iHelm Porject Summary Report

Mohamed Abuella

December 7, 2023

Contents

Executive Summary

To meet the urgent requirements for the climate change mitigation, several proactive measures of energy efficiency have been implemented in maritime industry. Many of these practices depend highly on the onboard data of vessel's operation and environmental conditions.

In this project, a high resolution onboard data from passenger vessels in short-sea shipping (SSS) have been collected and preprocessed.

We first investigated the available data to deploy it effectively to model the physics of the vessel, and hence the vessel performance. Since in SSS, the weather measurements and forecasts might have not been in temporal and spatial resolutions that accurately representing the actual environmental conditions.

Then, We proposed a data-driven modeling approach for vessel energy efficiency. This approach addresses the challenges of data representation and energy modeling by combining and aggregating data from multiple sources and seamlessly integrates explainable artificial intelligence (XAI) to attain clear insights about the energy efficiency for a vessel in SSS.

After that, the developed model of energy efficiency has been utilized in developing a framework for optimizing the vessel voyage to minimize the fuel consumption and meeting the constraint of arrival time.

Moreover, we developed a spatial clustering approach for labeling the vessel paths to detect the paths for vessels with operating routes of repeatable and semi-repeatable paths.
This report summarizing the major methods and findings of iHelm project.

1 Introduction

1.1 General Background

Short-Sea Shipping (SSS) is a commercial transportation mode that does not involve intercontinental cross-ocean. The SSS provides a cost-efficient and environment-friendly alternative for transportation by utilizing inland and coastal waterways to transport the commercial freight [4].

On the other hand, the SSS produces some negative effects on the natural habitats and polluting the air along the coasts of populated cites [10]. As a response to this, the International Maritime Organization (IMO) have conducted many studies and recommended standards and imposed policies for the maritime sector to reduce the carbon dioxide (CO_2) to 40% by 2030 and cut 50% of all GHGs by 2050, based on the emissions in 2008 [7].

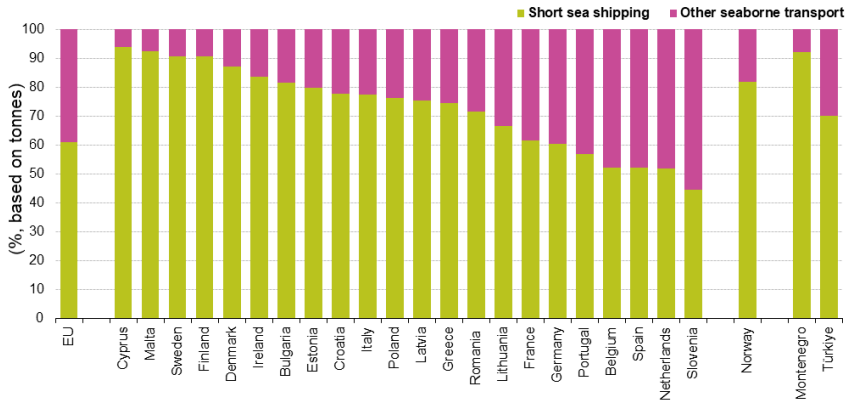


Figure 1: European short-sea shipping of freight versus total sea transport, in 2021 [11].

Furthermore, COVID-19 pandemic has accelerated the digitalization of the entire shipping industry globally, and hence attracted a profound consideration to data collection and prepara-

tion stages [8]. The operational and some environmental conditions can be accessed through an Automatic Identification System (AIS) messages, which is a service developed by the International Maritime Organization (IMO) in 2002 to record the sensor measurements and send the vessel position information for the traffic between other ships and neighboring shores [12].

In a broader perspective, for improving the vessel's energy efficiency and harnessing more fuel savings and less GHG emissions, there are mainly two procedures. The first strategy is in the ship design stage, where the ship is built to obtain a body and equipped machinery that work efficiently. The second strategy is during the ship operation over the water or at the ports. This latter procedure can be achieved by adopting energy management plans that optimally enhance the energy efficiency and fuel consumption [13].

This project proposes a data-driven framework for modeling and optimizing approaches to improve energy efficiency in short-sea shipping. In addition, a framework for vessel path identification has been developed for route planning and management of resources in maritime industry.

1.2 Main Outcomes

The main outcomes of this project can be summarized as follow:

- Modeling of energy efficiency: Develop a data-driven model for voyage energy efficiency, including:
  A spatiotemporal aggregation of operation and navigation data from onboard and external sources to capture the impact of both spatial and temporal factors on voyage energy efficiency.
  Introduce an efficiency score that considers both total fuel consumption and voyage duration to measure the voyage energy efficiency.

- Data clustering: Clustering the data of voyages and sorting them based on their efficiency scores. This clustering enables the voyage optimization algorithm to learn more insights for better actions, either by selecting the best voyages or by eliminating the worst voyages.

- Time-series analysis models and comparative analysis: Four time-series based models are implemented as algorithms of voyage speed optimization. Then, a rigorous evaluation of their performance is conducted across different data clusters and using metrics that account for voyage efficiency.

- Practical implication: Demonstrate the significant effectiveness and practicality of the proposed approach for fixed-route vessels in short sea shipping, where the options for obtaining efficient voyages are limited. The approach also aligns with the guidance of domain experts, adhering to safety and traffic considerations.

- The path clustering approach has a proven added value for clustering semi-defined paths.

- The hierarchical-based clustering approach has a customized parameter to determine the number of path classes, thereby enhancing the flexibility and adaptability of the framework, allowing users to tailor it to their specific needs.

- The path clustering approach is robust and interpretable by applying a similarity measure that reduces the influence of noise or outliers and offers a clear interpretation of path clustering.

- The path clustering approach has a customized parameter to determine the number of path classes, thereby enhancing the flexibility and adaptability of the framework, allowing users to tailor it to their specific needs.

- The framework is a data-driven solution that can be used as a valuable asset for informed decision-making in route planning and optimization, traffic management, and resource allocation.

2 Data Collection and Preparation

2.1 Data Collection

The ship's onboard data have been received from our industry partner CetaSol AB in Gothenburg [1]. The data has been gathered over a period of 15 months, between January 2020 and March 2021. It has a 3Hz frequency and records about the ship's position, course direction, and speed. It is also including some of operational and meteorological data, such as fuel rate, engine speed, torque, acceleration, wind speed and direction.

Some information about the ship and its voyage can be found on Marine Traffic website [5]. Other weather variables such as wave height and sea current speed and direction have been collected from external sources, Copernicus Marine Service [2] and Stormglass [3] APIs.

Table 1: The navigational variables and their data sources.

Variable	Source	Variable	Source
Latitude	Onboard	WindSpeed_cps	Copernicus
Longitude	Onboard	WindDirection_cps	Copernicus
SpeedOverGround	Onboard	WaveHeight	Copernicus
HeadingMagnetic	Onboard	WaveDirection	Copernicus
Pitch	Onboard	WindSpeed_sg	Stormglass
Roll	Onboard	WindDirection_sg	Stormglass
WindSpeed_onb	Onboard	CurrentSpeed	Stormglass
WindDirection_onb	Onboard	CurrentDirection	Stormglass

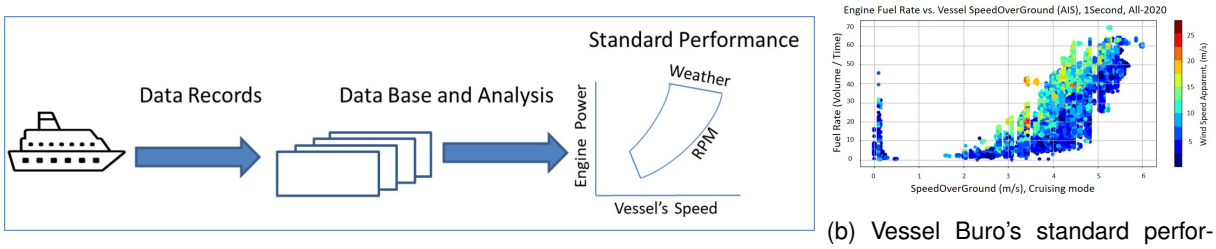
2.2 Data Preparation and Validation

The external weather data are past forecasts (hindcasts), which have reanalysed to become hourly in temporal resolution and with 0.25 to 0.5 degree as a spatial resolution. Trilinear interpolation in time and space dimensions has been applied on external weather data to be more suitable for time and position frames of the given vessel routing. Therefore, the weather and onboard data are used in this analysis with a temporal resolution of 1-minute in average.

The data validation is conducted through the cruising-speeds mode is to reduce the other vessel effects on the fuel consumption, and thus, producing graphs that can be then compared with the general ship's standard performance. The operational and weather data validation is carried out visually, as shown in Figure 2.

3 Methodology

Energy efficiency modeling and data clustering is a vital component of our framework. The primary objective is to identify and sort the voyages based on their efficiency scores. Then, train the models with the sorted data clusters iteratively, to distill insights from the voyages with different behaviours. Thus, the trained models will gain valuable insights into the performance and operational patterns of the vessel.



