Contents lists available at ScienceDirect

Communications in Transportation Research

journal homepage: www.journals.elsevier.com/communications-in-transportation-research

Full Length Article

Validation and correction of auto-logged position measurements

Angelos Ikonomakis[a,b,*], Ulrik Dam Nielsen[b,e], Klaus Kähler Holst[a], Jesper Dietz[c], Roberto Galeazzi[d]

[a] *Maersk R&D, Copenhagen, Denmark*
[b] *DTU Mechanical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark*
[c] *Maersk Line Fleet Performance, Copenhagen, Denmark*
[d] *DTU Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark*
[e] *Centre for Autonomous Marine Operations and Systems, NTNU AMOS, Trondheim, Norway*

ARTICLE INFO

Keywords:
Linear interpolation
GNSS coordinates
AIS coordinates
Measurement error
Sensor accuracy

ABSTRACT

Accurate position measurements are extremely valuable in the shipping industry for various reasons such as safety (collision avoidance), security (situational awareness), fuel-saving (weather identification), punctuality (route prediction), etc. Although GNSS (Global Navigation Satellite System) receivers installed on-board the ships are proven to be highly accurate, the data logging process may occasionally be problematic, mainly due to the complexity of the measurements and the decimal precision that is required. Data were collected from 3 years of operations of 228 Maersk Line container vessels and an analysis reveals that there is a substantial amount ($\approx$ 20%) of historical position measurements sent to shore that does not reflect reality. In the study, the sources of the faulty logged position measurements are categorized and an interpolation methodology is proposed to validate and correct them by using AIS (Automatic Identification System) data.

1. Introduction

1.1. Background

Satellite navigation is a very important asset in modern positioning systems and is the only system that can provide a ship's absolute position relative to the earth's geocentric coordinate system (Chang et al., 2020). Consequently, shipping companies should be extremely cautious about taking good care of their navigation systems in order to avoid losing coordinates data. There is a great number of internal processes in shipping companies that rely on good quality position measurements. The most critical are related to safety and security like for instance collision avoidance (Hu et al., 2007), motion prediction in ports (Shimizu and Pedersen, 2006; Johansen and Fossen, 2016) and motion precision in warships (Núñez et al., 2017). Having correct and reliable GPS coordinates, however, is also critical when it comes to performance evaluation assessing fuel consumption. The operating cost of a ship is mainly influenced by bunker fuel and lubricating oil prices which account for 50%–60% of it (Perera and Guedes Soares, 2017). In a recent paper from Ikonomakis et al. (2021b) it was mentioned that in order to accurately calculate the sea currents' projection on a ship's hull, which can be used as model input on a fusion model of fuel optimization, one needs to be extremely careful with the quality of the position measurements due to plausible mapping mistakes on the metocean grids of the weather providers. By improving the precision in position measurements, performance assessment and planning, including weather routing could be made more accurate leading to a reduction of the fuel consumption.

GNSS (Global Navigation Satellite System) is the standard generic term for satellite navigation systems that provide automated geospatial position with global coverage. This term includes GPS (US), GLONASS (Russia), Galileo (EU), Beidou (China) and other regional systems like QZSS (Japan) and IRNSS or NavIC (India). GNSS is a term used worldwide and its advantage, facilitated by having access to multiple satellite networks, is accuracy, redundancy and availability at all times (Heukelman, 2018; Venezia, 2015). However, there are several identified GNSS error sources and consequences. The work by Karaim et al. (2018) classifies them into (i) signal propagation errors, (ii) clock-related errors, (iii) system errors, (iv) international error sources, (v) user equivalent range error and (vi) dilution of precision. The (i) signal propagation errors refer to those errors caused by the relative shift between the satellite and receiver location at signal transmission time and signal reception time due to the Earth's rotation. The (ii) clock-related errors refer to the

[*] Corresponding author. DTU Mechanical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark.
E-mail address: angelos.ikonomakis@maersk.com (A. Ikonomakis).

https://doi.org/10.1016/j.commtr.2022.100051
Received 14 December 2021; Received in revised form 28 January 2022; Accepted 28 January 2022
Available online 10 March 2022

time-drift errors caused by the clock accuracy between the satellite and the GNSS receiver clocks. The (iii) system errors are caused by the overall nature of the system, e.g., the shape of orbital planes and receiver structure. The (iv) international error sources refer to errors imposed by the service provider or an attack on the system. The (v) user equivalent range error (UERE) is a metric used to quantify the total effect of the remaining errors on pseudorange measurements (Noureldin et al., 2013). The (vi) dilution of precision (DOP) is the metric that evaluates the geometry of visible satellites. The better the geometry is, the lower the DOP.

In shipping, the most frequent measurement loss/modification effects derive from jamming and spoofing which are part of the (iv) international error sources. Jamming is a kind of white noise interference, causing loss of accuracy and potentially loss of positioning (Morong et al., 2019). Spoofing is an intelligent form of interference that fools the GNSS receiver into computing a wrong location (Psiaki and Humphreys, 2016). Various researchers in the past have proposed solutions for jamming (Gao et al., 2016; Medina et al., 2019) and spoofing (Akos, 2012; Broumandan et al., 2012; Schmidt et al., 2016; Fukuda et al., 2021) which correspond to the most frequent a priori data loss in navigation systems (a priori refers to the instance before a position data point reaches the GNSS receiver). Others like Liang et al. (2019) explored the issue by building the missing trajectory using a Random Forest model to identify missing data and an LSTM (Long Short-Term Memory)-based supervised learning method for trajectory reconstruction. In Ryu et al. (2016), the authors focused on improving the accuracy of the position data by integrating INS (Inertial Navigation System) measurements using an EKF (Extended Kalman Filter) and a UKF (Unscented Kalman Filter). In a recent study from Zhang et al. (2021) the authors tried to impute 20 Greenland GPS time series using missForest, which is a new machine learning method for data imputation.

In addition to what is listed above, the authors of this paper have recently identified another problematic issue which refers to the posterior modification/loss of the position data, assuming that the signal has successfully arrived at the GNSS receiver. This makes it more of an internal data processing error than a GNSS error. Apparently, for a position data point to reach the shipping company's data centre at shore it travels through a long path that varies depending on the ship type. This path might transform the format of the position data point several times before reaching the data centre. There are claims of choosing AIS (Automatic Identification System) position data over owned measurements on mathematical models of high precision due to low trust in the latter. This is unreasonable given that sometimes, AIS receives satellite coordinates from the same GNSS antenna as the data monitoring and recording system of the ships.

1.2. *Objective and scientific contribution of study*

This paper intends to initiate an open dialogue on the posterior data loss and modification of position measurements which occurs after the signal has arrived at the GNSS receiver. Firstly, it describes the posterior position measurements path through the data recording system of the ship, leading to shore. This path is proven to be maleficent for the position measurements in about 20% of the vessels tested among the fleet considered. Based on a literature review, it is believed that no study to date has examined the posterior data loss/modification of position measurements. In the second part, the paper introduces ways of identifying and correcting faulty position observations both in vessel-specific (real-time data corrections) and in shore-specific (historical data correction) modes. It should be acknowledged that an earlier version of this work was presented at HullPiC conference in June 2021 at Gubbio, Italy (Ikonomakis et al., 2021a).

2. Data

The study includes data recordings from 228 container ships that have been collected over a three-year period (2017–2020) during operations in the majority of the world's larger oceans. The data is divided into two categories; the CAMS (Control Alarm Monitoring System) dataset which consists of ≈ 25 million rows of sensor data recordings stored in the company's database, and the AIS dataset which is sourced from two external providers consisting of ≈ 50 million rows of data. An overview of the datasets is given in Table 1.

It should be noted that the reason for sourcing and merging the AIS dataset from two providers was to get as many valid points as possible to maximize the average frequency. Thus, the final AIS dataset consists of three measurement types (i) terrestrial, (ii) satellite, and (iii) dynamic. In terrestrial, the data is broadcast on a common international VHF frequency. In satellite, the data is received through the satellite navigation network. Finally, for heavy traffic regions such as the “South China Sea” or the “English Channel”, AIS signals collide resulting in position detection failures and inaccurate reporting. The dynamic measurement type solves this issue.

As a supplementary note, it can be mentioned that the same dataset, although slightly smaller (189 ships), has also been used to investigate what sea conditions ships *really* encounter, see Nielsen and Ikonomakis (2021), noticing that all corrupted position data were excluded.

3. Problem formulation

3.1. *Problem source*

The position signal that reaches the GNSS receiver is transformed into an NMEA sentence. NMEA is an acronym for the National Marine Electronics Association. Today, NMEA is a standard data format supported by all GNSS manufacturers. Particularly, the NMEA sentence is in printable ASCII form and may include information such as time, position, speed, water depth, etc. (NMEA, 2021). An example of an NMEA sentence is shown below.

$GPGGA,181908.00,3404.7041778,N,07044.3966270.
W,4,13,1.00,495.144,M,29.200,M,0.10,0000*40.

