UNIVERSITY OF
HOHENHEIM

University of Hohenheim

Faculty of Business, Economics and Social Sciences

THAS Report

Automatic Identification System (AIS)

Name:	**Leonard Hegemann**
Matriculation No.:	**1034442**
Study Program:	**B.Sc. Digital Business Management, 4th Semester**
Name:	**James Geter**
Matriculation No.:	**1029952**
Study Program:	**B.Sc. Digital Business Management, 4th Semester**

Course: Introduction to Data Science with R and RStudio

June 26, 2025

1 Introduction

The Automatic Identification System (AIS) is a global tracking system used in the maritime industry to monitor vessel movements and identify ships in real-time. Originally developed to improve navigational safety and avoid collisions, AIS now supports a wide range of applications, including marine traffic management, environmental monitoring, and maritime security.

AIS works by having each vessel equipped with a transponder that regularly broadcasts messages via VHF (very high frequency) radio. These messages are picked up by coastal base receiving stations (terrestrial AIS) and satellites (satellite AIS), enabling coverage both near coastlines and across open oceans. The broadcast rate depends on vessel activity: fast-moving ships send updates every few seconds, while stationary ones may transmit less frequently.[1]

In this project, we analyze AIS data made available through a PostgREST API at `https://aidaho-edu.uni-hohenheim.de/aisdb` with the goal of providing a broad data overview and developing short-term movement forecasts. The API offers two main tables: `/ais_static` for vessel metadata and `/ais_dynamic` for real-time movement. One can also access stored procedures through the `/rpc/` endpoint and retrieve metadata via the OpenAPI description at the root. Using this data, we implement sampling techniques, clean and visualize ship trajectories, evaluate data quality, and develop short-term movement forecasts.

2 Data Overview

2.1 Description of AIS Data Fields

AIS data are structured into two main components: `ais_static` and `ais_dynamic`. The `ais_static` table holds one record per vessel, identified by its MMSI (Maritime Mobile Service Identity) and, when available, a permanent IMO (International Maritime Organization) number. It includes the ship's name, call sign, flag state, and physical attributes such as length, width, draught, and ship type, including both code and label. Additional fields cover the latest known destination, estimated time of arrival (ETA), and the timestamp of the last metadata update.

The `ais_dynamic` table contains time-series data on vessel movement. Each row logs the MMSI, geographic coordinates (latitude and longitude), speed over ground (SOG), and course over ground (COG). It also includes the heading, navigational status, maneuver indicators, rate of turn (ROT), and a flag for positional accuracy. The `collection_type` field indicates whether the message was received via satellite or a coastal base receiver station. Together, these datasets provide a complete view of a ship's identity and behavior over time.

2.2 Descriptive Statistics of the `ais_static` Table

In this section, we examine the structure of the `ais_static` data by quantifying missing values and analyzing the most common entries in key categorical fields. We highlight the dominant ship types and flag states and investigate frequent values in the `destination` field to identify potential issues such as placeholder strings and malformed inputs.

Table 1 summarizes the numeric fields in the `ais_static` table. It shows that the table contains metadata for 220,685 vessels, each uniquely identified by their MMSI. These values vary in the number of digits, which stand for the specific vessel and the type of AIS station.[2] Further, only about half of the records include an IMO number and draught, where draught averages 4.77 meters, including values up to 25.5 meters. The width and the length also exhibit extreme maxima, ranging up to 1022 meters and widths up to 126 meters, significantly beyond typical vessel sizes, suggesting possible data entry errors or anomalous cases.

Variable	**Average**	**Min**	**Max**	**Count**
MMSI	390,688,877.73	2,010,002	775,999,039.0	220,685
IMO	5,283,478.56	0	9,999,999.0	112,961
Draught	4.77	0	25.5	112,961
Ship Type Code	44.10	0	255.0	207,614
Length	61.90	0	1,022.0	207,892
Width	11.41	0	126.0	207,892

Table 1: Numeric Descriptive Statistics in `ais_static` columns

Table 2 presents the top five ship types and flag states by absolute count, highlighting the dominant vessel classes and national registrations. “Other” and “Cargo” together account for over half of all vessels, indicating a strong presence of general-purpose and commercial ships in the dataset. In terms of national registration, China leads with 67,208 flagged ships—out of 229 distinct flag states—followed by the USA, Panama, Norway, and Indonesia. This reflects a highly skewed distribution toward a few dominant flags, likely influenced by trade volume and registration practices, as most merchant ships fly Panama’s flag because foreign owners seek to bypass the stricter maritime regulations of their home countries.[5]