(a) General data analytics for ship's standard performance [9].

(b) Vessel Buro's standard performance

Figure 2: Vessel's data analytics and standard performance graph for the case study

3.1 Energy Efficiency Modeling

This part presents a mathematical and visual overview of the fundamental theoretical background that forms the basis for modeling vessel energy efficiency—an indispensable element within our comprehensive framework.

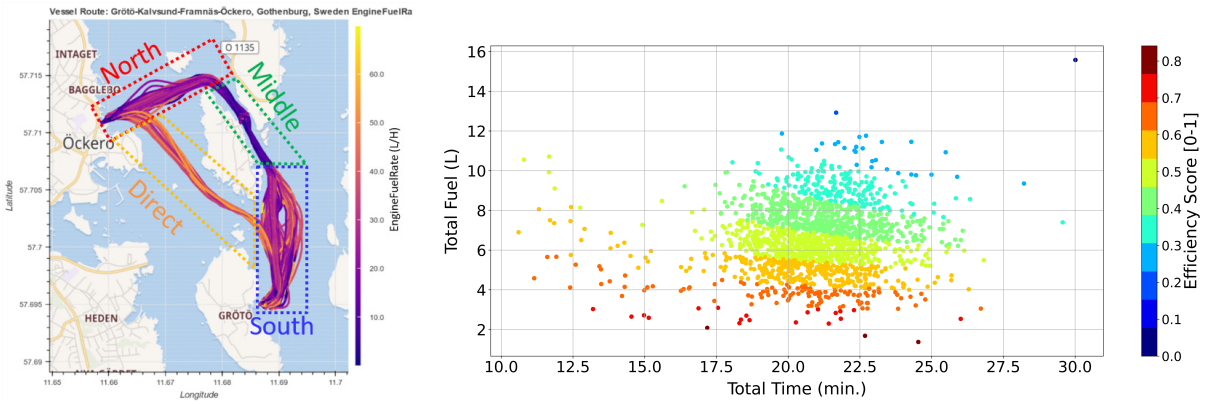
To estimate the vessel's energy efficiency in this framework of voyage optimization, we employed a previously developed model equipped with artificial intelligence (XAI) and machine learning techniques. More details about this energy efficiency modeling approach can be found in [6].

The efficiency score ($\mathrm{Eff}_{\mathrm{score}}$) is calculated from the normalized total fuel and time for every voyage, as following:

$$\mathrm{Eff}_{\mathrm{Score}} = 1 - \frac{2 \times [Fuel_{Tl_{Nm}} \times Time_{Tl_{Nm}}]}{[Fuel_{Tl_{Nm}} + Time_{Tl_{Nm}}]} \tag{1}$$

The efficiency score considers the proportional reduction in both fuel consumption and time, assessing the vessel's efficient use of resources during the voyage.

Figure 3 facilitates to visualize the process of aggregation for the vessel's voyages. First by illustrating the voyages in terms of space (i.e., latitude and longitude) as shown in Figure 3a, and second by representing the aggregated voyages as points in new dimensions of Efficiency Score versus total fuel and total time. The aggregated data and its new dimensions are projected as in Figure 3b.



(a) Vessel's route.

(b) Data are projected by Efficiency Score vs. fuel and time.

Figure 3: The vessel voyages and the aggregated data projected by the Efficiency Score.

The representation of voyages in terms of fuel and time is done by adopting the concept of the Efficiency Score (Eff-Score). The efficiency scores for all vessel's voyages are presented in Figure 3b. It is evident that the voyage with lower fuel and shorter time have higher efficiency scores, and vice versa.

Algorithm 1: Modeling of Energy Efficiency and Clustering of Voyages Data Based on Their Energy Efficiency

Data: Voyages data of the vessel
Result: Clusters of Voyages
Load the operational and navigational data, including speed, course, fuel, position, distance, and weather;
Tag the datapoints to its corresponding voyage, V_{id};
for *each voyage* V_i *in voyages data* **do**
    Calculate total fuel consumption and time for V_i;
    Normalize total fuel and time for V_i based on their maximum values of all voyages;
    Calculate the *Eff-Score* as described in Eq. (1), and assign it to all datapoints of this voyage V_i;
Initialize four empty lists for each cluster: $Top75Pr$, $Top50Pr$, $Top25Pr$, $Top10Pr$ (Percentiles of Eff-Scores);
for *each data point in all data* **do**
    Extract the Eff-Score of the data point;
    if *Eff-Score* $\geq$ *0.4070* **then**
        Append the data point to $Top75Pr$;
    else if *Eff-Score* $\geq$ *0.4623* **then**
        Append the data point to $Top50Pr$;
    else if *Eff-Score* $\geq$ *0.5220* **then**
        Append the data point to $Top25Pr$;
    else if *Eff-Score* $\geq$ *0.5730* **then**
        Append the data point to $Top10Pr$;

There are four voyage data clusters, namely $Top75Pr$, $Top50Pr$, $Top25Pr$, and $Top10Pr$, as shown in Figure 4, These clusters are categorized on their respective Eff-Score percentiles, enabling a structured analysis of voyage efficiency across various percentile groups.

We introduced a data-driven approach for modeling of vessel energy efficiency, by integrating data from various sources and employing explainable artificial intelligence (XAI) and artificial neural network (ANN) to gain clear insights into a vessel's energy efficiency in SSS. For more details, refer to our publication [6].
The workflow for modeling of vessel energy efficiency is shown in Figure 5.

3.2 Framework of Voyage Optimization

One of the main objective of this project is to improve the vessel voyage by optimizing its speed to enhance the vessel's energy efficiency. In other words, reducing the vessel's fuel consumption within constrained arrival time.
The framework of the developed approach for improving the vessel voyages is depicted in Figure 6.

For the purpose of vessel voyage optimization, we need a model to optimize the vessel's speed profile for improving the vessel's energy efficiency. This ideal model should mainly be able to:

- Model the temporal dependencies in vessel speed profiles.
- Incorporate external variables, such as weather conditions, into the modeling process.
- Adapt to changing conditions to provide real-time optimized speed profiles.

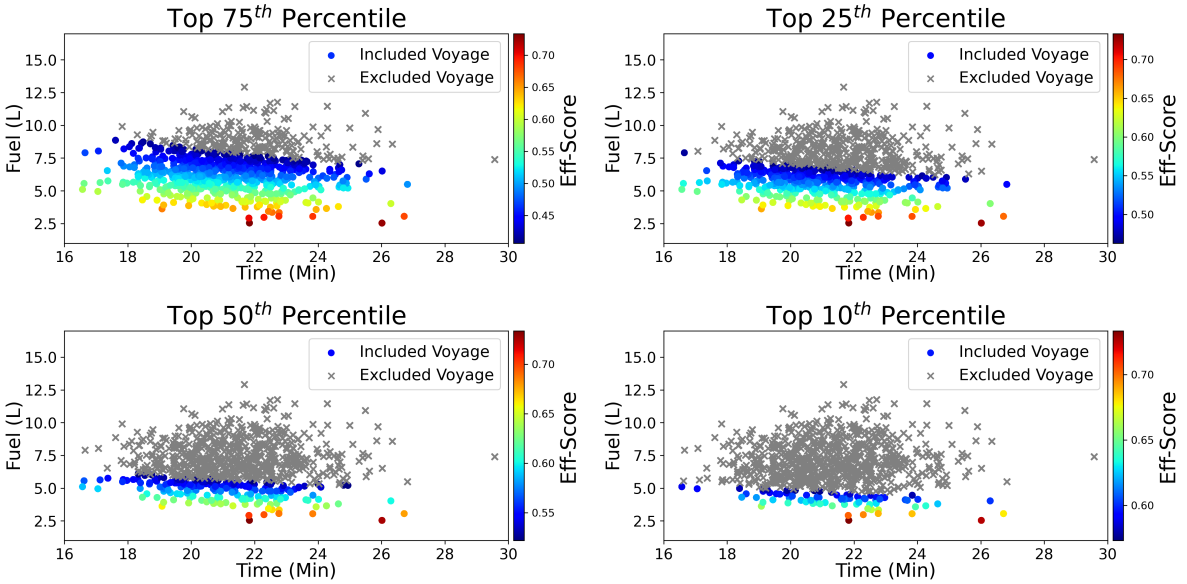


Figure 4: Four clusters of voyages based on their efficiency.

In order to meet these requirements, we adopt several time-series analysis models, specifically they are: Long Short-Term Memory (LSTM), Dynamic Time Warping (DTW), k-Nearest Neighbors (kNN), and Hidden Markov Model (HMM).

3.3 Framework of Path Identification

The theoretical background and description of the underling methodology of our proposed framework are covered within this section.
The framework of vessel path identification is depicted in Figure 7.

