Some Methods for Addressing Errors in Static AIS Data Records

Steven D. Meyers*[a], Yasin Yilmaz[b] and Mark E. Luther [a]

[a]*Center for Maritime and Port Studies*
University of South Florida,
St. Petersburg, FL, USA
smeyers@usf.edu, mluther@usf.edu

[b]*College of Engineering*
University of South Florida,
Tampa, FL, USA
yasiny@usf.edu

*corresponding author

Abstract

The Automatic Identification System (AIS) provides essential services in support of maritime domain awareness. Accurate AIS values for hull dimension and type are often critical for safe and efficient management of ship traffic, and for development of new artificial intelligence maritime algorithms. AIS variables are subject to fault from multiple sources, ranging from bad weather to human error. New heuristic methods for correcting ship draft, beam, and class were developed and evaluated, using AIS data in the vicinity of large Florida ports as a test bed. Novel low order polynomials for 9 broad functional vessel classes yielded predicted values for draft and beam as functions of vessel length. The majority of relative differences between predicted and reported values were <0.1. A logistic regression (LR) multiclass classification scheme using the residuals from these polynomial predictions generally showed good agreement between estimated and reported vessel class. The LR scheme demonstrated skill in verifying AIS-transmitted classification, detecting incorrectly classified vessels, and flagging those with incorrect draft or operating near an extreme draft. A diagnostic of reports whose classification had very low and very high confidence suggested directions for further improvement of the algorithm. A new hierarchy for processed AIS data is proposed.

Keywords: automatic identification system; multiclass classification; vessel identification; logistic regression; maritime domain awareness

Introduction

The Automatic Identification System (AIS) is a maritime vessel recognition scheme originally designed to increase situational awareness between vessels, and between vessels and ports (Harre, 2000; Murk, 1999). Through the AIS, vessels transmit their identifying information every few minutes using automated radio signals. Two general categories of data are provided by the AIS: static and dynamic. Static variables are typically fixed quantities, including the Maritime Mobile Service Identity (MMSI) number, length (L), beam (B), draft (D), and type (Y), though the draft of cargo and tanker ships can change when material is offloaded or onloaded. Crew members are responsible for entering the static values into the AIS transmitter. Dynamic variables include time of transmission, vessel position, speed over ground, and heading. These are typically entered into the report automatically by instrumentation.

AIS data can be accessed in real-time using specialized receivers that pickup broadcasts within a ~50 km radius, or with a slight delay through data service companies such as Pole Star USA, Marine Traffic, GateHouse Maritime, and others that access the ground-based as well as satellite AIS receivers. These companies often provide small amount of AIS data to researchers without charge. Processed AIS data in US coastal waters is also available, sometimes with a significant delay but without cost, from Marine Cadastre (marinecadastre.gov/ais), a combined service of the U.S. Department of Commerce's National Oceanic and Atmospheric Administration (NOAA) Office for Coastal Management and the U.S. Department of the Interior's Bureau of Ocean Energy Management (BOEM). Regardless of the provider, most of these data are offered with little to no error flagging or correction. This may be because objective error handling routines for AIS data are still under development, most of which have focused on the dynamic variables. There have been few publications regarding the static AIS variables in this context. Adoption of a standard set of handling routines would facilitate AIS usage in a range of applications. The outline for such a system is proposed at the end of this article.

AIS data have become essential to the monitoring and management of global vessel traffic, as well as in academic and private sector maritime research programs (Tu et al., 2017; Yang et al., 2019). The latter encompasses many areas of maritime operations, including relatively simple maps of vessel traffic density (Demšar and Virrantaus, 2010; Shelmerdine, 2015), predicting future routes and collision avoidance (Chen et al., 2018; Rong et al., 2019; Silveira et al., 2013; Wang et al., 2013), predicting arrival times (Dobrkovic et al., 2016; Jahn and Scheidweiler, 2018; Xin et al., 2019), and detecting anomalous vessel movement (Liu, 2015; Oh et al., 2018; Sidibé and Shu, 2017). Lim et al. (2018), Robards et al. (2016), and Zhou et al. (2019) provide reviews of AIS applications, many of which utilize artificial intelligence / machine learning where AIS records are used as a source of training data.

Incomplete or inaccurate AIS reports can confound studies of maritime operation. Such faulty data arise from multiple causes, such as human error, instrument failure, an overwhelmed transmission spectrum, and atmospheric interference (Emmens et al., 2021; Harati-Mokhtari et al., 2007). Processed AIS data may also be subject to errors or inconsistencies in sorting, filtering, or transcription. Most previous studies have focused on detection of dynamic AIS errors (Bošnjak et al., 2012; Sun et al., 2021; Zhao et al., 2018). Of relevance to this study, Guo et al. (2021) used kinematically-based cubic polynomials to model trajectories and determine errors in vessel position and speed by their generic "distance" from the model. There have been few publications that focused on correcting static AIS errors. Wang et al. (2021) applied the Random Forest algorithm to AIS static values to identify five vessel classes. Sheng et al. (2018) developed a logistic regression binary classifier that discriminated between Cargo and Fishing class vessels based on their position, course, and speed near Shantou, China. Steidel et al. (2019) suggested correcting AIS Destination data using a combination of automated and direct communication with each vessel. Atypical B vs. L values were used to manually identify 3 misclassified, misreported, or unusually large vessels in a narrowly defined group of bulk carriers (Smestad et al., 2017).