`Ship Type`		`Flag`	
Typ	**Count**	**Flag**	**Count**
Other	58,068	China	67,208
Cargo	52,051	United States	13,756
Fishing Vessel	33,745	Panama	7,638
Tanker	16,548	Norway	7,234
Tug	13,354	Indonesia	6,538

Table 2: Top 5 Ship Types and Flag States by Absolute Count

Table 3 shows the number of missing values per column of the `ais_static` dataset. Several fields like `IMO`, `Draught`, `Destination`, and `ETA` have over 100,000 missing entries, making it hard to use them for analysis. In contrast, key fields like `MMSI` and `Flag` are complete. This means basic identification works well, but for applications like tracking ships or predicting arrival times, the missing data can cause problems and must be handled carefully.

Column	Explanation	Missing Values
`MMSI`	Unique nine-digit vessel identifier	0
`IMO`	Permanent ship identifier	107,724
`Draught`	Vertical hull depth (m)	107,724
`Ship Type Code`	Numeric code for vessel category	13,071
`Length`	Vessel length overall (m)	12,793
`Width`	Vessel width overall (m)	12,793
`Name`	Registered ship name	14,426
`Call Sign`	Maritime radio call sign	45,411
`Flag`	Vessel flag state	0
`Ship Type`	Textual vessel category label	14,875
`Destination`	Reported next port or location	108,967
`ETA`	Estimated time of arrival	131,871
`Static Updated At`	Timestamp of last metadata update	26,809

Table 3: Number of Missing Values in `ais_static`

To evaluate the reporting quality of the `destination` field, we list its five most frequent entries. Table 4 shows that the top two values are long placeholder strings, followed by the only recognizable ports—SHANGHAI and ZHOUSHAN—highlighting the widespread use of defaults or malformed entries in almost 18,000 records combined. This absence of valid destination information prevents port operators, logistics planners, and regulatory agencies from accurately predicting ship arrivals and allocating resources, leading to potential delays, congestion, and increased operational costs.

Count	destination
16,962	@@@@@@@@@@@@@@@@@@@@
676	0@@@@@@@@@@@@@@@@@@@
592	SHANGHAI@@@@@@@@@@@@
459	ZHOUSHAN
398	@

Table 4: Top 5 Values of `destination`

2.3 Descriptive statistics of the `ais_dynamic` table

With 63,337,967 rows, the `ais_dynamic` table is too large to fully analyze the dataset. To overcome this problem, we implemented a sampling routine that selects random 5-minute intervals from the 24th January 2024 and collects up to 100 AIS messages per interval until a 1,000-row sample is reached.

This method corresponds to **simple random sampling (SRS)**: each 5-minute interval has an equal probability of selection. SRS ensures unbiased estimates of population means and an unbiased and straightforward variance estimation. However, it does not account for structural patterns like regional clusters or time-of-day effects. Alternative methods such as *stratified sampling*, which divides the population into predefined subgroups (strata) and samples independently within each stratum to ensure representativeness, or *cluster sampling*, where entire groups (e.g., regions or time blocks) are randomly selected and all or some elements within them are sampled, can improve efficiency when such structural patterns are present.

Descriptive Statistics of Sampled Data

In this section, we summarize the numeric and non-numeric distributions and data completeness of our 1,000-point AIS dynamic sample to identify variability, outliers, and missingness.

Table 5 summarizes basic statistics for numeric fields in the sampled `ais_dynamic` data. Most fields are complete, but variables like `Speed`, `Course`, and `Heading` have fewer than 1,000 observations, and `Status`, `Maneuver`, and `ROT` appear in only 842 cases—suggesting partial transmission gaps. Notably, vessel speeds average 3.62 knots but reach unrealistic values above 100 knots, and heading values exceed the typical 0–360° range, indicating the presence of outliers or encoding errors.