3.3.1 Problem Formulation

The equations (2-4) serve as a mathematical representation to describe the clustering of vessel paths. It is worth mentioning that the clustering process is conducted sequentially, point by point, while the labeling of path classes is performed to the entire voyage. Therefore, each voyage has a single path class label.

$$Voyages \in Path\ Classes \tag{2}$$

$$Voyages = \{ts_1[p_1, p_2, \ldots, p_n], ts_2[p_1, p_2, \ldots, p_n], \ldots, ts_j[p_1, p_2, \ldots, p_n]\} \tag{3}$$

$$Path\ Class\ Set = \{class_1, class_2, \ldots, class_k\} \tag{4}$$

where:
$Voyages$: a collection of time series representing the voyage of the vessel taken through a path with a predicted clustered class.
ts_j: a time series of a voyage j a sequence of n data points, where each data point p represents the vessel position and is defined by a pair of coordinates, namely latitude and longitude.
n: a time duration of each voyage, it can be different from voyage to another.
j: a total number voyages.
$Path\ Class\ Set$: a set of k classes into which the path of voyages is being clustered.

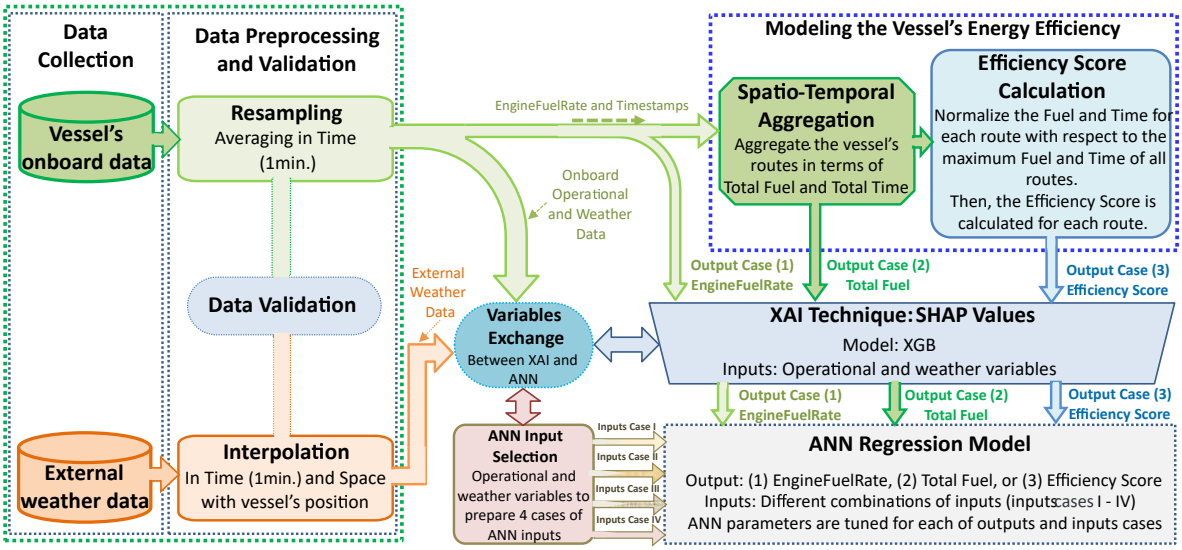


Figure 5: Workflow of modeling and analysis of energy efficiency in short-sea shipping

3.3.2 Distance-Based Method

The similarity between two paths is measured by the average nearest neighbor distance (ANND), as shown in Eq. (5).

$$ANND(i,j) = \frac{1}{n_i}\sum_{k=1}^{n_i} Distance(P_i^k, NN(P_j^k)) \tag{5}$$

where:
$ANND(i,j)$: is the average nearest neighbor distance between path i and path j, present in the distance matrix at row i and column j. It is a symmetric, meaning that $ANND(i,j)$ is the same as $ANND(j,i)$
$Distance(P_i^k, NN(P_j^k))$: The distance between the k^{th} point in path i, denoted as P_i^k, and its corresponding nearest neighbor point in path j, indicated as $NN(P_j^k)$. n_i is the total number of points in path i.
The measure $Distance$ is an Euclidean distance. However, for longer curved routes, Haversine or Great-circle distance is more suitable.

The ANND, as expressed in Eq (5), is computed by averaging the distances between each point in one path and its nearest neighbor in the other path.
Then, the similarity value (i.e., ANND) of this pair of paths is stored as an element in the distance matrix.
A lower ANND indicates that the paths within a cluster are more similar. The distance matrix will have dimensions $(m \times m)$, where m is the number of paths.
For instance, the computed distance matrix for a set of 12 paths is illustrated in Figure 16.

After the construction of distance matrix, machine learning (ML) technique is applied for clustering the paths based on their corresponding values in distance matrix. The ML techniques that we used are K-means, Gaussian Mixture Model (GMM), and Hierarchical clustering method.

3.3.3 Segmented Gaussian Likelihood Method

In addition to identifying the vessel's path, to better understand how the vessel changes its paths, we employ Gaussian distributions on several distinct segments of the vessel route. The

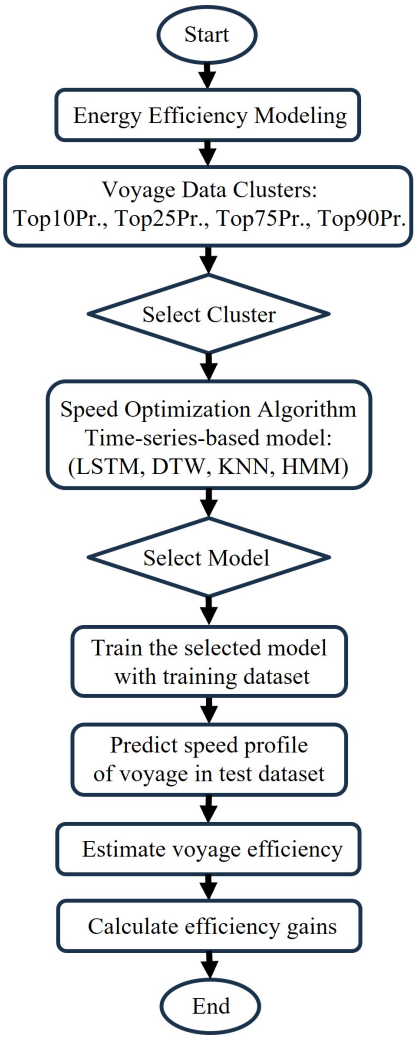


Figure 6: Framework of vessel voyage optimization.

process of this technique can be summarized as follows:

- Utilize a training dataset comprising vessel position information that should adequately represent all potential paths of the vessel route.
- Divide the route into different distinct segments.
- Train a single GMM model for each segment to find the Gaussian distributions of all segments of the route.
- Estimate likelihoods of the segments by using the trained GMM models with their corresponding segments of each vessel voyage in test dataset.
- Label the path classes based on the estimated likelihood at the unique segments of the route.

Figures 21, 22, and 23 illustrate the steps of implementing the segmented Gaussian likelihood method.

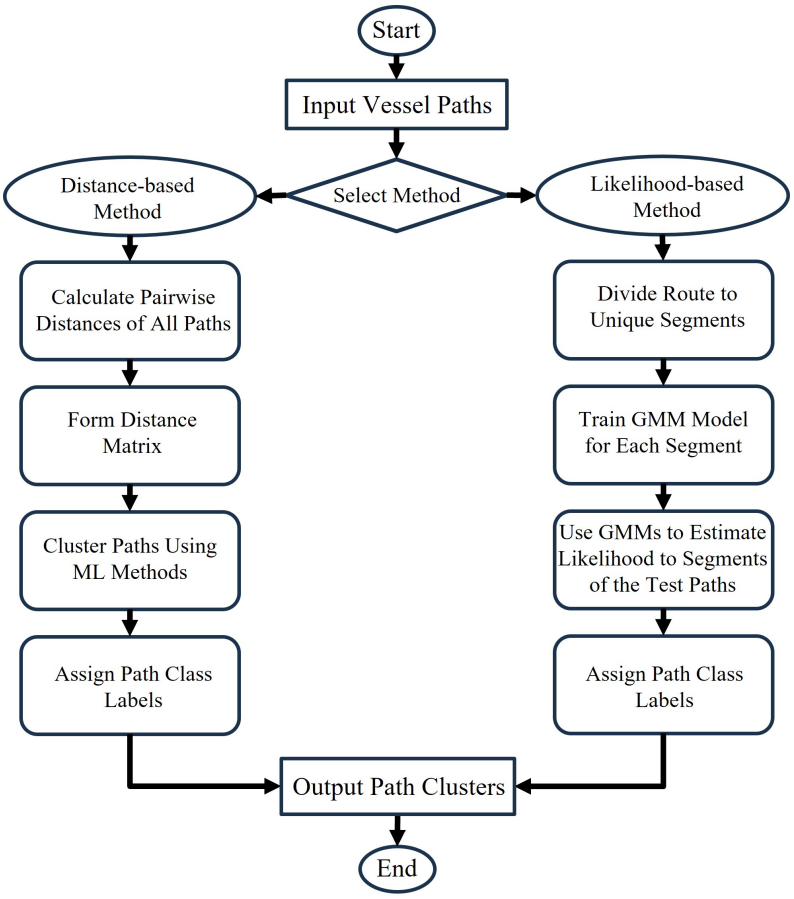


Figure 7: Framework of vessel path identification.

