil Consumption. *IEEE Access*, 11:132576–132589, 2023.

[10] Miyeon Jeon, Yoojeong Noh, Yongwoo Shin, O-Kaung Lim, Inwon Lee, and Daeseung Cho. Prediction of ship fuel consumption by using an artificial neural network. *Journal of Mechanical Science and Technology*, 32(12):5785–5796, December 2018.

[11] Pavlos Karagiannidis and Nikos Themelis. Data-driven modelling of ship propulsion and the effect of data pre-processing on the prediction of ship fuel consumption and speed loss. *Ocean Engineering*, 222(108616):108616, February 2021.

[12] Young-Rong Kim, Min Jung, and Jun-Bum Park. Development of a Fuel Consumption Prediction Model Based on Machine Learning Using Ship In-Service Data. *Journal of Marine Science and Engineering*, 9(2):137, January 2021.

[13] Lin Lei, Zecheng Wen, and Zhongbo Peng. Prediction of Main Engine Speed and Fuel Consumption of Inland Ships Based on Deep Learning. *Journal of Physics: Conference Series*, 2025(1):012012, September 2021.

[14] James G. MacKinnon, Morten Ørregaard Nielsen, and Matthew D. Webb. Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, 232(2):272–299, 2023.

[15] Christos Papandreou and Antonis Ziakopoulos. Predicting VLCC fuel consumption with machine learning using operationally available sensor data. *Ocean Engineering*, 243:110321, 2022.

[16] Smart Freight Centre / Global Logistics Emissions Council. Ghg emission factors for inland waterway transport (iwt). `https://www.smartfreightcentre.org`, 2018. STC-NESTRA report for the GLEC Framework. Accessed July 2, 2025.

[17] Smart Freight Centre / Global Logistics Emissions Council. Global logistics emissions council. `https://www.smartfreightcentre.org/en/our-programs/emissions-accounting/global-logistics-emissions-council/`, 2025. Accessed July 2, 2025.

[18] Zeynep Tacar, Noriyuki Sasaki, Mehmet Atlar, and Emin Korkut. An investigation into effects of gate rudder® system on ship performance as a novel energy-saving and manoeuvring device. *Ocean Engineering*, 218:108250, 2020.

[19] Tresco Engineering GmbH. Fuel savings from smart steering with trackpilot, 2024. Company presentation and promotional material, accessed October 2025.

[20] Wikipedia contributors. Simpson's paradox — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/Simpson%27s_paradox`, 2025. Accessed: 2025-10-29.

[21] Zhi Yuan, Jingxian Liu, Yi Liu, Yuan Yuan, Qian Zhang, and Zongzhi Li. Fitting Analysis of Inland Ship Fuel Consumption Considering Navigation Status and Environmental Factors. *IEEE Access*, 8:187441–187454, 2020.

[22] Zhi Yuan, Jingxian Liu, Qian Zhang, Yi Liu, Yuan Yuan, and Zongzhi Li. Prediction and optimisation of fuel consumption for inland ships considering real-time status and environmental factors. *Ocean Engineering*, 221:108530, February 2021.

[23] Yongjie Zhu, Yi Zuo, and Tieshan Li. Modeling of Ship Fuel Consumption Based on Multisource and Heterogeneous Data: Case Study of Passenger Ship. *Journal of Marine Science and Engineering*, 9(3):273, March 2021.

A Dataset Variables

This appendix documents all variables used in the analysis. It distinguishes between *original* variables provided by Shipping Technology and *computed* variables derived as part of this thesis. For each variable, the meaning, aggregation method, unit, type, and (for computed variables) purpose are described.

Unless otherwise specified, the underlying sensor sampling rate is 1 Hz. Trajectory linestrings are sampled at 15-second intervals.

A.1 Original Variables

The original dataset consists of one record per traversal of a directed river segment (edge). The following variables are provided:

- **edge** — Directed river segment identifier, defined by entry and exit nodes of a Voronoi polygon.
  Aggregation: constant per traversal.
  Unit: none.
  Type: string.