Variable	**Count**	**Mean**	**Standard deviation**	**Min**	**Max**
MMSI	1000	405,485,708.7390	162,602,556.8566	2,442,102.0000	775,628,000.0000
Latitude	1000	20.2225	29.3217	-77.4419	76.6064
Longitude	1000	31.0414	83.2480	-179.4346	179.4666
Speed	960	3.6162	6.1591	0.0000	102.3000
Course	960	200.9089	115.5125	0.0000	360.0000
Heading	960	293.1531	180.7258	0.0000	511.0000
Status	842	2.4299	4.0164	0.0000	15.0000
Maneuver	842	0.0309	0.2318	0.0000	3.0000
ROT(Rate Of Turn)	842	60.8124	76.0818	0.0000	255.0000

Table 5: Numeric Descriptive Statistics of the `ais_dynamic` data sample

To further assess data completeness and categorical patterns in our sampled AIS messages, we now summarize non-numeric fields by their total count, number of unique values, most frequent category, and frequency.

Table 6 shows that timestamps (`Created At`, `Message Timestamp`, `Position Updated At`) are spread over dozens of values but cluster at a few peaks, `Accuracy` is marked FALSE in 557 of 1,000 records, and `dynamic_classA` is the predominant collection source with 637 entries. The prevalence of `Accuracy = FALSE` suggests limited positional precision, reducing reliability for detailed spatial analysis. Moreover, the dominance of `dynamic_classA` indicates a sample bias toward large commercial vessels, underrepresenting smaller ship types.[2]

Variable	Count	Unique	Top	Top Count
Created At	1,000	936	2024-01-24T15:40:37.806+00:00	4
Message Timestamp	1,000	10	2024-01-24T02:58:00+00:00	100
Position Updated At	1,000	51	2024-01-24T03:14:00+00:00	98
Accuracy	1,000	2	FALSE	557
Collection Type	1,000	3	dynamic_classA	637

Table 6: Non-Numeric Descriptive Statistics of the `ais_dynamic` data sample

2.4 Influence of the `collection_type`

The picture changes noticeably when filtering the AIS data to include only points with `collection_type = "satellite"`. Satellite-based AIS captures vessel positions far beyond the reach of coastal receivers, leading to broader spatial coverage—especially over open ocean areas. However, this advantage comes with trade-offs. In congested coastal regions and ports, satellite AIS often provides less detail due to weaker signal quality, transmission collisions, and lower message frequency. As a result, vessel movements in these areas may appear sparse or fragmented. The overall number of observations is also typically lower in satellite data, as it is more susceptible to coverage gaps and interference.[4] These differences, which we can see in Figure 1 and Figure 2, reflect the technical and operational limitations of satellite-based AIS compared to terrestrial systems.