4 Implementation of Frameworks

In this section, we present specific case studies demonstrating how the frameworks of the iHelm project, applied with real-world data, provide practical insights and solutions.

4.1 Case Study of Voyage Efficiency Modeling and Optimization

The implementation of our approach of a time-series analysis-based voyage optimization framework for a fixed-route vessel of our case study is depicted in Figure 6, and the step-by-step process is described by algorithms 1 and 2.
For more detailed information about setting up the models and their specifications including various parameters, you may refer to the source codes, which are developed in Python 3.9.7 to produce the results of this study. These source codes are available at: https://halmstaduniversity.box.com/s/3cuabxcu8l5h2yrt57nj69arkaumw4gq

4.2 Case Study of Path Identification

We also used another dataset with multiple classes of vessel paths for the case study of path identification. This dataset was collected from a ship named Cinderella II, which operates in the Stockholm archipelago. The dataset spans over five months, from July to November 2022, and it comprises information of 124 voyages of this vessel, connecting between two main ports of Vaxholm in the east and Sodra in the west.

Aggregate the datapoints for each voyage to determine the overall path class.
Statistical analysis is conducted to figure out how the vessel paths are different in terms of fuel,

Algorithm 2: Speed Optimization Models for Improving Voyage Efficiency

Data: Refer to **Algorithm 1** for data processing and clustering.
Add $\text{SOG}_{\text{Meas.}}$ and Weathers to Inputs.
for *each* C_k *in CT (sorted by Eff-Score)* **do**
  Set training dataset to voyages $\in C_k$;
  Set test dataset to voyages $\notin C_k$;
  for *each* ***model*** *in [LSTM, DTW, KNN, HMM]* **do**
    Train the **model** with training dataset C_k;
    ; // Skip DTW
    for *each Voyage* v_i *in test dataset* **do**
      if *model is LSTM* **then**
        Recall trained LSTM to predict $\text{SOG}_{\text{Pred.}_i}$.
      else if *model is DTW* **then**
        for *each voyage* v_j *in training dataset cluster* C_k **do**
          Measure the similarity of voyage v_i compared to v_j;
        Set $\text{SOG}_{\text{Pred.}_i}$ to $\text{SOG}_{\text{Meas.}_j}$ of the most similar v_j;
      else if *model is KNN* **then**
        Recall the trained kNN to predict $\text{SOG}_{\text{Pred.}_i}$;
      else if *model is HMM* **then**
        Recall the trained HMM to estimate three weather states.
        if *Weather is Calm* **then**
          Set $\text{SOG}_{\text{Pred.}_i}$ to **max**(SOG_{Calm});
        else if *Weather is Moderate* **then**
          Set $\text{SOG}_{\text{Pred.}_i}$ to **mean**($\text{SOG}_{\text{Moderate}}$);
        else
          Set $\text{SOG}_{\text{Pred.}_i}$ to **min**($\text{SOG}_{\text{Rough}}$)
      Energy Efficiency Estimation:;
      Estimate voyage v_i efficiency of both $\text{SOG}_{\text{Meas.}_i}$ and $\text{SOG}_{\text{Pred.}_i}$.;
    Evaluation Stage:;
    Calculate efficiency gains of test voyages by **model** trained with C_k;

time, distance and speed.
After calculating the distance matrix for the paths, we applied k-means, GMM, and Hierarchical clustering to detect the path classes.

5 Results and Discussion

The vessel has a fixed-route which starts from the southern port to the northern port or vice versa. This route can be divided into four segments, specifically North, Middle, South, and Direct, as depicted in Figure 3a.
Cruising speeds are more common in North, South, and Direct segments of the vessel's route. Meanwhile, in the Middle segment, the vessel typically operates at maneuvering speeds, due to the presence of two ports located on west and east sides of the canal.

5.1 Statistical Analysis

We have first conducted a statistical analysis on the dataset. Figure 8 illustrates some important statistic for all aggregated voyage, with regard to the accumulated fuel, time, and distance at

different route segments. The route segments are depicted in Figure 3a.

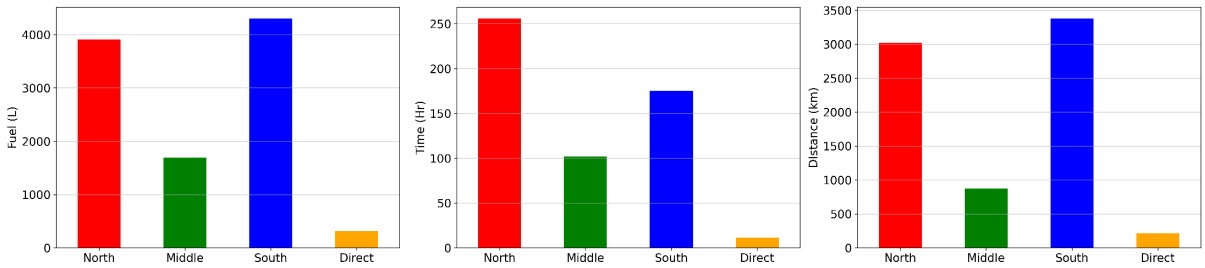


Figure 8: Barplots for statistics of Fuel, Time, and Distance in different route segments

As it can be seen from Table 2, the difference of fuel consumption of the and cruising speeds is 5.47%, so that and also based on the recommendations from domain experts in compliance with maritime regulations including safety and traffic considerations, it might be more practical to primarily focus on optimizing cruising speed.

Table 2: Statistics of the dataset for different speed modes

Variable	Speeds		Difference (%)
	All	Cruising	
Fuel, total (Liter)	1329.2	1256.62	5.47%
Time, total (Hour)	48.04	22.32	53.57%
Distance, total (km)	608.72	349.2	42.68%
Speed, average (m/s)	2.67	1.67	37.5%

5.2 Modeling of Energy Efficiency

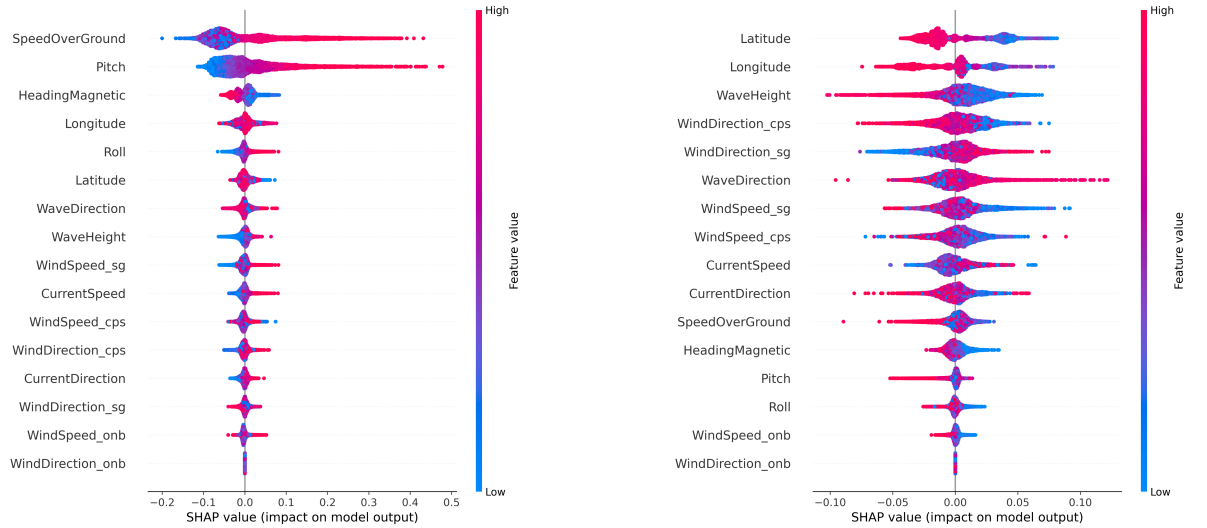
The first step in optimizing the model is to identify the best set of input parameters. We consider four cases of ANN inputs, where each case consists of different combinations of operational and weather variables. Further details about these ANN input cases are provided in Tables 3 and 1.

Table 3: Description of the four input cases of ANN.

Inputs Case	Operational Variables	Weather Variables	
		onboard data	external sources
I	Vessel's location, speed, and direction are used for all cases	wind	—
II		—	wind, wave, and current
III		wind	wave and current
IV		wind	wind, wave, and current

The Beeswarm plot in Figure 9b indicates that the vessel's location has the most significant impact on the Efficiency Score. Therefore, a spatial analysis was conducted to identify the impact of various combinations of operational and weather variables on the Efficiency Score concerning the vessel's location.

The results are shown as heatmaps in Figure 10, revealing that the direct route from south to north or vice versa, located on the open sea, is particularly susceptible to the impact of weather conditions. Thus, for this direct section of vessel routes, the estimation of Efficiency Score, as shown in Figure 10b, has the highest accuracy with different input combinations.