All NMEA sentences start with the $ character, and each data field is separated by a comma. GP stands for GPS position (e.g., GL would denote GLONASS). The next value 181908.00 is the timestamp (UTC time in hours, minutes and seconds) followed by 3404.7041778, the latitude in the DDMM.MMMMM format. Here it should be mentioned that decimal places are variable. N denotes north latitude. 07044.3966270 is the longitude again in DDDMM.MMMMM format and W denotes west longitude (Gakstatter, 2015). The rest of the values will not be interpreted due to irrelevance to this study.

Table 1
Type, description, frequency, value range, median and units for CAMS and AIS datasets.

Type	Description	Sampling Time	Range	Median	Unit
CAMS	Time	10 min	01/01/2 017–22/02/2 020	—	(UTC datetime)
	ImoNo	10 min	228 unique vessels	—	(-)
	Latitude	10 min	−49–61	25.024 7	(°)
	Longitude	10 min	−180–180	23.732 8	(°)
	SOG	10 min	0–26	12.6	(kn)
AIS	Time	Uneven	01/01/2 017–22/02/2 020	—	(UTC datetime)
	ImoNo	Uneven	228 unique vessels	—	(-)
	Latitude	Uneven	−50–61	32.126 4	(°)
	Longitude	Uneven	−180–180	9.918 8	(°)
	SOG	Uneven	0–25	12.3	(kn)
	COG	Uneven	0–360	188	(°)

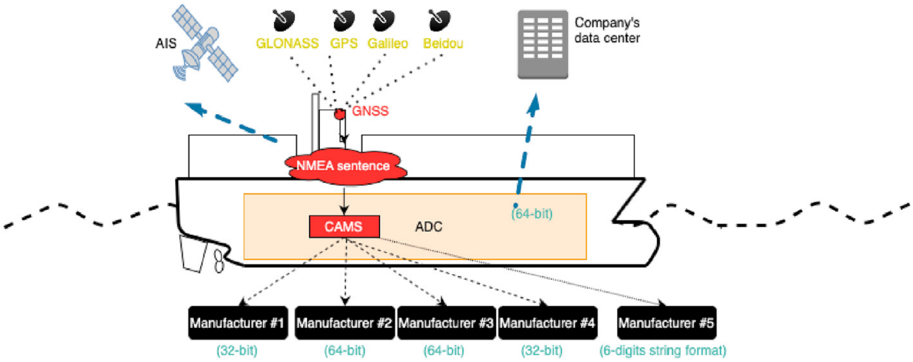


Fig. 1. Position measurement path on board a vessel before reaching shore. The bit-rate format is indicated for each of the 5 CAMS manufacturers.

This NMEA sentence is initially received by the AIS transceiver. AIS is an automatic tracking system that is used by vessel traffic services (VTS) supplementing the marine radar for collision avoidance. AIS transceivers can be tracked by AIS base stations located along coastlines or, when out of range of terrestrial networks, through a growing number of satellites (Contributors, 2021a). According to the literature, the AIS transmission rate is relative to the ship's speed and ranges from 5 to 180 s (ITU, 2014).

In order for the data deriving from multiple sensors to be monitored and stored, Maersk is using a data processing system called the ADC (Auto Data Collector) which is the source from where the data is sent to shore. Within ADC there is a system called CAMS. CAMS is responsible for connecting the sensors, normalizing and converting each data point (NMEA sentences), aggregating it into either 1 s or 10 min and later logging it and transmitting it to multiple internal services. Whenever there is a good internet connection, ADC sends the 10-min aggregates to shore. Depending on the vessel class, CAMS is bought from a list of manufacturers, each with distinct characteristics. The main characteristic that distinguishes them is the encoding/decoding memory format. They are either 32-bit or 64-bit. In some unique cases, the memory format is even lower, accepting only 6-digit numbers. The memory format of ADC is 64-bit. While converting among memory formats, sometimes data quality is degraded significantly (Haithcoat, 1999). Locations closer to the Equator are more sensitive due to the oval shape of the earth and need to have at least 4 valid degree decimal points to get a precision within 10m. Limiting the memory format to 6-digit numbers, automatically increases the precision within 100m. Fig. 1 illustrates the path a data point follows before reaching the company's data centre along with the bit-rate format for each of the 5 CAMS manufacturers.

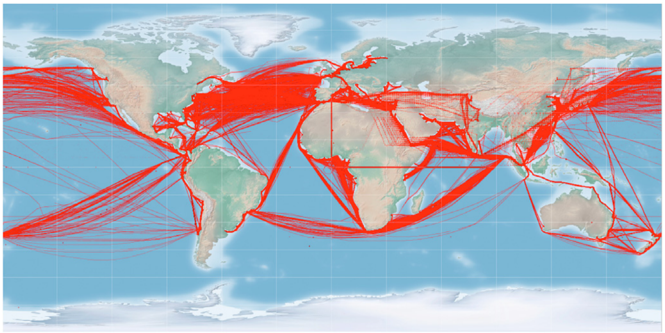


Fig. 2. Position measurements of 228 container ships during 3 years of operation (2017–2020).

Before decomposing the various sources that degrade the position data, a map with plotted trajectories from all the vessels of the dataset illustrates the gravity of the problem. Fig. 2 shows the coordinates registered from 228 container ships during 3 years of operation (2017–2020). It is evident that there are multiple measurements registered in locations away from a common vessel's trajectory.

A whole range of different sources degrading the ship's position data were identified. The most prominent are:

N/E issue: It refers to when the longitude values do not turn to negative when the vessel crosses the prime meridian towards the western hemisphere. As a result, location measurements get packed on the upper right quadrant of the map where both longitude and latitude values are positive. Fig. 3 illustrates the registered trajectory of a vessel with a CAMS system experiencing such a problem. The issue is present in classes where CAMS manufacturers #2 and #3 are installed. In this case, the error is immeasurable since the measurements could drift for thousands of kilometres.

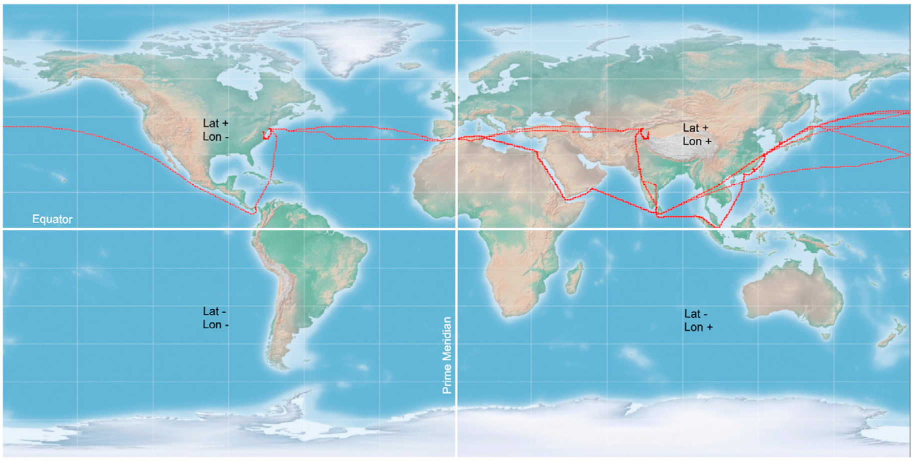


Fig. 3. Registered trajectory of a vessel with “N/E issue”.

Timestamp	Latitude[°]	Longitude[°]
2020-05-22 04:00:00	7,087135	76,308050
2020-05-22 04:10:00	7,074442	76,337243
2020-05-22 04:20:00	7,063569	76,366003
2020-05-22 04:30:00	7,052612	76,395564
2020-05-22 04:40:00	7,041450	76,425406
2020-05-22 04:50:00	7,030998	76,454078
2020-05-22 05:00:00	7,020308	76,483448
2020-05-22 05:10:00	7,009499	76,513218
2020-05-22 05:20:00	6,598878	76,541891
2020-05-22 05:30:00	6,587945	76,570984
2020-05-22 05:40:00	6,576370	77,000341
2020-05-22 05:50:00	6,564854	77,028670

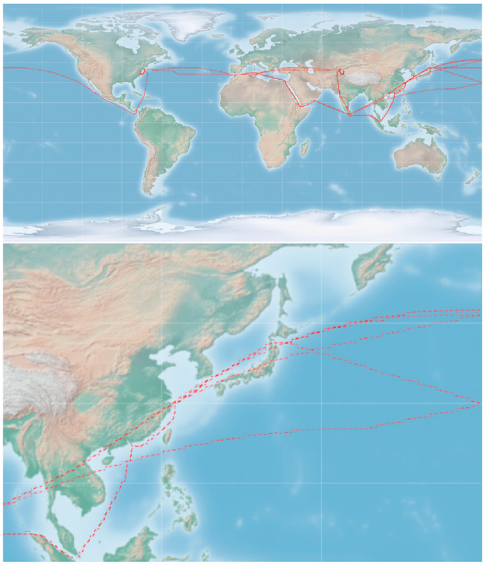


Fig. 4. Vessel with "zig-zag issue" on stored position data. On the left, there is a data table indicating the "jumps" in red colour. On the right, there are two maps; below is a zoomed version of the one above. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Zig-Zag issue: It refers to when either longitude or latitude values go beyond 0.599999. In this case, the next step is 1.000000 based on the logic that 1° is made up of 60 min. That is obviously not true, as it should simply be 0.600000, followed by 0.600001. Every time this happens we see a "jump" of 0.4° of either latitude or longitude, which relates to ≈ 40 km of error. Fig. 4 illustrates the location of the vessel on the right part with a zoomed version on the bottom right to have a clearer view of the zig-zagging effect. On the left, there is a data table that indicates the "jumps" in red colour. Here, we should mention that the previous source "N/E issue" is also visible. The issue is present again in classes where CAMS manufacturers #2 and #3 are installed.