- **prev_edge** — Identifier of the previous edge traversed by the ship.
  Aggregation: constant per traversal.
  Unit: none.
  Type: string.

- **next_edge** — Identifier of the next edge entered after this traversal.
  Aggregation: constant per traversal.
  Unit: none.
  Type: string.

- **device_id** — Unique identifier of the ST-BRAIN device onboard (e.g. `st-7ct479`). Each device is specific to a single ship and serves as a proxy for the vessel and its typical crew.
  Aggregation: constant per traversal.
  Unit: none.
  Type: string.

- **center_3035** — Midpoint of the traversal in EPSG:3035 (ETRS89 / LAEA Europe), expressed as `POINT (x y)`.
  Aggregation: geometry-derived.
  Unit: meters (x, y).
  Type: string.

- **center_4326** — Midpoint of the traversal in EPSG:4326 (WGS84), expressed as `POINT (lon lat)`.
  Aggregation: geometry-derived.
  Unit: degrees.
  Type: string.

- **linestring_3035** — Polyline of the vessel trajectory through the edge in EPSG:3035. Points are sampled at 15-second intervals.
  Aggregation: sampled every 15 seconds.
  Unit: meters (x, y).
  Type: string.

- **linestring_4326** — Same as above, expressed in EPSG:4326 (WGS84).
  Aggregation: sampled every 15 seconds.
  Unit: degrees.
  Type: string.

- **start_time** — UTC timestamp marking when the vessel enters the edge, rounded to the 15-second sampling grid.
  Aggregation: first sample in edge.
  Unit: seconds (UTC).
  Type: datetime.

- **end_time** — UTC timestamp when the vessel exits the edge, rounded to the 15-second sampling grid.
  Aggregation: last sample in edge.
  Unit: seconds (UTC).
  Type: datetime.

- **contains_stop** — Boolean flag indicating whether vessel speed dropped below 1 knot for more than 1 minute during the traversal. Used to filter traversals with stops.
  Aggregation: rule-derived.
  Unit: none.
  Type: boolean.

- **distance** — Length of the trajectory along the edge.
  Aggregation: geometry-derived.
  Unit: meters.
  Type: float.

- **avg_speed** — Average of instantaneous speeds measured by onboard sensors during traversal. All speeds below 1 knot are not recorded and thus excluded from the average.
  Aggregation: mean.
  Unit: knots.
  Type: float.

- **total_time** — Duration of the traversal, `end_time` − `start_time`, rounded to the 15-second grid.
  Aggregation: difference.
  Unit: seconds.
  Type: float.

- **run_id** — Internal ETL identifier, not used analytically.
  Aggregation: none.

Unit: none.
Type: string.

- **med_draft_front** — Median immersion at a fixed bow sensor point. Median is used to suppress sensor noise due to turbulence and air bubbles.
  Aggregation: median.
  Unit: centimeters.
  Type: float.

- **med_draft_back** — Median immersion at a fixed stern sensor point.
  Aggregation: median.
  Unit: centimeters.
  Type: float.

- **med_weight** — Median cargo mass transported during the traversal.
  Aggregation: median.
  Unit: kilograms.
  Type: float.

- **avg_engine_rpm** — Average revolutions per minute across all engines.
  Aggregation: mean.
  Unit: rpm.
  Type: float.

- **avg_engine_load** — Average relative engine load at the current rpm, expressed as a percentage of the maximum at that rpm.
  Aggregation: mean.
  Unit: percent.
  Type: float.

- **avg_engine_fuel_consumption** — Mean of instantaneous fuel-rate readings (liters per hour) averaged over time and engines.
  Aggregation: mean of per-engine means.
  Unit: L/h.
  Type: float.

- **med_depth** — Median water depth under the hull measured by the depth sensor (local vertical datum). Median is used to mitigate noise from bubbles, vegetation, or debris.
  Aggregation: median.
  Unit: meters.
  Type: float.

- **waterlevel_tiel_waal** — River water level from the Tiel–Waal gauge. Each station uses its own vertical datum; values are not directly comparable across gauges.
  Aggregation: sampled and joined.
  Unit: centimeters.
  Type: float.