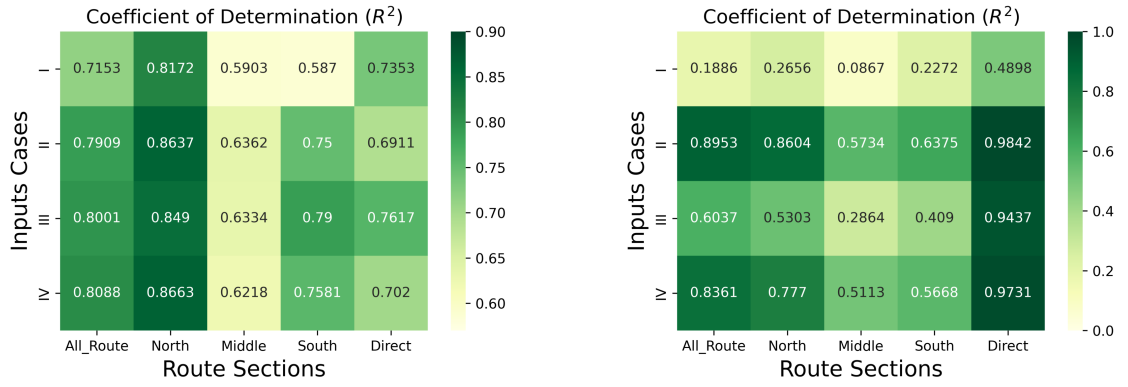
(a) Model's output is EngineFuelRate, (R^2=0.7615)

(b) Model's output is Efficiency Score, (R^2=0.8324)

Figure 9: Beeswarm plots of SHAP values for XGBoost regression model with different outputs

Meanwhile, in the north section, where strong either head or tail wind is more frequent (in this area, west winds dominate) with respect to the vessel route, the Efficiency Score estimation has the second highest accuracy, as in Figure 10b.

In the other case, when it comes to estimating EngineFuelRate, as shown in Figure 10a, the results are not accurate. For instance, the direct sections of the route are not achieving the highest accuracy, even though they are supposed to experience more weather conditions than other sections due to these sections being the most similar to an open sea.



(a) EngineFuelRate

(b) Efficiency Score

Figure 10: Results (R^2) for ANN regression with EngineFuelRate and Efficiency Score as outputs across different input cases in relation to varying vessel's route sections.

5.3 Improving Voyage Efficiency

After that we have implemented the framework for improving the vessel's energy efficiency, as presented in Section 3.2. Then, we evaluate the results, and for a sake of a fair comparison, we are injected both the actual measured and optimized speed profiles into the same estimation model of energy efficiency to predict fuel and time before and after the improving framework of energy efficiency is being implemented. Once the fuel and time estimated, we compute the efficiency to determine how much energy has been saved.

One of our main metrics to evaluate the model performance of voyage efficiency optimization is the gain of efficiency scores, as represented in (6).

$$Eff._{Gain} = \frac{Eff.Score_{Pred.} - Eff.Score_{Meas.}}{Eff.Score_{Meas.}} \times 100 \tag{6}$$

Where $Eff.Score_{Meas.}$ and $Eff.Score_{Pred.}$ represent the voyage efficiency obtained with measured and predicted speed profiles, respectively.

Table 4: Average efficiency gains (Eff. Gains %, see Eq. 6) and counts of improved voyages (Eff. Improved #) out of 162 voyages in the test dataset.

Cluster	Efficiency Metric	LSTM	DTW	KNN	HMM
Top10Pr	Eff. Gains (%)	2.61	3.20	2.13	6.05
	Eff. Improves (#)	134	127	114	139
Top25Pr	Eff. Gains (%)	2.38	3.23	1.58	1.30
	Eff. Improves (#)	129	128	107	107
Top50Pr	Eff. Gains (%)	0.97	2.58	0.98	7.34
	Eff. Improves (#)	100	117	106	140
Top75Pr	Eff. Gains (%)	-0.84	2.28	0.50	9.31
	Eff. Improves (#)	60	119	93	141
Average	Eff. Gains (%)	1.28	2.82	1.30	6.00
	Eff. Improves (#)	105.75	122.75	105.00	131.75

As shown in Table 4, the HMM model achieves the highest average efficiency gain of 6.00%, followed by the DTW model (2.82%), the KNN model (1.30%), and the LSTM model (1.28%). In terms of the number improved voyages out of 162 voyages in test dataset, the HMM model also improves the energy efficiency of the most average number of improved voyages (131.75 out of 162 voyages).

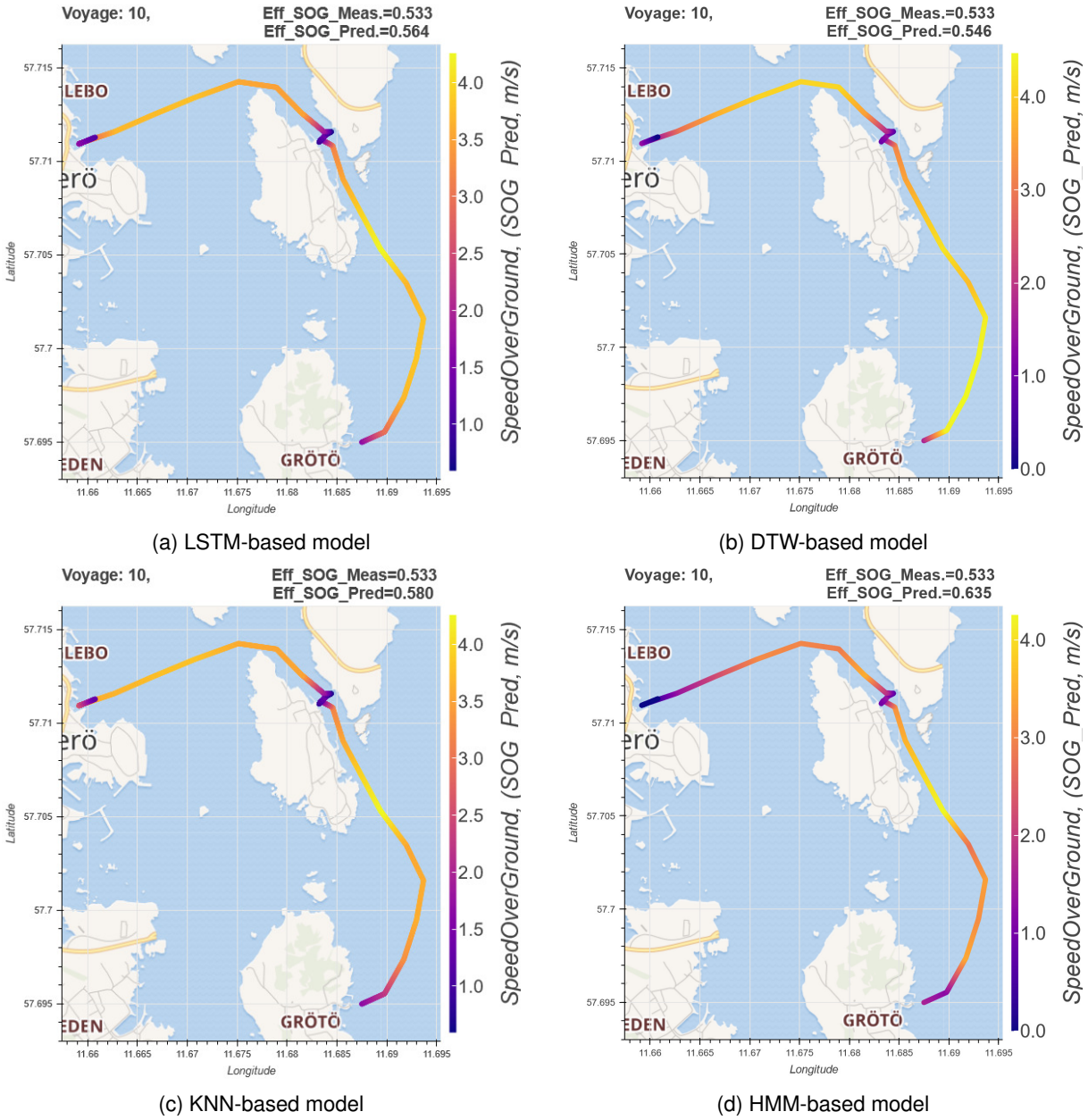
The HMM model achieves its best performance when trained on the Top75Pr cluster, which includes voyages with lower Eff-Scores and frequently encountered adverse weather conditions. Such performance underscores the HMM model's capability to learn the hidden patterns between the vessel speed and weather states, ultimately facilitates for developing more efficient speed profiles.

We also present the results by plots illustrating the predicted speed profiles for a test voyage, which are generated by four time-series based models that incorporate weather data as inputs. These models were trained using data from the Top10Pr cluster.