Scatter issue: It refers to when there are plenty of random iterations of either longitude or latitude scattered around the globe away from the regular vessel's trajectory. Fig. 5 illustrates the registered trajectory of the vessel with "scatter issue". Source "N/E issue" is again visible in this vessel's trajectory. The issue is present in vessel classes where CAMS manufacturer #2 is installed. The error for this issue is immeasurable since the scattering on the investigated vessels was not within a specific threshold from the real trajectory.

Drift issue: It refers to when parts of the vessel's trajectory are shifted a few degrees towards either East/West or North/South affecting both longitude and latitude values. Fig. 6 illustrates the registered trajectory of the vessel with "drift issue". Source "N/E issue" is again on top, given that the vessels are using the same CAMS system. The issue is present again in classes where CAMS manufacturer #2 CAMS is installed. In this case, the error can be assigned as immeasurable since the variance of the investigated vessels on the locations with prominent drifting was up to 5000 km.

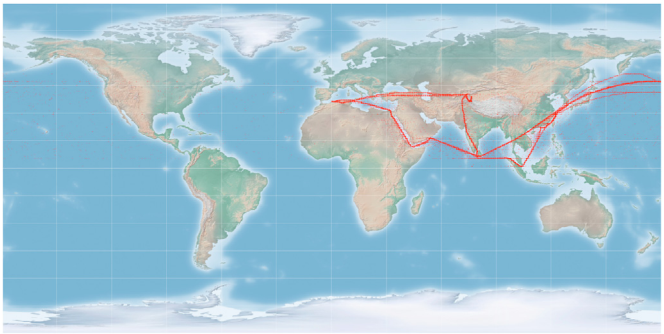


Fig. 5. Registered trajectory of a vessel with "scatter issue".

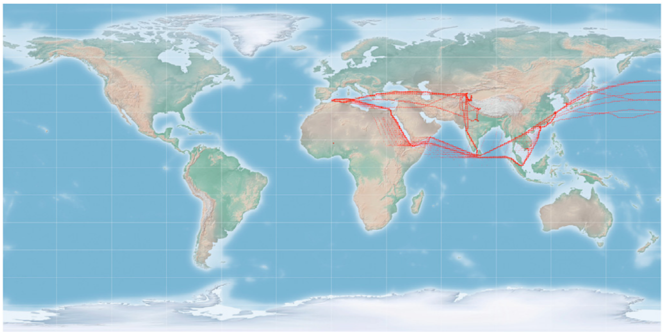


Fig. 6. Registered trajectory of a vessel with "drift issue".

Frozen issue: It refers to when either longitude or latitude values are frozen to 0°. Fig. 7 illustrates the registered trajectory of the vessel with "frozen issue". The issue is present in classes where CAMS manufacturer #4 CAMS system is installed. The measurement error is immeasurable for this case.

Bit-rate issue: Besides the visible issues that were described and showed on the above maps, CAMS manufacturers #1, #4 and #5 experience the bit-rate conversion degradation on the position data as described earlier in the beginning of the subsection. In Fig. 8, there is a graphical representation of the issue explaining either of the two cases that might cause the "Bit-rate issue" given the installed equipment. In this case, the measurement error could scale up to 100 m.

In this project, 6 issues were identified in total. It is believed that these were the most obviously distinguishable issues and that there might be more either in this or in another company's dataset. In the next subsection, the authors introduce the indicators built to identify faulty position measurements in any dataset, regardless of the underlying issue.

3.2. Validation indicators

Based on Gakstatter (2015) the SOG (Speed Over Ground) is included in the NMEA sentence which means that the SOG is computed using

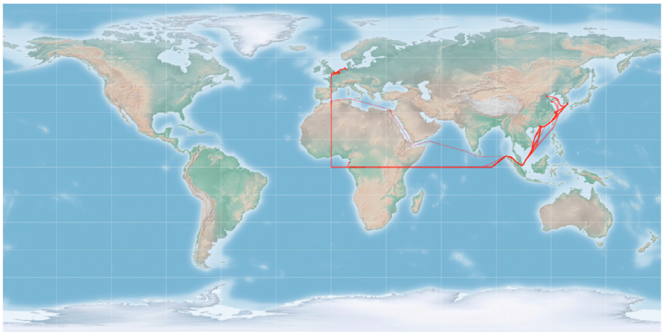


Fig. 7. Registered trajectory of a vessel with “frozen issue”.

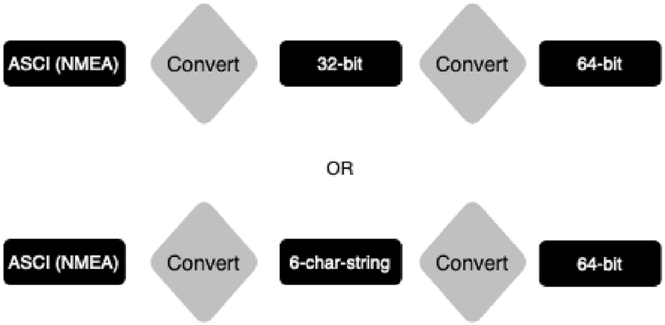


Fig. 8. Graphical representation of the “bit-rate issue”.

internal estimation processes by the navigation devices. It is also known that AIS and CAMS systems record values of SOG from −10 to 50 knots (with 99.9% of our dataset being positive values). On top of that, SOG is registered with a single decimal point which makes it less vulnerable to accidental measurement degradation (averaging, bit-rate conversion, etc.). Taking into account the previous statements and due to speed's physical relationship with position, SOG makes an ideal measurement for comparison purposes. Thus, by computing the distance a vessel has travelled between consecutive data logging/transmission (10 min between one another) with two different methodologies (one using the position measurements and one using the SOG), the two outputs should theoretically match. Based on the latter the following distance, deviation indicators were designed to validate that assumption:

$$D^c = d_p^c - d_s^c, \tag{1}$$

$$D^a = d_p^a - d_s^a. \tag{2}$$

Equations (1) and (2) describe how the distance deviation indicators D^c and D^a are composed. The superscripts $\{c\}$ and $\{a\}$ stand for CAMS and AIS, respectively. On the right part of the equations, d is distance and the subscripts $\{p\}$ and $\{s\}$ denote the two distance calculation methodologies using position measurements and SOG, respectively. In short, in Equation (1) CAMS data is used to compute the distance deviation D^c which is calculated by subtracting the distance between consecutive coordinates using the SOG methodology d_s^c from the same distance using the position measurements methodology d_p^c. In Equation (2), there is the same computation, but by using the AIS data.

For the position measurement methodology, given that the time between consecutive iterations is not longer than 10 min based on Table 1, the distance between two points is computed using the haversine formula (Contributors, 2021b), assuming that the earth is a perfect sphere with a radius $R = 6378.2$ km. For the SOG methodology, the travelled distance is computed as the integral of the SOG within the time period of 10 min.

After having introduced that both D^c and D^a are the main indicators of this study, there is a third one utilized as a safety indicator. This is d_p^b denoting the distance between AIS and CAMS coordinates. The subscript $\{p\}$ indicates the position measurement methodology and the superscript $\{b\}$ stands for “between”. Fig. 9 shows how d_p^c, d_s^c, d_p^a, and d_s^a distinguish from d_p^b.