- **waterlevel_koblenz** — River water level from the Koblenz gauge, subject to the same local-datum remark.
  Aggregation: sampled and joined.
  Unit: centimeters.
  Type: float.

- **ala_sessions** — Number of separate intervals of ALA activation during the traversal.
  Aggregation: count.
  Unit: unitless.
  Type: float.

- **n_engines** — Number of main engines onboard.
  Aggregation: constant per ship.
  Unit: unitless.
  Type: float.

- **ala_active_seconds** — Total duration of ALA activation during the traversal.
  Aggregation: sum.
  Unit: seconds.
  Type: float.

- **build_year** — Year in which the ship was built.
  Aggregation: constant per ship.
  Unit: year.
  Type: integer.

- **ship_type** — Vessel category: cargo, tanker, container, passenger, or roll-on/roll-off.
  Aggregation: constant per ship.
  Unit: none.
  Type: string.

- **dimensions_width** — Registered ship width.
  Aggregation: constant per ship.
  Unit: meters.
  Type: integer.

- **dimensions_length** — Registered ship length.
  Aggregation: constant per ship.
  Unit: meters.
  Type: integer.

A.2 Computed Variables

In addition to the original variables, a number of derived variables were computed for the purposes of data cleaning, feature engineering, and analysis.

- **trav_id** — Sequential numeric identifier uniquely representing each traversal. Computed as an incremental index after grouping by `device_id` and sorting by `start_time`.
  Aggregation: constant per traversal.
  Unit: none.
  Type: integer.
  Purpose: uniquely identify traversals and enable joins across datasets.

- **max_turn_angle_deg** — Maximum local turning angle between consecutive segments of the trajectory, computed from `linestring_4326` as the maximum absolute difference in heading between successive 15-second samples.
  Aggregation: maximum.
  Unit: degrees.
  Type: float.
  Purpose: detect unrealistic GPS artefacts or extreme rudder movements during data cleaning.

- **prev_trav_id** — Identifier of the previous traversal by the same vessel on the reverse edge, obtained by matching `device_id` and `prev_edge` and selecting the record whose `end_time` is closest (and before) the current `start_time`.
  Aggregation: constant per traversal.
  Unit: none.
  Type: integer.
  Purpose: infer dynamic context such as whether the vessel was accelerating after a stop.

- **next_trav_id** — Identifier of the following traversal by the same vessel on the next edge, computed analogously to `prev_trav_id` using `next_edge`.
  Aggregation: constant per traversal.
  Unit: none.
  Type: integer.
  Purpose: detect decelerating or braking phases in subsequent traversals.

- **fuel_consumption** — Total fuel consumed during the traversal:

$$\text{fuel\_consumption} = \text{avg\_engine\_fuel\_consumption} \times \text{n\_engines} \times \frac{\text{total\_time}}{3600}.$$

  Aggregation: derived scalar.
  Unit: liters.
  Type: float.
  Purpose: base fuel metric for total consumption comparisons.

- **fuel_consumption_lh** — Mean total fuel rate:

$$\text{fuel\_consumption\_lh} = \text{avg\_engine\_fuel\_consumption} \times \text{n\_engines}.$$

  Aggregation: derived scalar.
  Unit: L/h.

Type: float.
Purpose: normalise fuel usage to an hourly rate.

- **fuel_consumption_km** — Fuel consumption per kilometre travelled:

$$\text{fuel\_consumption\_km} = \left(\frac{\text{fuel\_consumption}}{\text{distance}}\right) \times 1000.$$

Aggregation: derived scalar.
Unit: L/km.
Type: float.
Purpose: main normalised fuel metric used as model target.

- **fuel_consumption_ton_km** — Fuel consumption per kilometre per ton of displaced mass (ship deadweight, cargo, ballast, equipment):

$$\text{fuel\_consumption\_ton\_km} = \left(\frac{\text{fuel\_consumption\_km}}{\text{total\_mass}}\right) \times 1000.$$

Aggregation: derived scalar.
Unit: L/(ton·km).
Type: float.
Purpose: mass-normalised fuel metric.