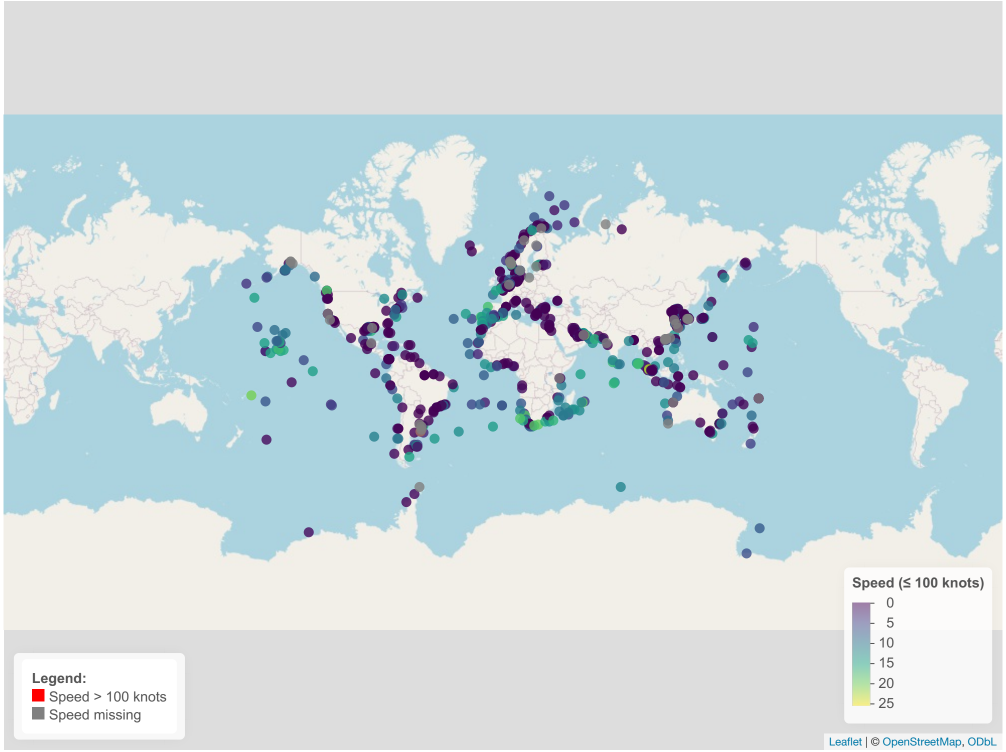

Figure 1: Map of all Sampled AIS Points

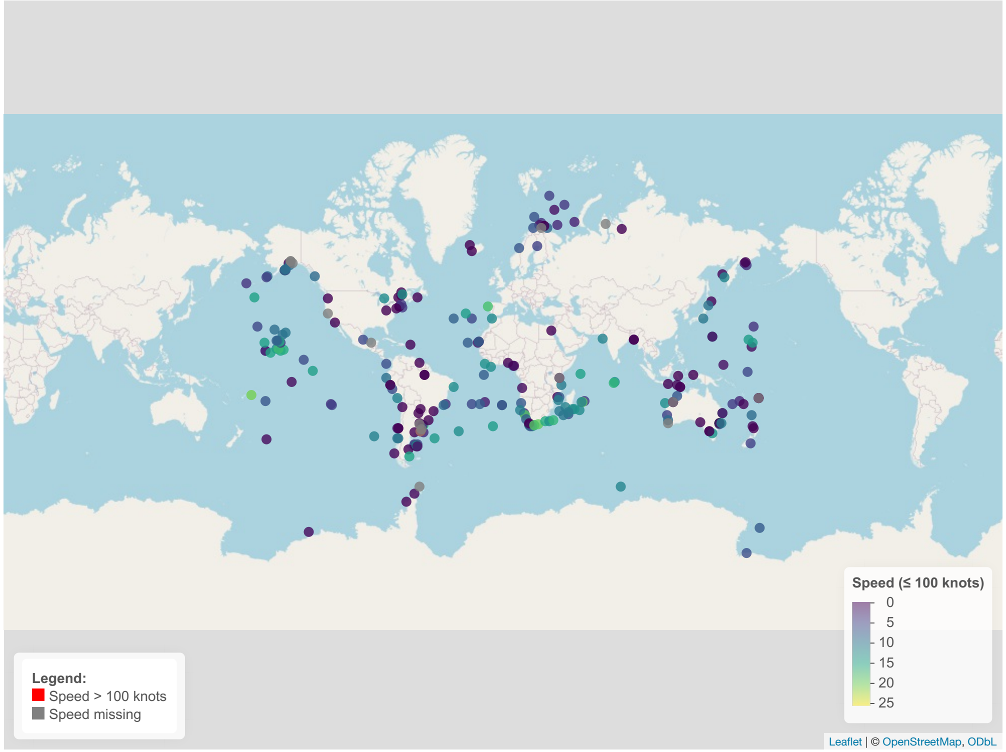

Figure 2: Map of AIS Points with `collection_type = 'satellite'`

2.5 Special MMSI values

The AIS data for MMSI 2579999 shows position points scattered all over the globe, with abrupt jumps between continents in just a few minutes. According to the "Types and classes of AIS – AIS Base Station", MMSI 2579999 corresponds to a shore-based AIS Base Station as it follows the MMSI format '00MIDXXXX'.[2] These base stations, and especially MMSI 2579999 does not transmit actual vessel positions but instead sends special-purpose AIS messages, including test signals and AtoN (Aid to Navigation) broadcasts.[6] Test signals are used to check whether AIS receivers are functioning correctly and are typically logged automatically when a receiver—on land or aboard a satellite—is started or reset. AtoN messages represent navigational aids such as buoys, beacons, or virtual markers intended to assist vessels in safe navigation. Table 7 supports this finding, as one can see that all values except for the flag are missing. Thus, MMSI 2579999 does not stand for a vessel.