In Figs. 10 and 11, one can see the IQRs of both D^c and D^a indicators categorized by encoded ship names V_{imo} and classes V_{class}. Each of the 228 boxes indicated by I.001, I.002, …,I.228 in the y-minor-axes represents the IQR of either D^c or D^a of each ship of the dataset. The boxes are categorized by 32 ship classes shown in different colour indicated in the y-major-axes by C.01, C.02, …, C.32. The grey dots represents data outliers for values outside 99.3% of the distribution. The requirements for a ship's position measurements to be of good quality are the following:

Requirement 1 *The mean value of $D^{c,a}$ distribution should be close to 0 m with small variance.*

Requirement 2 *Based on internal company research and considering the results from the boxplots in Figs. 10 and 11, it has been decided that the IQR (interquartile range) of the $D^{c,a}$ of each ship should not exceed 300 m, given that IQR represents the 50% of the distribution. The 300 m limit is a parameter based on the dataset of this study. It should be adjusted according to each company's precision requirements.*

Starting from Fig. 10 and having in mind the above requirements, it is evident from the boxplots of classes such as C.03, C.07, C.16 which have ships with very narrow boxplots and with mean values close to 0, that these are the ships with good quality position CAMS measurements on the database. Classes like C.10, C.13, C.18, C.30 and C.32 with medium size boxplots close to the 300 m threshold are questionable and need further investigation. Finally, position measurements from classes like C.02, C.06, C.11, C.12 and C.17 (only a few ships) are clearly problematic and require replacement. By comparing the results of D^c and D^a with Fig. 11, it must be pointed out that the AIS dataset is indisputable but not better than the good quality boxplots of some classes of D_c. Nonetheless, this does not mean that the faulty CAMS measurements can be blindly replaced with their AIS equivalents mainly because the data logging in AIS is not registered in even time intervals like in CAMS. To better understand the difference between D_c and D_a, Fig. 12 shows the distributions by class for both indicators. The classes on D_a are normally distributed with similar standard deviation among each other. In the contrary, those of D_c are normally distributed only for a few classes, not mentioning the standard deviation which are different in almost all of them.

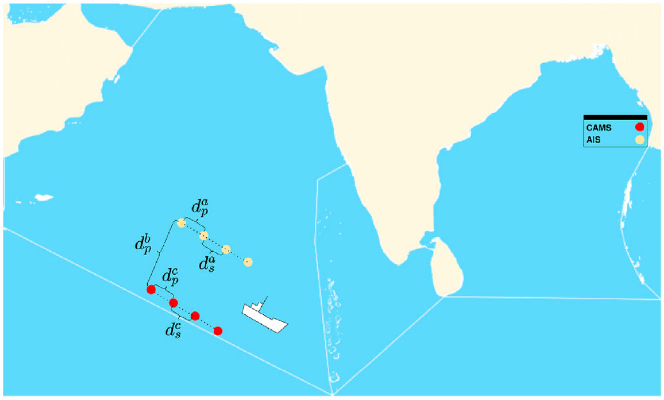


Fig. 9. Sample trajectory with position measurements intervals and the main elements of the validation indicators.

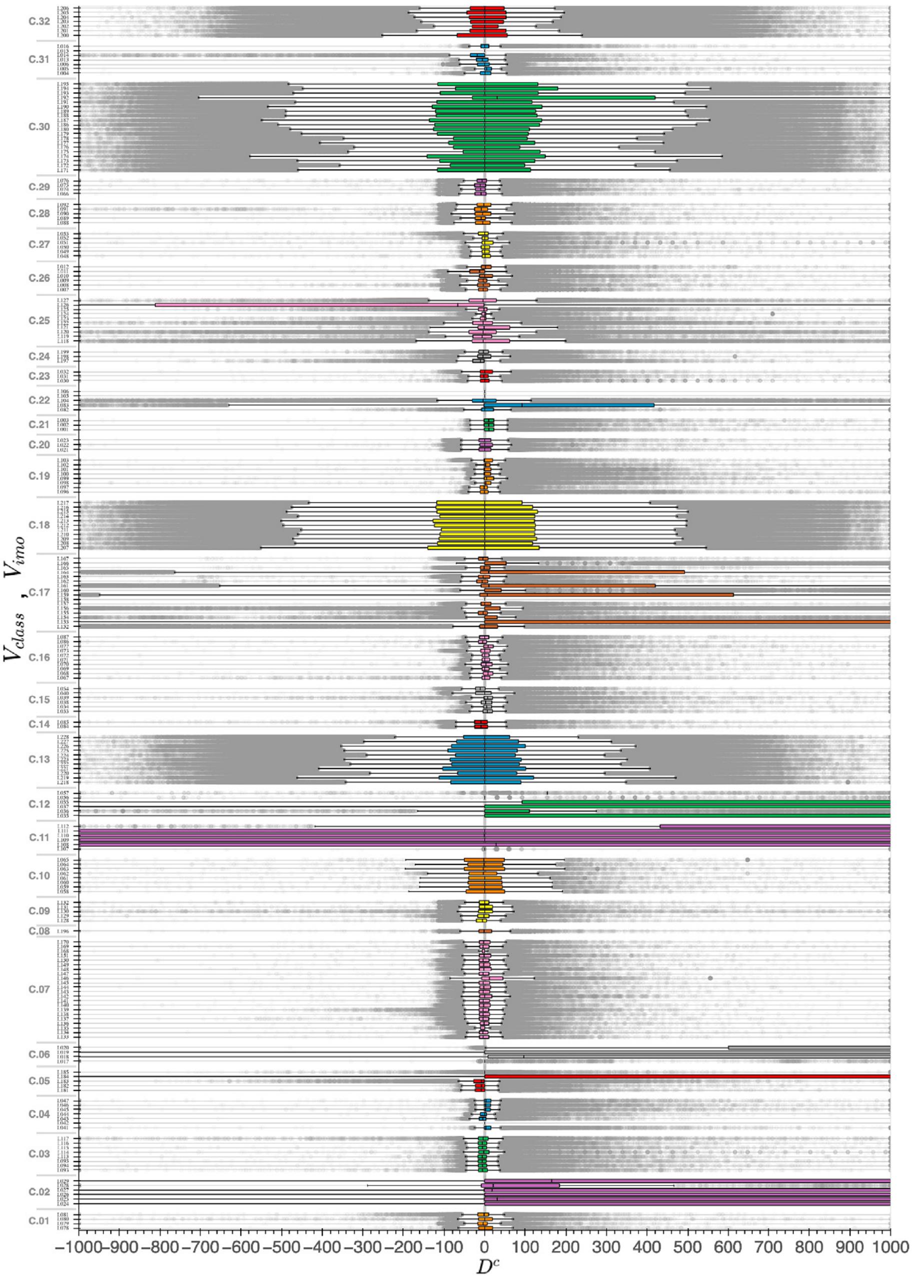


Fig. 10. CAMS distance deviation D^c per vessel V_{imo}, categorized by class V_{class}. Each boxplot represents one vessel's data among I.001 to I.228 over a three-year period (2017–2020).

4. Methodology

4.1. Position measurements conversion models

Geodetic position coordinates are measured in degrees. As mentioned in Section 3.1, degrees close to the equator are longer in meters than those close to the poles. That makes any potential usage of the degree unit problematic for various reasons. In order to avoid errors deriving from estimating degrees over meters, the geodetic coordinates (φ, λ) must be converted into geocentric (Cartesian) coordinates (X, Y) and then after any modification, they are turned back to geodetic. The fundamental notation from the models introduced in Vermeille (2002) is described below:

$$a, b, c = \text{semi} - \text{major axis, semi} - \text{minor axis, eccentricity of reference ellipsoid} \tag{3}$$

$$X, Y, Z = \text{Cartesian geocentric coordinates} \tag{4}$$

$$\lambda, \varphi, h = \text{geodetic longitude, geodetic latitude, geodetic height} \tag{5}$$

It should be noted that the rest of the notation used in this section that is not mentioned above (such as $n, p, q, r, s, t, u, v, w, k$ and D) is utilized as helping variables to avoid the complexity of long equations. Consequently, the coordinate transformation from geocentric to geodetic is described as