- **med_draft** — Average of bow and stern draft medians:

$$\text{med\_draft} = \frac{\text{med\_draft\_front} + \text{med\_draft\_back}}{2}.$$

Aggregation: mean of medians.
Unit: centimeters.
Type: float.
Purpose: represent vessel immersion for buoyancy and resistance modelling.

- **med_trim** — Draft difference between bow and stern:

$$\text{med\_trim} = \text{med\_draft\_front} - \text{med\_draft\_back}.$$

Aggregation: difference of medians.
Unit: centimeters.
Type: float.
Purpose: describe trim-induced resistance effects.

- **computed_speed** — Mean geometric speed over ground:

$$\text{computed\_speed} = \frac{\text{distance}}{\text{total\_time}}.$$

Aggregation: derived scalar.
Unit: m/s.
Type: float.
Purpose: speed estimate independent of speed-sensor bias.

- **environment** — Categorical descriptor of flow direction relative to river current (`UPSTREAM`, `DOWNSTREAM`, `TIDAL`), computed using a proprietary function supplied by Shipping Technology.
  Aggregation: constant per edge.
  Unit: none.
  Type: string.
  Purpose: key control variable for hydrodynamic regime.

- **lat** — Latitude extracted from `center_4326`.
  Aggregation: constant per traversal.
  Unit: degrees.
  Type: float.
  Purpose: spatial reference for mapping.

- **lon** — Longitude extracted from `center_4326`.
  Aggregation: constant per traversal.
  Unit: degrees.
  Type: float.
  Purpose: spatial reference for mapping.

- **rough_cog** — Approximate course-over-ground, computed from the first and last points in `linestring_4326`.
  Aggregation: derived scalar.
  Unit: degrees.
  Type: float.
  Purpose: rough orientation metric.

- **ala_active_percentage** — Ratio of ALA-active duration to total traversal time:

$$\text{ala\_active\_percentage} = \frac{\text{ala\_active\_seconds}}{\text{total\_time}}.$$

  Aggregation: ratio.
  Unit: fraction (0–1).
  Type: float.
  Purpose: quantify fraction of time under autonomous control.

- **ala_on** — Boolean indicator for full ALA activation, defined as `True` if ala_active_percentage $>$ 0.9, else `False`.
  Aggregation: logical rule.
  Unit: none.
  Type: boolean.
  Purpose: main binary treatment variable for causal inference.

- **prev_trav_contains_stop** — `True` if the previous traversal (via `prev_trav_id`) contains a stop; `Null` if no preceding traversal exists.
  Aggregation: boolean condition.
  Unit: none.

Type: boolean.
Purpose: detect accelerating situations.

- **next_trav_contains_stop** — `True` if the following traversal (via `next_trav_id`) contains a stop; `Null` if no subsequent traversal exists.
Aggregation: boolean condition.
Unit: none.
Type: boolean.
Purpose: detect decelerating or braking phases.

- **total_mass** — Estimated total displaced mass of the vessel:

$$\text{total\_mass} = \rho\, L\, W\, d,$$

where $\rho = 1000\,\text{kg/m}^3$, L = `dimensions_length`, W = `dimensions_width`, d = med_draft.
Aggregation: computed per traversal.
Unit: kilograms.
Type: float.
Purpose: proxy for loading condition and hydrodynamic resistance; used in defining `fuel_consumption_ton_km`.

- **hc_mean_abs_deg** — Mean absolute change in heading between consecutive 15-second samples along the trajectory. Heading is computed for each segment, wrapped to $[-180°, 180°]$, and the mean of absolute differences is taken.
Aggregation: mean.
Unit: degrees.
Type: float.
Purpose: measure of average trajectory curvature / steering activity.

- **hc_std_abs_deg** — Standard deviation of the absolute heading changes along the trajectory.
Aggregation: standard deviation.
Unit: degrees.
Type: float.
Purpose: indicates variability of steering activity during the traversal.