MMSI 412420898 is assigned to a real vessel, but its AIS track shows some unusual data patterns. In particular, there are sudden jumps and isolated points that lie far from the main route. These are common issues in AIS data and are usually caused by GPS errors such as multipath effects, poor satellite coverage, or receiver noise. Such errors can briefly produce incorrect position reports, resulting in sharp "spikes" or straight-line connectors in the visualized track. These points do not reflect the vessel's true movement and are considered outliers in the data.[3] Table 7 underlines, that MMSI 412420898 is assigned to a real vessel.

Variable	**2579999**	**412420898**
IMO	NA	0
Name	NA	ZHOU YUAN YU 2601@@@
Call Sign	NA	BZUQ9@@
Flag	Norway	China
Draught	NA	0
Ship Type Code	NA	30
Ship Type	NA	Fishing Vessel
Length	NA	43
Width	NA	8
ETA	NA	2024-01-01T01:01:00+00:00
Destination	NA	A
Static Updated At	NA	2024-01-24T23:29:14+00:00

Table 7: Static Vessel Information from `ais_static` in long format

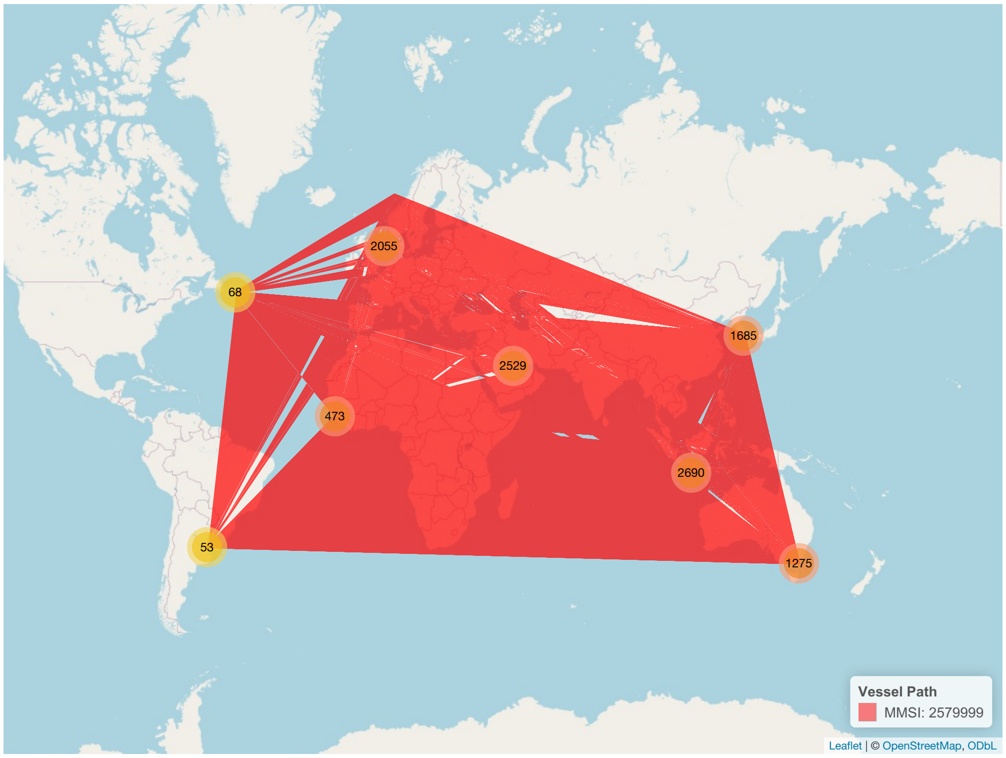

Figure 3: Map of MMSI 2579999 showing position points scattered all over the globe