$$X = (h + n)\cos\varphi \cos\lambda, \tag{6}$$

$$Y = (h + n)\cos\varphi \sin\lambda, \tag{7}$$

$$Z = \left(h + n - c^2 n\right)\sin\varphi, \tag{8}$$

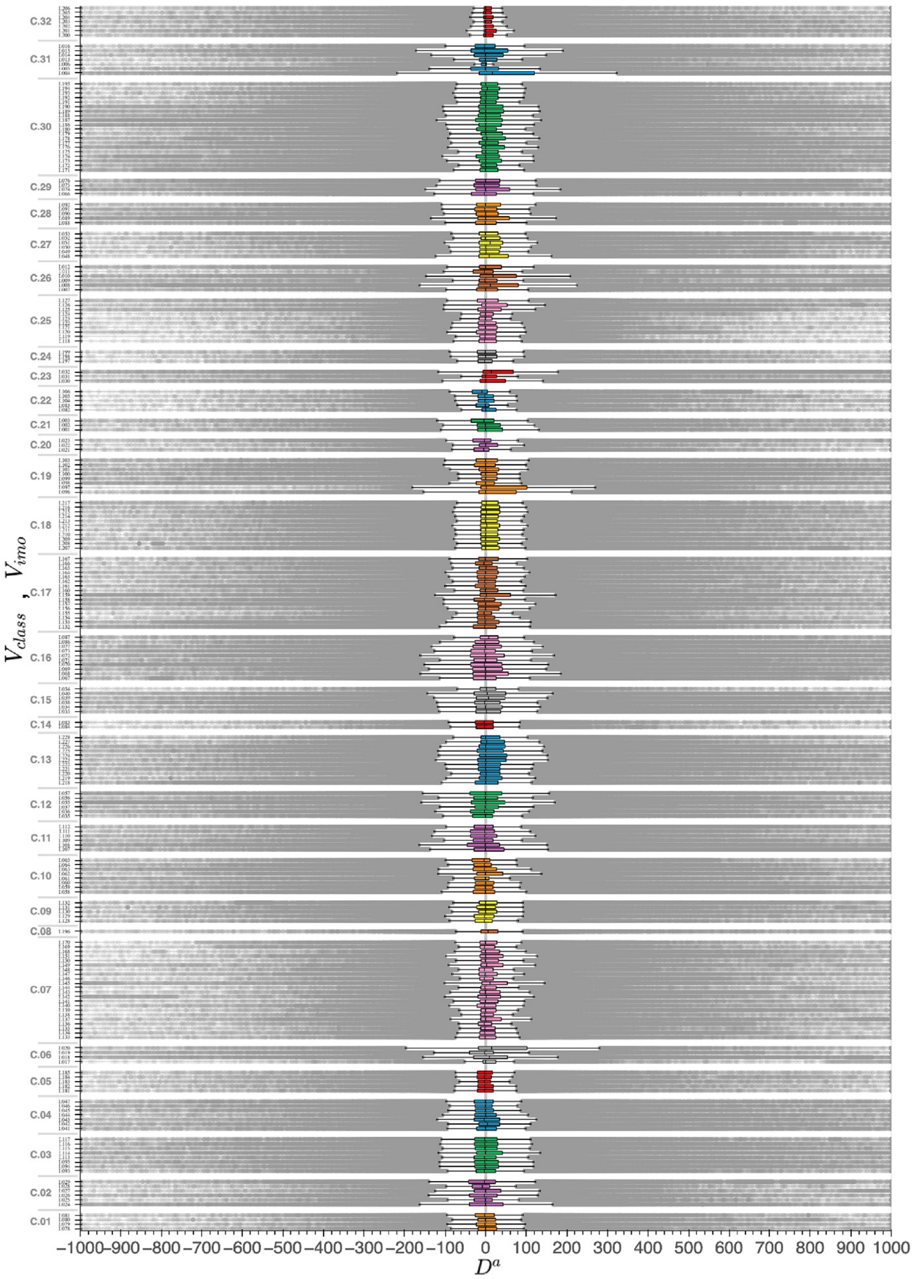


Fig. 11. AIS distance deviation D^a boxplot per vessel V_{imo}, categorized by class V_{class}. Each boxplot represents one vessel's data among I.001 to I.228 over a three-year period (2017–2020).

where:

$$n = \frac{a}{\sqrt{1 - c^2 \sin^2 \varphi}}. \tag{9}$$

To transform geocentric to geodetic coordinates, given that (X, Y, Z) is known, firstly the values of k and D are computed by the following sequence of formulae:

$$p = \frac{X^2 + Y^2}{a^2}, \tag{10}$$

$$q = \frac{1 - c^2}{a^2} Z^2, \tag{11}$$

$$r = \frac{p + q - c^4}{6}, \tag{12}$$

$$s = c^4 \frac{pq}{4r^3}, \tag{13}$$

$$t = \sqrt[3]{1 + s + \sqrt{s(2 + s)}}, \tag{14}$$

$$u = r\left(1 + t + \frac{1}{t}\right), \tag{15}$$

$$v = \sqrt{u^2 + c^4 q}, \tag{16}$$

$$w = c^2 \frac{u + v - q}{2v}, \tag{17}$$

$$k = \sqrt{u + v + w^2} - w, \tag{18}$$

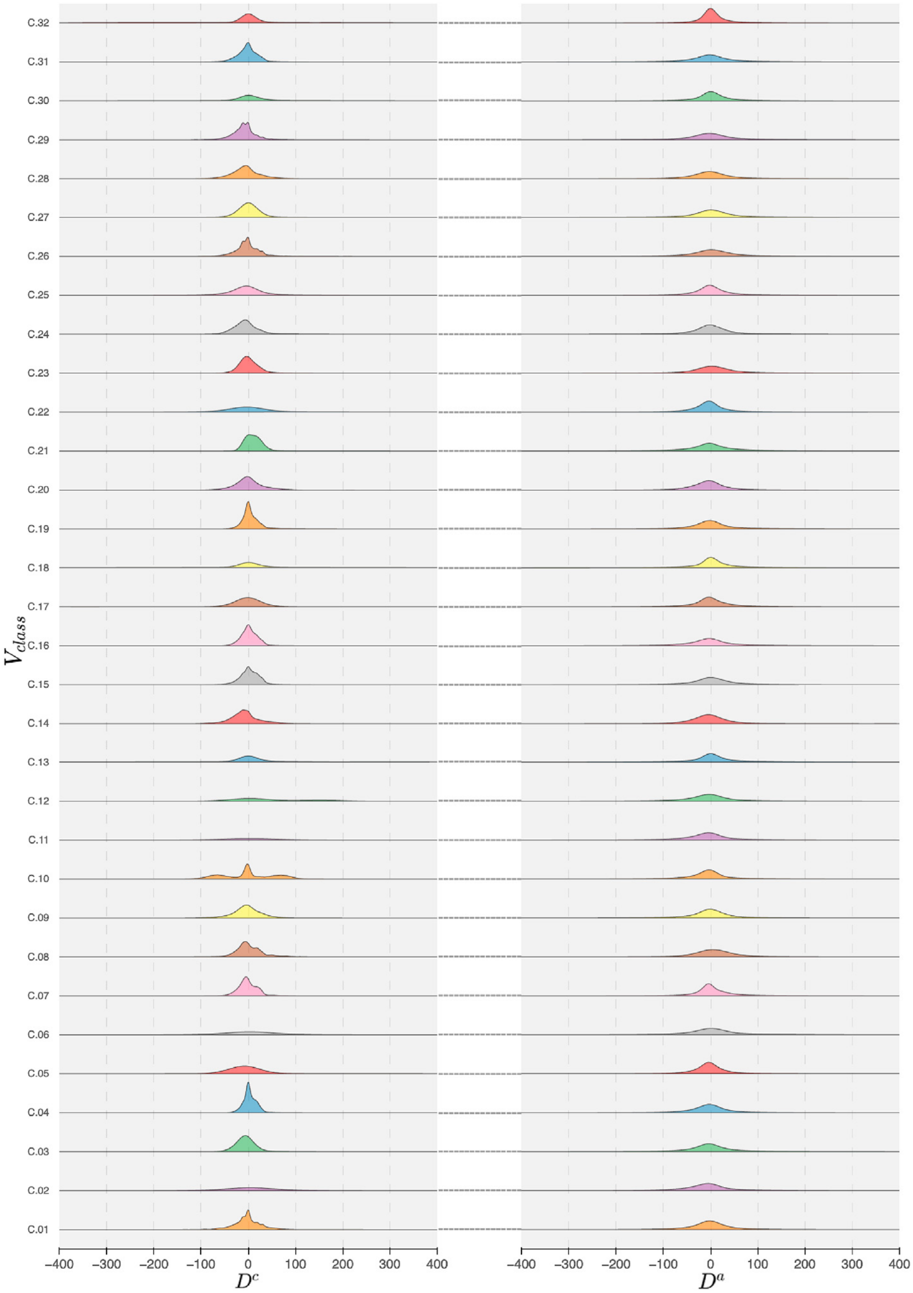


Fig. 12. CAMS distance deviation D^c (left) and AIS distance deviation D^a (right) distribution per class V_{class}. Each distribution represents one class' data among C.01 to C.32 over a three-year period (2017–2020).

$$D = \frac{k\sqrt{X^2 + Y^2}}{k + c^2}. \tag{19}$$

Then, the geodetic coordinates λ, φ and h are formed by

$$\lambda = 2\,\arctan\frac{Y}{X + \sqrt{X^2 + Y^2}}, \tag{20}$$

$$\varphi = 2\,\arctan\frac{Z}{D + \sqrt{D^2 + Z^2}}, \tag{21}$$

$$h = \frac{k + c^2 - 1}{k}\sqrt{D^2 + Z^2}. \tag{22}$$

The computations in this paper were carried out with $a = 6378\,137$ m and $c = 0.081\,819\,191$.

4.2. Interpolation methodology

This algorithm was named after the underlined main function used to replace the missing vessel trajectories. The scope was to make a versatile and highly customizable algorithm that could provide a robust solution for industrial use with the potential to scale up in the future. Additionally, it should be possible to run in batch, meaning on datasets with multiple ships of similar characteristics.