In summary, the HMM-based model is the most effective model for improving energy efficiency for a vessel voyage in short sea. The HMM model is able to learn the complex relationships between the input features (e.g., speed and weather) and the output feature (Eff-Score), even in different weather conditions.

5.4 Path Identification

As shown Figure 7, we employ k-means, GMM, and Hierarchical clustering methods to identify path classes after computing the distance matrix of the vessel paths.
We evaluated the outcomes of the clustering techniques applied to a dataset comprising 124 voyages through visualization and tabulation, incorporating performance metrics such as precision, recall, and F1-score, along with the confusion matrix.



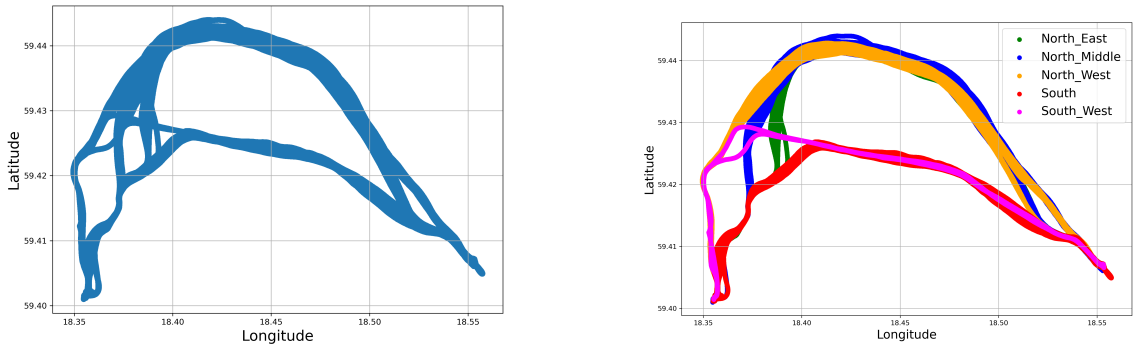
(a) LSTM-based model

(b) DTW-based model

(c) KNN-based model

(d) HMM-based model

Figure 11: Predicted speed profile for a test voyage. From four time-series based models incorporate weather data as inputs and are trained by Top10Pr cluster.



(a) The vessel route. (b) Main classes of path

Figure 12: The vessel route before and after applying path identification

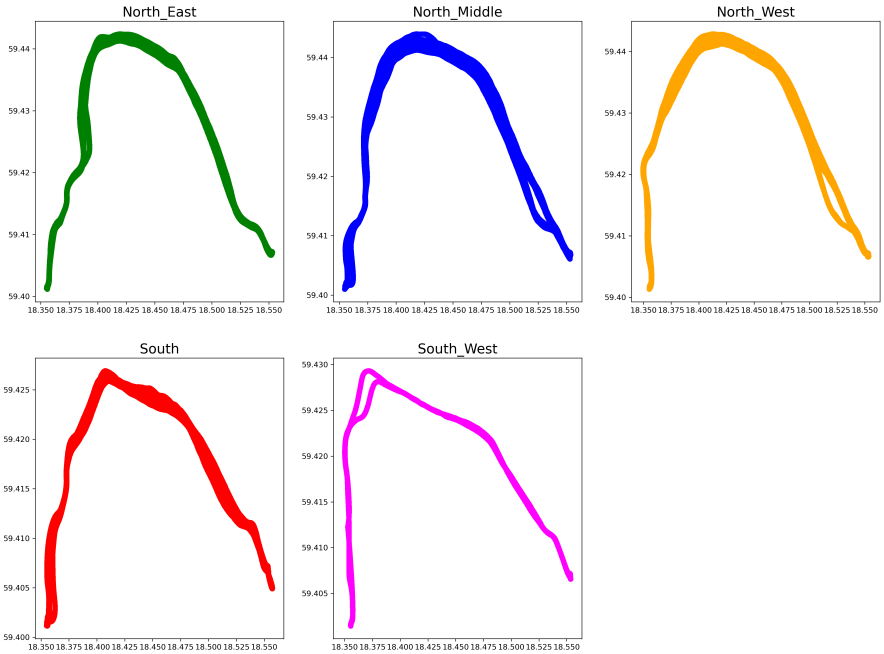


Figure 13: Individual Display of the Five Path Classes.

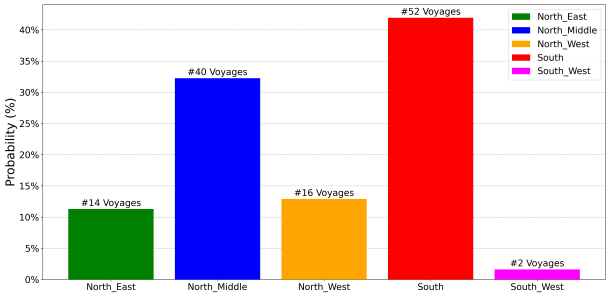


Figure 14: Distribution of voyages across the five path classes.

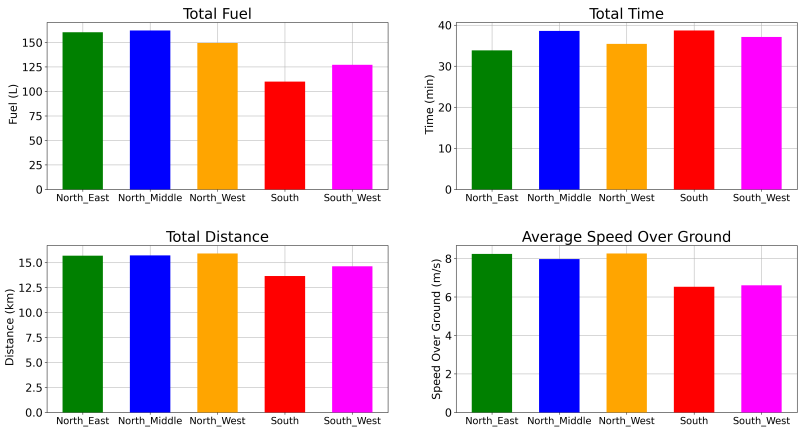


Figure 15: Average fuel and time, distance, and speed of five path classes

	North_East	North_Middle	North_West	South	South_West	North_East	North_Middle	South	South	South_West	North_West	North_Middle
North_East	0	3.42	8.51	11.73	16.42	1.24	2.74	11.51	11.34	15.38	8.46	3.27
North_Middle	3.31	0	6.57	13.13	15.66	4.12	1.03	13.02	12.92	14.26	6.61	0.86
North_West	8.58	6.43	0	18.02	10.1	8.96	6.88	18.07	18.07	10.68	0.39	5.83
South	13.74	16.78	20.75	0	5.99	12.84	16.18	0.47	0.64	4.96	20.6	16.11
South_West	21.37	21.17	14.67	8.6	0	20.12	21.21	8.73	8.9	1.03	14.56	20.43
North_East	1.09	3.96	8.49	9.83	14.36	0	3.31	9.61	9.51	13.32	8.42	3.67
North_Middle	2.57	1.03	6.66	12.97	15.97	3.41	0	12.84	12.75	14.67	6.73	0.98
South	13.09	16.21	20.1	0.45	5.79	12.22	15.62	0	0.43	4.82	19.97	15.6
South	13.61	16.71	20.68	0.64	6.02	12.74	16.11	0.47	0	5.12	20.52	16.06
South_West	19.92	19.91	14.9	7.83	1.03	18.63	19.95	7.95	8.19	0	14.77	19.28
North_West	8.37	6.29	0.4	18.31	10.64	8.67	6.77	18.37	18.31	11.21	0	5.73
North_Middle	3.07	0.87	6.44	12.82	15.69	3.77	0.99	12.7	12.62	14.29	6.53	0

Figure 16: Part of distance matrix, it is 12 paths.

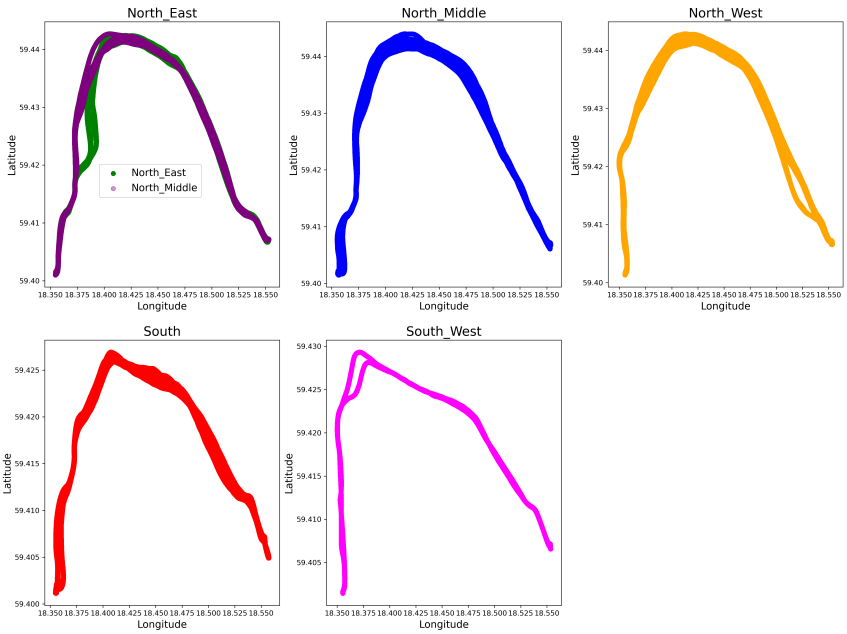


Figure 17: Results of K-means and GMM clustering to five path classes.