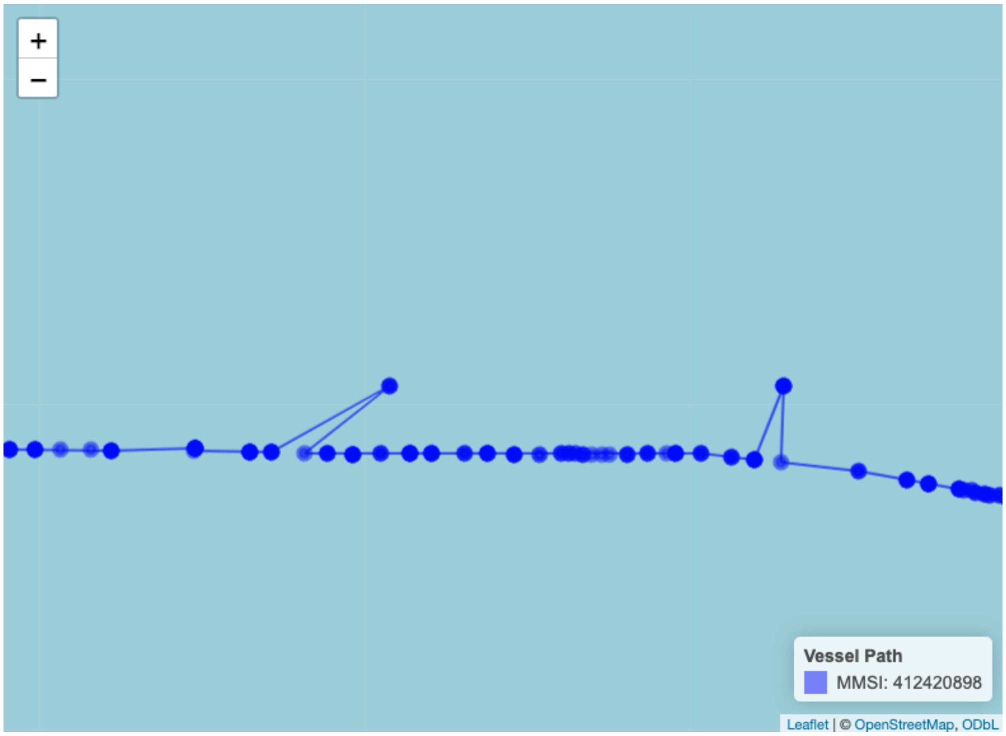

Figure 4: Map of MMSI 412420898 showing errors in GPS data

3 Forecasting

3.1 Predicting Vessel Positions

We developed a simple forecasting model based on the assumption of constant speed and course over a prediction interval Δt. The position projection uses the vessel's current speed v_n (in knots) and course θ_n (in degrees from north) to compute the change in position in nautical miles:

$$\Delta x = v_n \Delta t \sin(\theta_n), \quad \Delta y = v_n \Delta t \cos(\theta_n).$$

These changes are then converted into latitude and longitude degrees and added to the current position (ϕ_n, λ_n):

$$\phi_{n+1} = \phi_n + \frac{\Delta y}{60}, \quad \lambda_{n+1} = \lambda_n + \frac{\Delta x}{60\cos(\phi_n)}.$$

Using AIS data from MMSI 412420898, we filtered and sorted the messages by timestamp, projected future positions, and evaluated their accuracy using the great-circle (Haversine) distance with $R = 3440$ nautical miles, as it gives accurate distances between geographic points, accounting for the Earth's curve:

$$d\left((\phi_i, \lambda_i), (\hat{\phi}_i, \hat{\lambda}_i)\right) = 2R\arcsin\left(\sqrt{\sin^2\left(\frac{\hat{\phi}_i - \phi_i}{2}\right) + \cos(\phi_i)\cos(\hat{\phi}_i)\sin^2\left(\frac{\hat{\lambda}_i - \lambda_i}{2}\right)}\right),$$

Figure 5 shows the calculated predictions in comparison to the real positions.