Algorithm (1) sums up the methodology step-by-step including the rule for defining which of the two (AIS or CAMS) coordinates are the more accurate for the vessel in process. It starts by creating a vessel subset from each dataset, one for CAMS such that $vsl^c \subset CAMS$ and one for AIS such that $vsl^a \subset AIS$. Focusing on vsl^a, the algorithm initially converts the geodetic-to-geocentric coordinates as described in Section 4.1. Next, it creates a feature which is the time difference in hours between consecutive timestamps of the vsl^a subset. This feature combined with a minimum threshold of 10 h is used in the next step, where vsl^a is dynamically split into sub-segments.

Algorithm 1. Interpolation methodology algorithm

```
Algorithm 1: Interpolation methodology algorithm
input  : Two datasets, one with CAMS and one with AIS position measurements both including
         data from 228 container vessels
output: One dataset with both CAMS and AIS-corrected position measurements including $D^c$,$D^a$
         and $d_p^b$ validation indicators
1  foreach vessel in CAMS do
2      $vsl^c \subset$ CAMS
3      $vsl^a \subset$ AIS
4      GeodeticToGeocentric($vsl^a$)
5      diff $\leftarrow$ MeasureTimeDiff($vsl^a$)
6      $vsl^a_{lst} \leftarrow [vsl^a_1, vsl^a_2, \ldots, vsl^a_n], \quad n \in \mathbb{Z}^+$
7      foreach $vsl^a_j$ in $vsl^a_{lst}$ where $j \in \mathbb{Z}^+$ do
8          UpSample($vsl^a_j$)
9          LinearInterpolation($vsl^a_j$)
10         $vsl^a_{new} \leftarrow$ Concatenate($vsl^a_{lst}$)
11     end
12     GeocentricToGeodetic($vsl^a_{new}$)
13     $vsl^{c,a} \leftarrow vsl^c \bowtie vsl^a_{new}$
14     $d^c_p \leftarrow$ GreatCircleMethodology($vsl^{c,a}$)
15     $d^a_p \leftarrow$ GreatCircleMethodology($vsl^{c,a}$)
16     $d^c_s \leftarrow$ DifferentialSOGMethodology($vsl^{c,a}$)
17     $d^c_s \leftarrow$ DifferentialSOGMethodology($vsl^{c,a}$)
18     $D^c \leftarrow d^c_p - d^c_s$
19     $D^a \leftarrow d^a_p - d^a_s$
20     $d^b_p \leftarrow$ GreatCircleMethodology($vsl^{c,a}$)
21     if $\tilde{d}^b_p < 1000m$ and IQR($D^a$) > IQR($D^c$) then
22         prefer CAMS over AIS position measurements
23     else if $\tilde{d}^b_p < 1000m$ and IQR($D^a$) $\leq$ IQR($D^c$) then
24         prefer AIS over CAMS position measurements
25     else if $\tilde{d}^b_p \geq 1000m$ then
26         prefer AIS over CAMS position measurements
27     end
28 end
```

The split functionality is applied in order to avoid interpolation when the time distance between consecutive available AIS measurements is greater than 10 h. In such cases, this trajectory gap is left uncorrected in order to avoid interpolation over land. Fig. 13 illustrates an example AIS trajectory. The threshold's number has been chosen after a series of tests for optimal algorithm performance. A smaller number creates a substantial amount of splits of short scaled sub-segments resulting in an outstanding number of missing values. This step can be further improved by invoking a physical system that uses the course over ground (COG) to approximate the location in such long gaps.

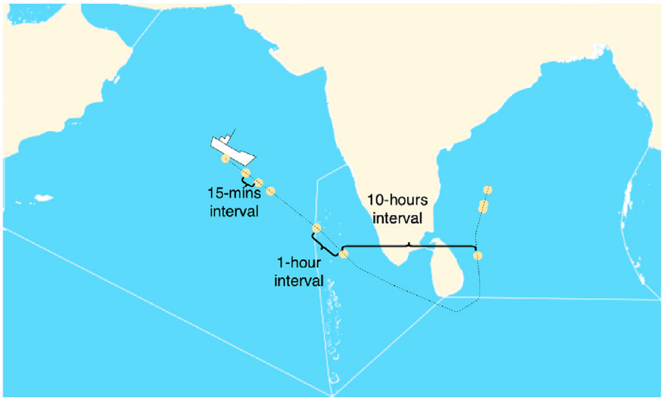


Fig. 13. Example of AIS coordinates with a long (10 h) gap included on the measurements. Yellow dots indicate the measurements and the black dashed line is considered to be the real trajectory of the vessel. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

After the split is over, each sub-segment is initially up-sampled to a 10-min interval frequency similar to that of vsl^c and then linearly interpolated using the deterministic methodology of polynomial interpolation for time-series data described in Lepot et al. (2017). Then all sub-segments are concatenated to form a vessel's dataset where the reverse procedure of geocentric-to-geodetic is applied as described in Section 4.1. Last step before creating the validation indicators D^c, D^a and the safety indicator d_p^b is to merge vsl^c with vsl^a_{new}. For easier interpretation, the flow chart in Fig. 14 illustrates the algorithm up to the point where the rule is applied in line 21.

Apart from the two requirements from Section 3.2 which have defined an initial step on how to validate the position measurements of a vessel, an additional rule has been formed based on a series of trials and errors. The rule is shown in Equation (23) below.

$$P(\varphi,\lambda) = \begin{cases} CAMS, & \tilde{d}_p^b < 1000\text{ m} \text{ and } IQR(D^a) > IQR(D^c) \\ AIS, & \left(\tilde{d}_p^b < 1000\text{ m} \text{ and } IQR(D^a) \leq IQR(D^c)\right) \text{ or } \left(\tilde{d}_p^b \geq 1000\text{ m}\right) \end{cases} \tag{23}$$

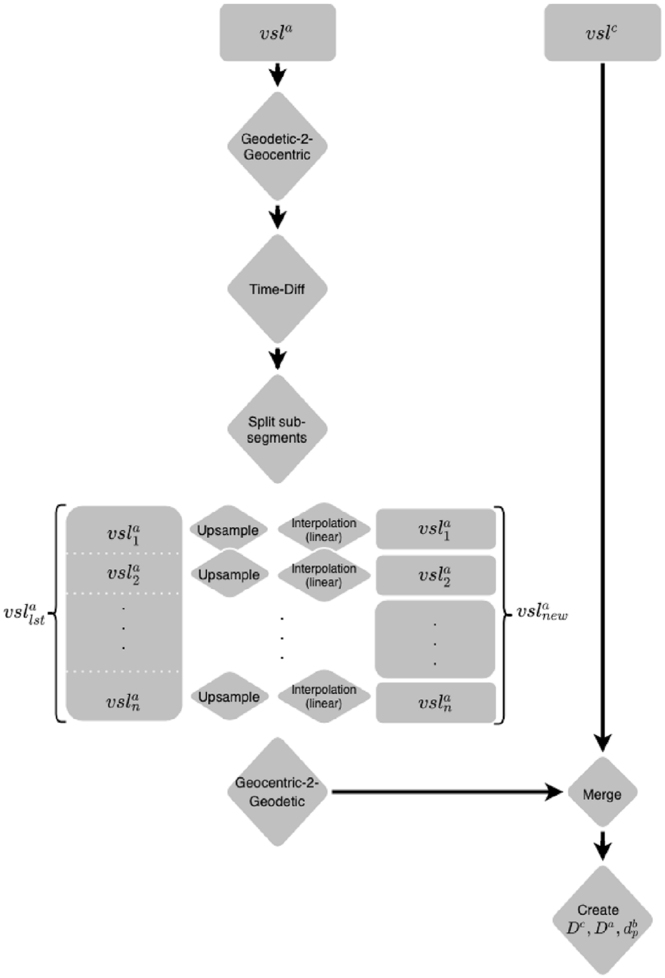


Fig. 14. Flow chart of the interpolation methodology.

The rule is generally sufficient to produce good results. In short, the Equation clarifies that the final position measurement of the historical dataset of one vessel can be drawn directly from the CAMS dataset if $\tilde{d}_p^b < 1000\,\mathrm{m}$ and $IQR(D^a) > IQR(D^c)$. In the contrary, when $\tilde{d}_p^b < 1000\,\mathrm{m}$ and $IQR(D^a) \leq IQR(D^c)$ or $\tilde{d}_p^b \geq 1000\,\mathrm{m}$ then the final dataset enclose position measurement processed in even time intervals derived from the interpolation.

5. Results

According to Algorithm (1)'s output, 18.4% of the fleet on the examined dataset of 228 vessels have been identified as faulty. This accounts for 19.2% of the dataset's total position measurements. The slight percentage difference occurs due to the variation in the number of measurement points from the different vessels. Some of them contribute with a few days of data while others contribute with as much as up to 36 months.