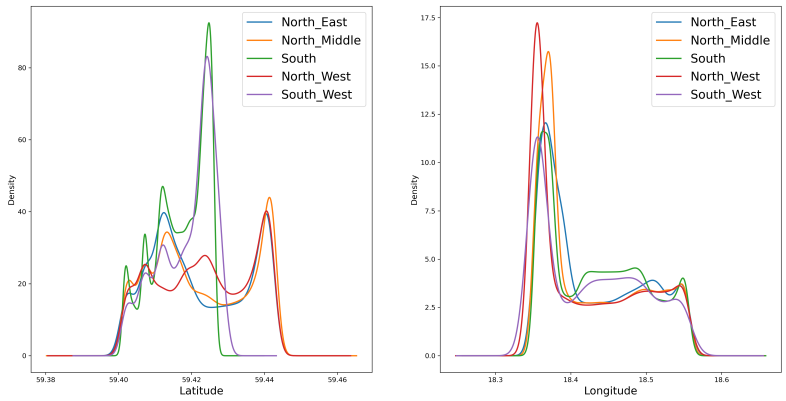


Figure 18: Probability distribution of location coordinates for correctly clustered paths by K-means or GMM.

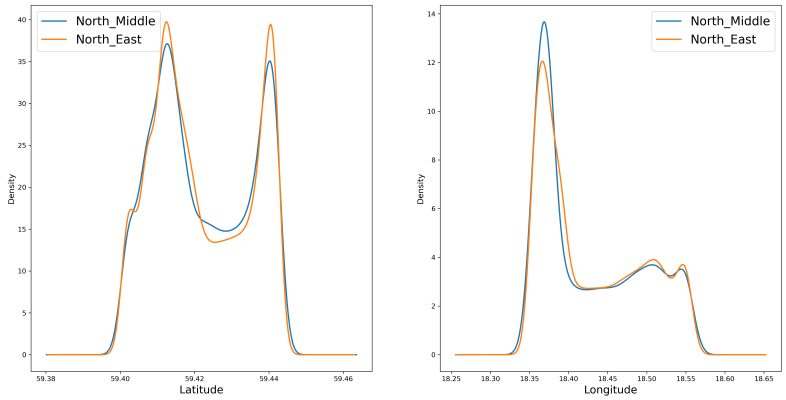


Figure 19: Probability distribution of location coordinates for misclustered paths by K-means or GMM.

Figure 20: Results of Hierarchical clustering to five path classes.

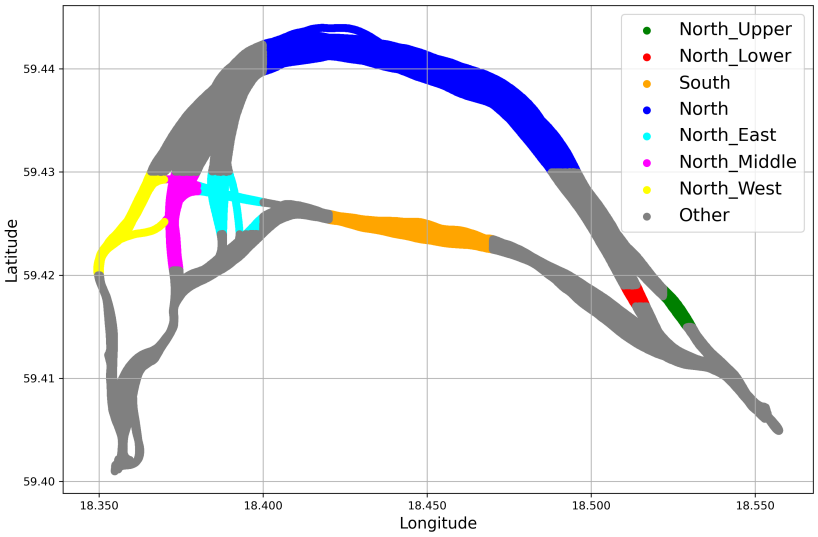


Figure 21: Division of the route into eight segments.

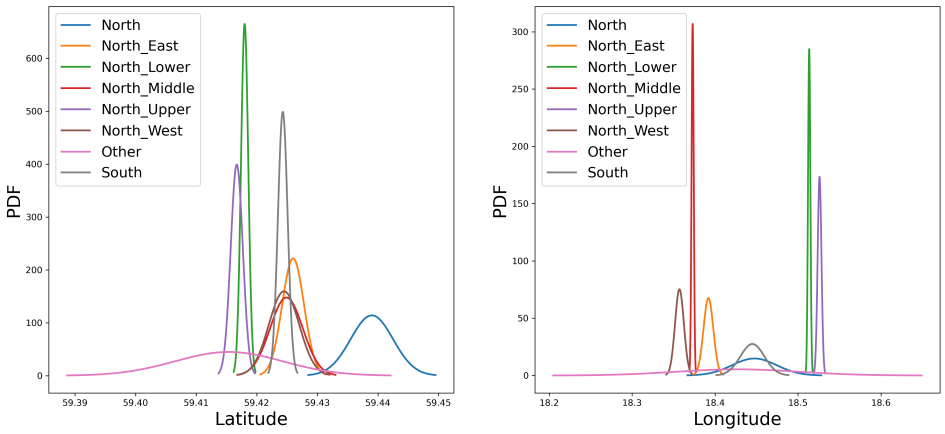


Figure 22: Probability distributions of location coordinates for the eight segments of the route.

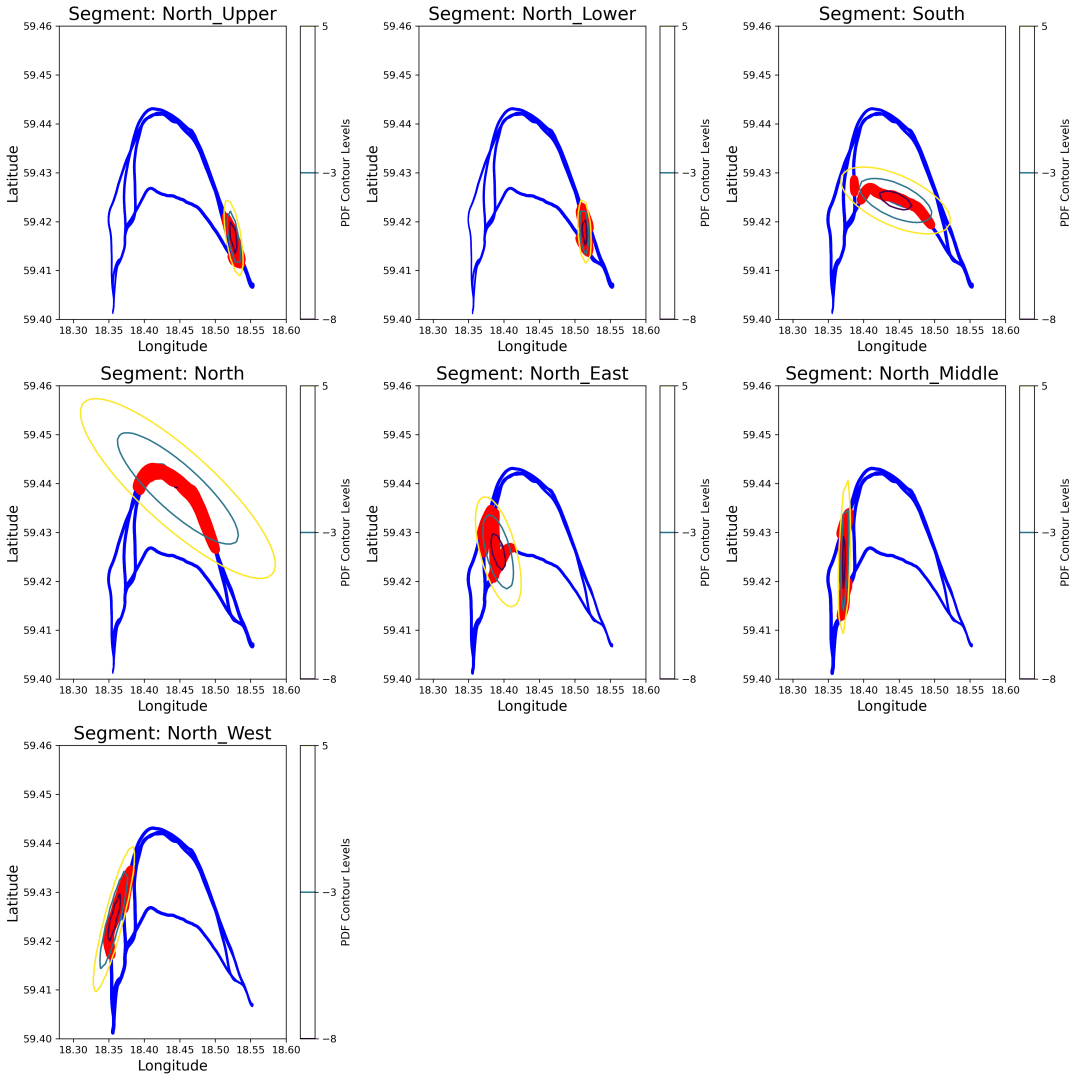


Figure 23: Gaussian distributions for seven segments of the route.