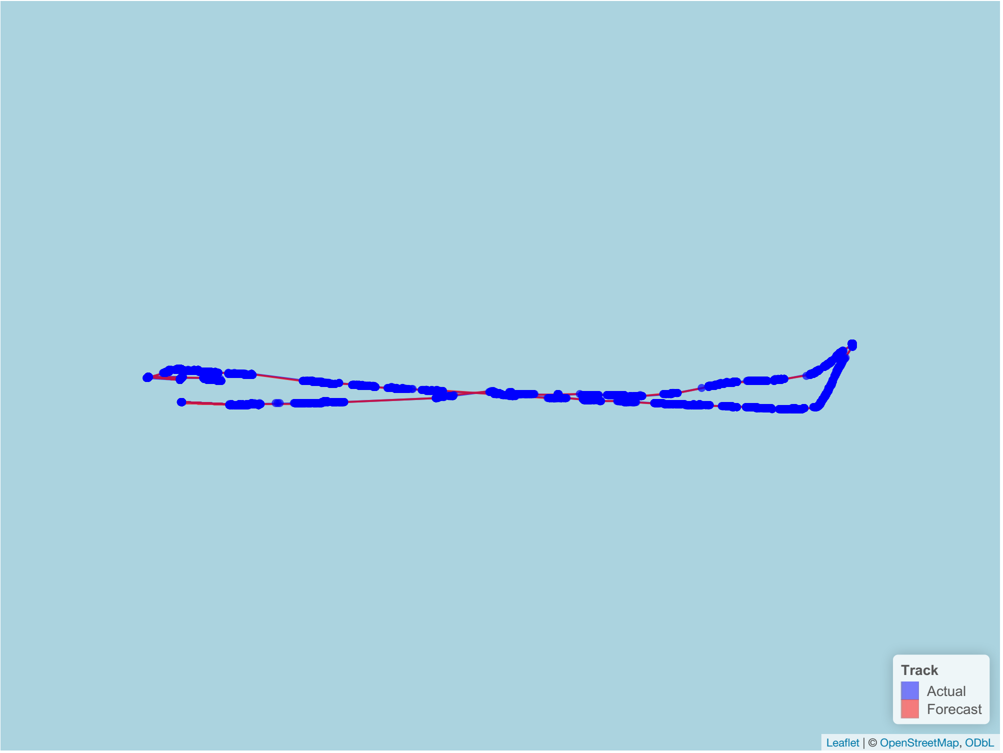

Figure 5: Comparison of Forecasted and Actual Positions for MMSI 41242089 ($\Delta t = 1\,\text{min}$) showing mostly accurate Forecasts

Figure 6 shows that the predictions don't always exactly match the real positions. One can see a slight lag between the predicted positions (red) and the actual positions (blue), which is expected since each forecast is based on the vessel's speed and course from one minute earlier. This assumption of constant motion leads to a small temporal offset, particularly noticeable when the vessel changes speed or direction. The placement of the predicted points consistently ahead of the actual ones from left to right also indicates that the vessel is moving eastward in the image.