Similar to Fig. 2 which shows the measurements from all 228 ships of the study, Fig. 15 illustrates the algorithm's output after the correction. Fig. 15 contains 80.8% of the position measurements from Fig. 2 since they were identified as CAMS by the algorithm's rule shown in Equation (23). The rest 19.2% of the position measurements were identified as AIS and were estimated using the interpolation methodology.

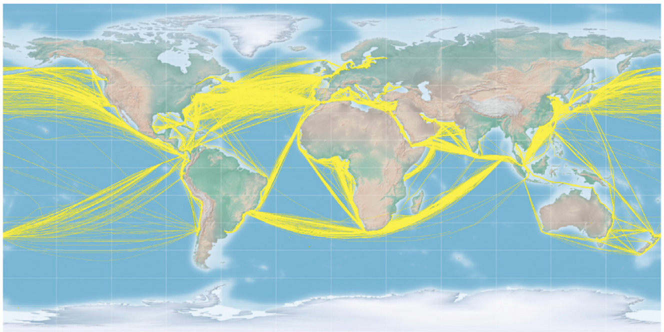


Fig. 15. Corrected position measurements of 228 container ships during 3 years of operation (2017–2020).

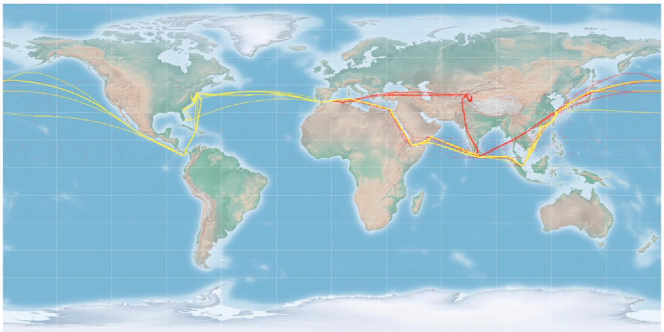


Fig. 16. Trajectory of vessel I.36 from class C.12. Red points refer to the CAMS registered coordinates and yellow refer to even-spaced AIS estimates of the interpolation methodology.

Driving the focus down to per-vessel case, the evaluation of the results could start from a ship with proven bad quality historical position measurements (vessel I.36 from class C.12) with its trajectory shown in Fig. 16. The ship has experienced “N/E issue”, “Zig-Zag issue” and “Drift issue” in its 3-year-long trajectory. The output of Algorithm (1) for this vessel can be seen in yellow-coloured points. In this case, the rule from Equation (23) decided to pick AIS even-spaced estimates over CAMS.

With a glimpse at the map of Fig. 16, it is clear that the yellow points eliminate all degrading sources mentioned for that vessel. It should be noted that the raw AIS coordinates as sourced from the external provider were missing by 64% when they merged with the CAMS dataset vsl^c. The interpolation methodology decreased the missing value percentage to $< 0.1\%$.

In the next examined ship, Algorithm 1 chose CAMS over even-spaced AIS estimates when deciding on the best quality coordinates. In fact, when the boxplots (referring to each ship's D^c and D^a IQRs) from Figs. 10 and 11 are compared for vessel I.48 of class C.27, D^c's boxplot is narrower than that of D^a's which confirms the second scale of the rule in Equation (23). Fig. 17 showcases the vessel's trajectory both in CAMS and in AIS even-spaced estimates.

It is clear from Fig. 15 that most of the major issues described in Section 3.1 were eliminated. However, there are still a few visible land crossing measurements that require further attention. In line with the acquired knowledge from applying alternative algorithms (Linear/Non-linear Kalman Filter) during the process of building the interpolation methodology, it can be concluded that a hybrid model invoking a state-space model into the interpolation methodology could bring better results in historical data validation and correction. Additionally, it could form the fundamentals to build a real-time model that can validate and correct position measurements right before the data has been recorded into the ship's database, serving the scope of vessel-specific mode.

6. Summary and conclusions

This paper has been conducted to tackle the a posteriori data loss/modification of position measurements which refers to the posterior modification/loss of the position data after the signal has successfully

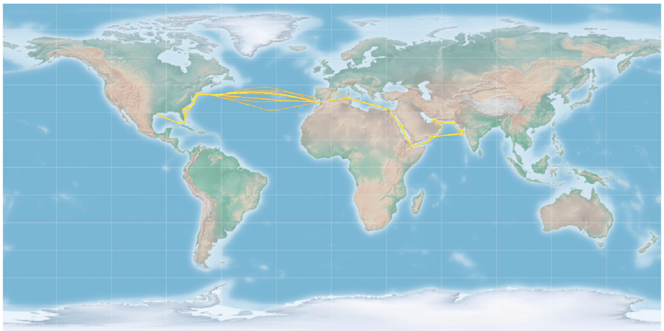


Fig. 17. Trajectory of vessel I.48 from class C.27. Red points refer to the CAMS registered coordinates and yellow refer to even-spaced AIS estimates of the interpolation methodology.

arrived at the GNSS receiver. The faulty measurements were categorized into 6 groups; (i) the “N/E issue”, (ii) the “Zig-Zag issue”, (iii) the “Scatter issue”, (iv) the “Drift issue”, (v) the “Frozen issue” and finally (vi) the “Bit-rate issue”. Some groups were only noticeable in CAMS from specific providers while others were apparent in ships with distinct CAMS installations.

Furthermore, three validation indicators were created with which it can be identified if a ship carries faulty measurements in a specified time range. These validators were designed to be used in batch processing in shore-specific mode and not for individual data points.

Lastly, an algorithm/methodology for correcting the vessels with identified faulty measurements was designed. The algorithm uses interpolation as its main ingredient and its fundamental scope is to reverse the measurements back to their raw state, similar to how they looked before leaving the GNSS receiver towards the company's data centre. The methodology was proven to be a satisfactory foundation for a future real-time position (vessel-specific mode) measurements validation and correction algorithm mainly due to its highly customizable algorithm.

6.1. Future work

There are various processes that can be improved to achieve the optimal results on the matter of interest. Initially, the problem should be divided into two modes, either a vessel-specific or a shore-specific mode. Depending on the mode, the focus should either be on solving the problem in real-time regardless of any potential occurring issue (vessel-specific mode) or correct the historical position measurements (shore-specific mode) using a methodology similar to the algorithm introduced in this paper. Based on the above, shipping companies have the following options as far as position measurement validation and correction is concerned:

1. Go through each CAMS vendor (or any other internal system similar to CAMS that stands between the raw and the logged data) and fix each issue separately for each vessel. This option requires that the investigator is fully aware of what the issue is related to each vendor so as to avoid going through the same vessel again in case there is an issue that was not identified in time. This option requires a substantial amount of attention and effort not to mention the cost. For instance, a vessel was recently attended by a vendor to tackle the “Scatter issue” and the “N/E issue” but after fixing them, it was observed that the vendor accidently missed deploying a parameter that was functional before the intervention. On other vessels from classes with medium-sized boxplots as C.32, C30, C.18, C.13 and C.10 from Fig. 10, there is an ongoing internal investigation where the vendor is suspected of using a position measurement update threshold filter to display steady values. These are just a couple of examples to emphasize the complexity of this option.
2. Overtake CAMS and start registering raw data on a stable and remotely-configurable platform. Such work requires a similarly large amount of resources as the previous option but it sounds very promising in terms of reliability considering future interventions. The main issue here is that position measurement is only one out of the hundreds of data points that pass through CAMS. If CAMS is overtaken on position measurements, the rest of the data points should also be included in the new platform. Given that the position measurements are among the easiest to distinguish flaws on, due to their nature of being able to be plotted on a map, it is assumed that CAMS has equally modified other data points.
3. Focus on the vessel-specific mode and create a generic position validation/estimation system that should be deployed before CAMS. Established methods exist for this purpose that leverage state-space modelling and Bayesian filtering. The system should use inputs both from raw data and noon reports. After validating the position measurements, if approved the system will export the raw position. If not approved the system will be able to estimate the position, up to a specific time horizon. The noon report values will serve the scope of (i) setting up the system's initial conditions and (ii) providing regular validation feedback to the system estimates assuming that noon reports are the most reliable source of information the system can access in real-time, given that AIS might not be available for up to an entire day.
4. Focus on the shore-specific mode and improve the introduced algorithm. More specifically, the algorithm should be able to identify AIS gaps larger than 10 h where it should invoke a physical system using sensor fusion to estimate the position. Additionally, it should be extended into a per-month or a per-day state. Right now, the validation refers to the IQR of the data from the whole time that the ship has operated. Apparently, this is not a fair judgement because many CAMS systems have been fixed and after that, they register good quality data, or the opposite. Unfortunately, the proposed validators were built for batch corrections and not for individual data points.