Table 5: Result summary of K-means and GMM clustering to five path classes

Paths	Precision	Recall	F1-score	# Voyages	Confusion Matrix
North-East	0.7	1	0.824	14	14 / 0 / 0 / 0 / 0
North-Middle	1	0.85	0.919	40	6 / 34 / 0 / 0 / 0
North-West	1	1	1	16	0 / 0 / 16 / 0 / 0
South	1	1	1	52	0 / 0 / 0 / 52 / 0
South-West	1	1	1	2	0 / 0 / 0 / 0 / 2

Table 6: Result summary of Hierarchical clustering to five path classes

Paths	Precision	Recall	F1-score	# Voyages	Confusion Matrix
North-East	1	1	1	14	14 / 0 / 0 / 0 / 0
North-Middle	1	1	1	40	0 / 34 / 0 / 0 / 0
North-West	1	1	1	16	0 / 0 / 16 / 0 / 0
South	1	1	1	52	0 / 0 / 0 / 52 / 0
South-West	1	1	1	2	0 / 0 / 0 / 0 / 2

Table 7: Result summary of Gaussian distributions clustering to five path classes

Paths	Precision	Recall	F1-score	# Voyages	Confusion Matrix
North-East	1	1	1	14	14 / 0 / 0 / 0 / 0
North-Middle	1	1	1	40	0 / 34 / 0 / 0 / 0
North-West	1	1	1	16	0 / 0 / 16 / 0 / 0
South	1	1	1	52	0 / 0 / 0 / 52 / 0
South-West	1	1	1	2	0 / 0 / 0 / 0 / 2

The Hierarchical clustering performs better than K-means and GMM clustering.
The clustering by Gaussian distributions also yields profound results, and it might be a promising tool, especially for path clustering at different segments of the vessel route.

6 Conclusion

In this section, we present briefly the key findings of this project and provide recommendations for future research or practical implications for the maritime industry.

6.1 Modeling and Improving Voyage Energy Efficiency

- The modeling approach by using the Efficiency Score, instead of directly working with the EngineFuelRate onboard signal, is more effective in facilitating decision-making.
- The resulting model is based on a more comprehensive understanding of the critical factors that impact fuel consumption, both temporally and spatially, resulting in more dependable counterfactual predictions.
- The quantitative evaluation indicates that estimating the Efficiency Score produces more precise and less biased outcomes than estimating the measured EngineFuelRate.
- The study employs four distinct models: LSTM, DTW, KNN, and HMM, to optimize vessel speed profiles with the objective of enhancing energy efficiency in short sea voyages. The

key observation is that model performance varies significantly across these algorithms. However, the performance of the models varies depending on the data cluster used to train the model and the weather conditions.

- We developed a data-driven framework for optimizing vessel speed profiles to improve energy efficiency in SSS. The framework integrates a data-driven modeling approach to energy efficiency with the DTW algorithm. We evaluated the added value of the framework using a real-world dataset and found that it can effectively improve vessel energy efficiency, especially with limited options, which are common in short-sea shipping.
- DTW exhibits the ability to capture temporal dependencies within speed profiles, especially within the constraints of short-sea shipping where opportunities for actively controlling the vessel to enhance its energy efficiency are restricted
- Although the KNN can handle multivariate data and incorporating additional features like weather conditions, in this case study, the DTW performs better due to its specialized handling of time-dependent data and inherent patterns.
- The result findings emphasize that in terms of searching the best behavior of vessel from the observed data, the DTW exhibits superior performance compared to LSTM and KNN, since the DTW selects the best measured speed profiles. On the other hand, the HMM is the most effective approach in our study. where the HMM optimizes these measured speed profiles further by offering strategies to them informed by their weather states.
- The study also reveals that the HMM model exhibits notable adaptability to different weather states (Calm, Moderate, and Rough). In each weather state, the HMM consistently delivers efficiency gains, indicating its ability to adapt speed profiles according to varying environmental conditions. This adaptability is crucial for real-world maritime applications, where weather can change rapidly.

6.2 Vessel Path Identification

- The approach is able to identify the vessel paths with partially defined or unknown paths.
- In the distance-based method, the hierarchical clustering used in the approach outperforms k-means and GMM clustering techniques.
- The approach of hierarchical clustering includes a user-customized, a cut-off threshold, which allows desired control for the number of path classes, enhancing the flexibility and adaptability of the proposed approach.
- In the distance-based method, adopting ANND as a measure of similarity makes path clustering less affected by noise or outliers and provides a more intuitive interpretation of path similarity, ultimately enhancing the robustness and interpretability of our approach.
- The segmented Gaussian likelihood method is particularly useful for identifying and analyzing the vessel path alterations at different segments of the vessel route.
- The proposed approach is computationally efficient and has the potential to be a valuable tool for planning vessel paths. Accurate path identification can contribute to safer and more efficient maritime transportation practices, aiding in route planning, collision avoidance, and navigation optimization.
- Nevertheless, the framework has some potential limitations, such as the segmented Gaussian likelihood method exhibiting sensitivity to segment definition which could affect its salable performance, particularly in complex maritime scenarios. Moreover, while

the study case demonstrates that the framework is computationally efficient, it is essential to discuss any potential scalability issues, especially when dealing with large datasets, since the computational efficiency may vary depending on the dataset size and the nature of paths.

6.3 Future work and Recommendations

Future work could include the following:

- For modeling of vessel energy efficiency, considering spatial dimension, such as distance variable, might be worth of investigation.
- Including the use of heuristic algorithms such as genetic programming, for optimizing voyage efficiency.
- Exploring the scalability and real-world applicability of the proposed path clustering approach, as well as its integration with related systems of maritime transportation.
- Applying incremental map-matching algorithms for real-time vessel path identification.
- Future research in path detection may involve the study of graph theory and the application of evolutionary algorithms, including ant colony optimization.

Acknowledgment

This research work has been funded and supported by Vinnova. We would like to thank CetaSol AB for their support and for providing the resources necessary to conduct this research.
We also wish to thank the diverse group at the Center for Applied Intelligent Systems Research (CAISR), Halmstad University, for helpful discussions.

Supplementary Materials

The source codes that are implemented on Python 3.9.7 to produce the results are available at: https://halmstaduniversity.box.com/s/3cuabxcu8l5h2yrt57nj69arkaumw4gq

References

[1] CetaSol AB . [Online]. Available: https://cetasol.com.

[2] Copernicus Marine Service . [Online]. Available: https://marine.copernicus.eu.

[3] StormGlass API . [Online]. Available: https://stormglass.io.

[4] Development of short sea shipping. [Online]. Available: https://www7.transportation.gov/testimony/development-short-sea-shipping, February 2007.

[5] Marine Traffic. [Online]. Available: https://www.marinetraffic.com/en/ais/details/ships/shipid:1088282/mmsi:265513810/imo:8602713/vessel:BURO, July 2022.

[6] Mohamed Abuella, M Amine Atoui, Slawomir Nowaczyk, Simon Johansson, and Ethan Faghani. Data-driven explainable artificial intelligence for energy efficiency in short-sea shipping. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 226–241. Springer, 2023.

[7] Jeffrey Dankwa Ampah, Abdulfatah Abdu Yusuf, Sandylove Afrane, Chao Jin, and Haifeng Liu. Reviewing two decades of cleaner alternative marine fuels: Towards imo's decarbonization of the maritime transport sector. *Journal of Cleaner Production*, 320:128871, 2021.

[8] World Bank. *Accelerating Digitalization: Critical Actions to Strengthen the Resilience of the Maritime Supply Chain*. World Bank, Washington, 2020.

[9] John Carlton. *Marine propellers and propulsion*. Butterworth-Heinemann, Oxford, 2nd edition, January 2007.

[10] Patrick Donner and Tafsir Johansson. Sulphur directive, short sea shipping and corporate social responsibility in a eu context. *corporate social responsibility in the maritime industry*, pages 149–166, 2018.

[11] Eurostat. Short sea shipping - country level - gross weight of goods transported to/from main ports. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/mar_sg_am_cw/default/table?lang=en, 2023.

[12] C IMO. Fourth imo ghg study 2020, 2020.

[13] Thalis PV Zis, Harilaos N Psaraftis, and Li Ding. Ship weather routing: A taxonomy and survey. *Ocean Engineering*, 213:107697, 2020.