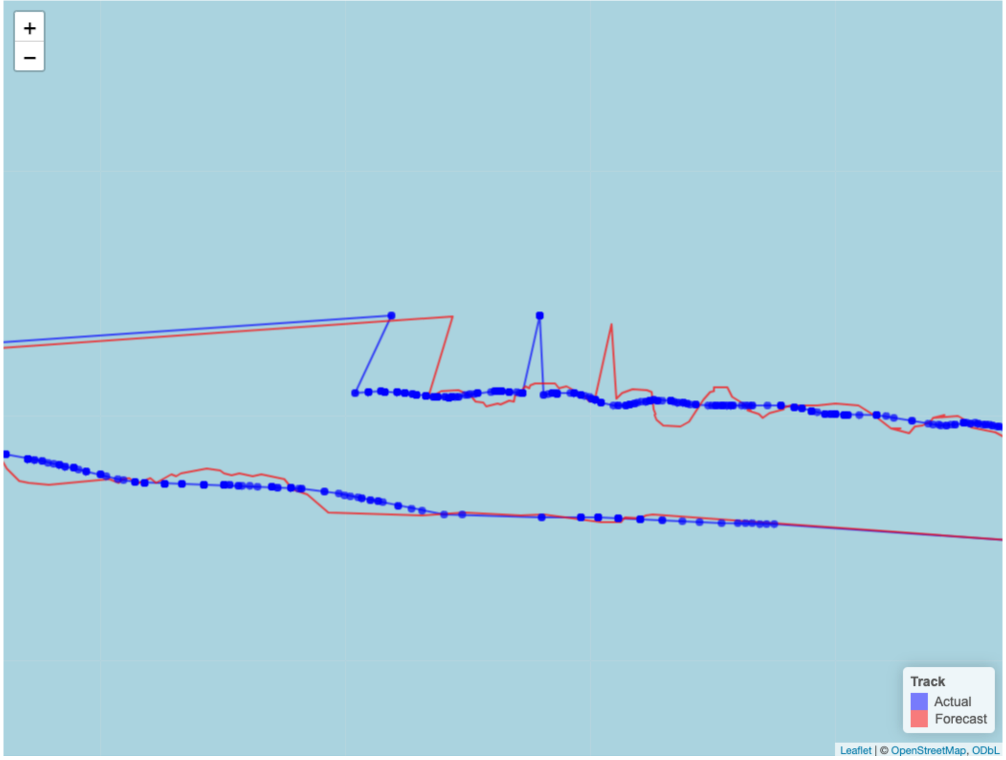

Figure 6: Zoomed in Comparison of Forecasted and Actual Positions for MMSI 412420898 ($\Delta t = 10\,\text{min}$) showing the deviation in Forecasts

The Mean Squared Prediction Error (MSPE) was then calculated as:

$$\text{MSPE} = \frac{1}{N}\sum_{i=1}^{N} d\left((\hat{\phi}_i, \hat{\lambda}_i), (\phi_i, \lambda_i)\right)^2.$$

The MSPE increased with forecast horizon, reaching 0.0465, 0.3677, and 1.1866 NM^2 for $\Delta t = 1$, 5, and 10 minutes, respectively.

As a baseline, we implemented a naïve model that assumes the vessel remains stationary. This approach performed slightly better at shorter intervals and significantly better at 10 minutes, with an MSPE of 0.1874 NM^2.

Raw AIS speed and course readings often contain high-frequency noise as referenced in the previous section 2.5. To mitigate these artifacts and produce more stable inputs for the kinematic projector, we applied Welford's online algorithm. This incremental method computes running means and variances in a single pass—without storing the full history—and is numerically stable even on streaming data.

From a statistical perspective, raw AIS speed and course values can be modeled as noisy observations: $x_t = v_t + \varepsilon_t$, where v_t is the vessel's true value and ε_t represents zero-mean random noise. By applying Welford's online algorithm, we compute a running mean $\bar{x}_t$ that reduces variance in the input data: $\text{Var}(\bar{x}_t) = \sigma^2/n$, compared to σ^2 for any single value. By smoothing speed and course estimates, we reduce spurious fluctuations and better capture the vessel's true recent motion.

The reduced 10-minute MSPE to 1.0587 NM^2 shows exactly this. It is an improvement over the raw kinematic model but still less accurate than the naïve predictor. These results indicate that smoothing can reduce short-term noise but may not adequately capture complex or drifting vessel behavior, where a stationary assumption remains more effective for moderate time horizons.

Repository Access

The full project repository, including code, data samples, and documentation, is available at:
`https://aidaho-edu.uni-hohenheim.de/gitlab/sofop40/AIDAHO_IDS_THAS_2025`

References

[1] Brousseau, Matthew R. (2021). *A Comprehensive Analysis and Novel Methods for On-Purpose AIS Switch-Off Detection.* Master's Thesis, Dalhousie University. `file-service://file-XjFniQmubBh6od3ZbwhD9W`

[2] GMDSS Test Equipment. (o. D.). *Shipborne Automatic Identification System (AIS).* GMDSS Radio Survey Blog. Retrieved June 22, 2025, from `https://gmdsstesters.com/radio-survey/ais/shipborne-automatic-identification-system-ais.html`

[3] Maritime Optima. (n.d.). *AIS and the Main Categories of AIS Challenges.* Retrieved June 22, 2025, from `https://maritimeoptima.com/insights/ais-and-the-main-categories-of-ais-challenges`

[4] Spire Global. (n.d.). *Not All AIS Data Is Equal: How to Identify and Ensure AIS Data Quality.* Retrieved June 22, 2025, from `https://spire.com/blog/maritime/not-all-ais-data-is-equal-how-to-identify-and-ensure-ais-data-quality/`

[5] BBC News. (2014). *Why do so many ships fly the Panama flag?* Retrieved June 23, 2025, from
`https://www.bbc.com/news/world-latin-america-28558480#:~:text=Most%20merchant%20ships%20flying%20Panama's,to%20employ%20cheaper%20foreign%20labour.`

[6] MarineTraffic. *Vessel Details: M02579999 (MMSI: 2579999).*
Retrieved June 23, 2025, from
`https://www.marinetraffic.com/en/ais/details/ships/shipid:23749/mmsi:2579999/imo:0/vessel:M02579999`