Evaluating the above options for any shipping company that faces similar position data quality issues, option 1 is considered a very costly option. Option 2 is considered the best one given the future of data quality as a whole but not the most cost-effective among the four. Option 3 is the most cost-effective next step. The new vessel-specific system would be useful even after the company decides to move on to option 2, given that it can be used as an additional validation of the new platform. By choosing option 4, the company would at least be able to trust the historical position measurements in order to use it for more accurate weather routing which will drive fuel consumption down with the least effort among the four options.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The idea to investigate issues on position measurements and design a system that validates and corrects them was formulated during discussions between DTU, Maersk R&D and Maersk StarConnect. The data collection has been a collaboration of Maersk R&D and Maersk StarConnect teams. The present work has been supported by “InnovationsFonden Danmark” with case number 8053-00231B and “Den Danske Maritime Fond” with case number 2018–060.

References

Akos, D.M., 2012. Who's afraid of the spoofer? gps/gnss spoofing detection via automatic gain control (agc). NAVIGATION. J. Inst. Navig. 59 (4), 281–290.

Broumandan, A., Jafarnia-Jahromi, A., Dehghanian, V., Nielsen, J., Lachapelle, G., 2012. Gnss spoofing detection in handheld receivers based on signal spatial correlation. In: Proceedings of IEEE/ION PLANS, pp. 479–487, 2012.

Chang, L., Niu, X., Liu, T., 2020. Gnss/imu/odo/lidar-slam integrated navigation system using imu/odo pre-integration. Sensors 20 (17), 4702.

Contributors, W., 2021a. Automatic identification system — wikipedia, the free encyclopedia. Online, 05-July-2021. https://en.wikipedia.org/wiki/Automatic_identification_system. URL.

Contributors, W., 2021b. Haversine formula. Online, 15-Sept-2021. https://en.wikipedia.org/wiki/Haversine_formula. URL.

Fukuda, G., Hatta, D., Guo, X., Kubo, N., 2021. Performance evaluation of imu and dvl integration in marine navigation. Sensors 21 (4), 1056.

Gakstatter, E., Feb 2015. What exactly is gps nmea data? URL. https://www.gpsworld.com/what-exactly-is-gps-nmea-data/.

Gao, G.X., Sgammini, M., Lu, M., Kubo, N., 2016. Protecting gnss receivers from jamming and interference. Proc. IEEE 104 (6), 1327–1338.

Haithcoat, T., 1999. Coordinate systems. Online, 20-Sept-2021. https://mospace.umsystem.edu/xmlui/bitstream/handle/10355/2950/CoordinateSystems.pdf?sequence=1&amp;isAllowed=y. URL.

Heukelman, C., Apr 2018. About ugps vs gnss - what engineers need to know for their designs. URL. https://www.semiconductorstore.com/blog/2018/GPS-vs-GNSS-What-Engineers-Need-To-Know-For-Their-Design-Symmetry-Blog/3083.

Hu, S., Fang, Q., Xia, H., Xi, Y., 2007. Formal safety assessment based on relative risks model in ship navigation. Reliab. Eng. Syst. Saf. 92 (3), 369–377.

Ikonomakis, A., Nielsen, U.D., Holst, K.K., Dietz, J., Galeazzi, R., 2021a. Historical position measurement validation and correction using ais data - an account from a larger shipping company. Proc. HullPIC 150–168.

Ikonomakis, A., Nielsen, U.D., Holst, K.K., Dietz, J., Galeazzi, R., 2021b. How good is the stw sensor? an account from a larger shipping company. J. Mar. Sci. Eng. 9 (5), 465.

ITU, Nov, 2014. M.1371: Technical characteristics for an automatic identification system using time-division multiple access in the vhf maritime mobile band. URL. https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.1371-5-201402-I!!PDF-E.pdf.

Johansen, T.A., Fossen, T.I., 2016. The exogenous kalman filter (xkf). Int. J. Control 90 (2), 161–167. https://doi.org/10.1080/00207179.2016.1172390. URL.

Karaim, M., Elsheikh, M., Noureldin, A., Rustamov, R., 2018. Gnss error sources. In: Multifunctional Operation and Application of GPS, pp. 69–85.

Lepot, M., Aubin, J.-B., Clemens, F.H., 2017. Interpolation in time series: an introductive overview of existing methods, their performance criteria and uncertainty assessment. Water 9 (10), 796.

Liang, M., Liu, R.W., Zhong, Q., Liu, J., Zhang, J., 2019. Neural Network-Based Automatic Reconstruction of Missing Vessel Trajectory Data, pp. 426–430.

Medina, D., Lass, C., Marcos, E.P., Ziebold, R., Closas, P., García, J., 2019. On gnss jamming threat from the maritime navigation perspective. In: 2019 22th International Conference on Information Fusion (FUSION). IEEE, pp. 1–7.

Morong, T., Puričer, P., Kovář, P., 2019. Study of the gnss jamming in real environment. Inter. J. Electro. Telecommunicat. 65 (1), 65–70.

Nielsen, U.D., Ikonomakis, A., 2021. Wave conditions encountered by ships—a report from a larger shipping company based on era5. Ocean Engineering. URL. https://www.sciencedirect.com/science/article/pii/S0029801821009690, 237.

NMEA, Jul 2021. Nmea-0183 standard. URL. https://www.nmea.org/content/STANDARDS/NMEA_0183_Standard.

Noureldin, A., Karamat, T.B., Georgy, J., 2013. Fundamentals of Inertial Navigation, Satellite-Based Positioning and Their Integration.

Núñez, J.M., Araújo, M.G., García-Tuñón, I., 2017. Real-time telemetry system for monitoring motion of ships based on inertial sensors. Sensors 17 (5), 948.

Perera, L., Guedes Soares, C., 2017. Weather routing and safe ship handling in the future of shipping. Ocean Eng. 130, 684–695.

Psiaki, M.L., Humphreys, T.E., 2016. Gnss spoofing and detection. Proc. IEEE 104 (6), 1258–1270.

Ryu, J.H., Gankhuyag, G., Chong, K.T., 2016. Navigation system heading and position accuracy improvement through gps and ins data fusion. J. Sens. 2016, 7942963.

Schmidt, D., Radke, K., Camtepe, S., Foo, E., Ren, M., 2016. A survey and analysis of the gnss spoofing threat and countermeasures. ACM Comput. Surv. 48 (4), 1–31.

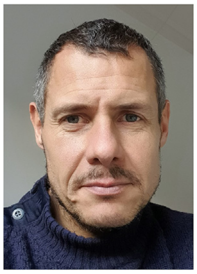
Shimizu, E., Pedersen, E., 2006. Robust tracking control for ship-to-ship operations. In: The Sixteenth International Offshore and Polar Engineering Conference. OnePetro.

Venezia, M., Dec 2015. What is the difference between gnss and gps? URL. https://www.semiconductorstore.com/blog/2015/What-is-the-Difference-Between-GNSS-and-GPS/1550.

Vermeille, H., 2002. Direct transformation from geocentric coordinates to geodetic coordinates. J. Geodes. 76 (8), 451–454.

Zhang, S., Gong, L., Zeng, Q., Li, W., Xiao, F., Lei, J., 2021. Imputation of gps coordinate time series using missforest. Rem. Sens. 13 (12), 2312.

Angelos Ikonomakis is a research scientist at A.P. Moller-Maersk, and an industrial PhD candidate at the Department of Mechanical Engineering, Technical University of Denmark. The focus of his research is on state space models, sensor fusion and ship propulsion performance models towards the reduction of the ecological footprint of ships.



Ulrik D. Nielsen is an associate professor at the Department of Mechanical Engineering, Technical University of Denmark, where he also obtained his Ph.D. degree in 2005, and where he has had the main part of his professional career. Since 2014 he has been affiliated with the Norwegian University of Science and Technology as an associate professor II at the Centre for Autonomous Marine Operations and Systems (NTNU AMOS). Ulrik Dam Nielsen works with energy efficiency and safety of ships, with a special interest in safe and efficient operation in waves.

Klaus K. Holst is a principal scientist at A.P. Moller-Maersk, an integrated transport and logistics company operating in 130 countries and today one of the worlds largest container shipping companies. He has a PhD degree from the Department of Biostatistics at the University of Copenhagen. The focus of his research is on causal inference, latent variable and state space models, and general aspects of computational statistics.

Jesper Dietz has been working within the maritime domain since 2000 where he graduated with a Ph.D. degree on naval architecture from the Technical University in Denmark (DTU). Key focus has been on noise and vibration investigation on commercial vessels and leisure yachts and building vessel performance products for the APMM Fleet. Current focus is to provide data-driven advice on how to get from A to B in the most safe, sailable and energy efficient way.

Roberto Galeazzi is Head of Centre for Collaborative Autonomous Systems, and Associate Professor of Control Theory and Technology at the Department of Electrical Engineering, Technical University of Denmark. He obtained a Ph.D. degree in Automation and Control in 2010 from the Technical University of Denmark. He researches the field of control and estimation theory, with focus on resilient autonomous marine systems. His research contributions are in motion control, sensor fusion, motion planning, and fault-tolerant control. He chairs the IFAC Technical Committee in Marine Systems.