- **hc_mean_rel** — Deviation of each traversal's mean heading change from the average of all traversals on the same edge:

$$\text{hc\_mean\_rel} = \text{hc\_mean\_abs\_deg} - \text{mean}(\text{hc\_mean\_abs\_deg over same edge}).$$

Aggregation: deviation from per-edge mean.
Unit: degrees.
Type: float.
Purpose: compare how smooth a traversal is relative to others on the same river segment.

- **hc_std_rel** — Deviation of each traversal's steering variability from the average variability on the same edge:

$$\text{hc\_std\_rel} = \text{hc\_std\_abs\_deg} - \text{mean}(\text{hc\_std\_abs\_deg over same edge}).$$

Aggregation: deviation from per-edge mean.
Unit: degrees.
Type: float.
Purpose: highlight traversals steered more or less consistently than typical for that edge.

B Top Regions of the Device-Adjusted CATE

Overview

The table below reports the highest-ranking regions according to the **device-adjusted Conditional Average Treatment Effect (CATE)**. Each region corresponds to a combination of the binned continuous variables (`avg_engine_rpm`, `avg_engine_load`, `med_depth`, `total_mass`) and the categorical attributes `environment` and `ship_type`.

For each region, the table lists:

- the feature-bin combination that defines the region;
- number of unique ships in that region (`n_ships_region`);
- number of traversals considered (`n_rows_region`);
- the device-adjusted CATE (`cate_centered`);
- its standard error and confidence interval bounds.

These entries highlight the operational regimes where ST-ALA provides the largest *additional* savings beyond each ship's own baseline average. As discussed in the main text, the top regions share a consistent pattern: *low RPM, low engine load, low total mass, and downstream environment*. These represent hydrodynamically favorable states where smoother steering corrections from ALA translate into relatively larger marginal gains. Importantly, these regions include substantial sample sizes and multiple ships, confirming that the observed effects are not driven by a single vessel.

rpm bin	load bin	depth bin	mass bin	env	ship type	n ships	n rows	CATE adj.	SE	CI low	CI high
0–33	0–33	67–100	0–33	DOWNSTREAM	tanker	24	3173	0.726991	0.005771	0.715797	0.738186
0–33	0–33	33–67	0–33	DOWNSTREAM	cargo	5	748	0.664006	0.010848	0.642743	0.685268
0–33	33–67	67–100	0–33	DOWNSTREAM	tanker	12	2793	0.601897	0.005277	0.591555	0.612239
0–33	0–33	67–100	33–67	UPSTREAM	container	5	141	0.591882	0.015202	0.553831	0.630074
0–33	0–33	67–100	33–67	TIDAL	tanker	16	650	0.566642	0.012129	0.542869	0.590415
0–33	0–33	33–67	0–33	DOWNSTREAM	tanker	13	11739	0.539884	0.000823	0.538270	0.541497
0–33	33–67	33–67	67–100	DOWNSTREAM	tanker	5	206	0.532617	0.012575	0.507971	0.557264
0–33	0–33	67–100	33–67	DOWNSTREAM	tanker	11	2046	0.513274	0.005316	0.502659	0.523689
0–33	33–67	33–67	67–100	UPSTREAM	container	5	771	0.47084	0.009642	0.451941	0.489739
0–33	0–33	33–67	33–67	DOWNSTREAM	tanker	14	1499	0.430628	0.008448	0.413998	0.447258
0–33	33–67	67–100	67–100	DOWNSTREAM	tanker	18	4805	0.409827	0.00504	0.399968	0.419725
0–33	0–33	33–67	33–67	DOWNSTREAM	container	5	901	0.413679	0.009187	0.395672	0.431686
0–33	33–67	33–67	33–67	TIDAL	tanker	10	153	0.421975	0.016985	0.388505	0.455085

Table 13: Top-ranked regions according to the device-adjusted Conditional Average Treatment Effect (CATE). Higher CATE values represent larger additional ALA-attributable fuel savings beyond each vessel's own baseline mean.

A Variable Dictionary

A full variable dictionary, including aggregation rules, units, and data types for every column in the dataset, will be provided here.