Figure 1. Map of peninsular Florida. The 5 largest ports are indicated.

This study examines some novel methods for correcting errors in static variables associated with hull dimension and type for many vessel classes. As demonstrated below, these variables were found to be interrelated and could be used to help determine missing values or detect inconsistencies in the group of values for many vessels. The methods examined start with simple heuristic drop-out replacement, but also include a new algebraic representation that takes advantage of the dependence between the static variables related to hull geometry, and a multiclass classification (MCC) scheme for confirming functional vessel class. The methods developed here can be used to flag or correct some missing or unusual static AIS variables.

Section 2 describes the AIS data used in this study. Restricting the analysis to underway vessels in the vicinity of large Florida ports (Figure 1) reduced computational cost for this initial analysis while retaining diversity of vessel types. Polynomial models and logistic regression are described as they relate to this study. Section 3 presents the geometric relations of hull dimensions found when partitioning by vessel functional class. The number of missing or inconsistent static values is then examined, and the potential use of polynomials to represent geometric hull relations and correct these errors is tested. This is followed by the development and testing of the new vessel classification system. Section 4 is a Discussion of the findings and how the methods employed might be adapted or improved. A new system of organizing processed AIS data is proposed.

2. Data and Methods

2.1 AIS Data

The AIS is divided into Class A and Class B. Class A transmissions have a range around 30-50 km, are prioritized by the system, and are mandatory for large and passenger vessels subject to the International Convention for the Safety of Life at Sea (SOLAS). Class B transmissions have a range ~16 km, are not prioritized, and are used by non-SOLAS craft, typically personal watercraft and some smaller, domestic commercial vessels.

AIS reports for the years 2015-2019 were obtain from Marine Cadastre who added Class B to their AIS records starting in 2018. Years prior only contained Class A reports. Also prior to 2018, L and B were provided to a precision of 0.01 m, but afterwards were provided as integer values. A relatively small subset of these reports was utilized in this analysis to facilitate development of the algorithms presented in this study.

Following Mitchell and Scully (2014), irregular polygonal Areas of Interest (AOI) around the five largest commercial ports in the state of Florida, Miami, Everglades, Jacksonville, Tampa, and Palm Beach, (Figure 1), were used to delimit a subset of AIS records. Vessel traffic is concentrated around ports. Extracting AIS records near them reduces the volume of records to be examined while retaining a breadth of sample comparable to that obtained from larger areas (e.g., the entire coast of Florida) that would include many of the same vessels as they traveled between ports. Each AOI included the port and its access waters and channels. AIS reports from all the ports were binned and analyzed collectively. Vessels that were slow or not moving (speed < 0.5 kn) for an entire year were not considered. This yielded a nominal 10^7 AIS reports per year of which <~0.01% lacked an MMSI, and were removed from the analysis. Some of the reports with missing MMSI provided an IMO number which could have been be used to check the vessel identification using an external database (Winkler, 2012), but the focus here was on exploiting relations between the geometric static values.

The unique MMSI and associated values of L, D, B, and Y reported in the AIS were determined. The number of vessels by class, and the number of vessels in each class with problems in their statics were found. For example, the number of vessels reporting both $D = 0$ and $D > 0$ (at different times) provided a measure of the utility for a direct replacement method. Calculating this same number but restricted to $L > 30$ m, eliminated many personal craft that have a higher rate of static AIS errors (Meyers et al., 2020), and helped focus the analysis on commercial and other ships more likely to be professionally maintained.

2.2 Functional Vessel Classes

Vessel identification in the AIS includes a choice from about 100 unique numbers that indicate vessel type such as search and rescue, recreational, cargo, and tanker, with the latter two further divided into a general type or one of several hazard classifications. Marine Cadastre organizes many of these AIS types into functional classes. A similar prescription was followed here, with each AIS report being labeled according to the class for the reported type (Table 1). About 10-15% of the vessels were not readily incorporated into a functional class (e.g., types 1005, 1007, 1018), so were not part of the class-based analysis. The number of unique vessels within each class was determined for each year 2015-2019 (Tables 2, 3). Large year over year changes in the relative number of vessels for some classes appear to have been associated with changes in the processing of the AIS data provided by Marine Cadastre. For example, in 2018 several Supply class vessels started reporting as type 90, which is 'unspecified', decreasing the number in the class. Similarly, many pilot and tender vessels made the opposite switch in 2018, changing from an unspecified type to one that fit within the Enforcement class as defined here, though most of these were smaller vessels (L<30 m) so did not impact the bulk of the analysis. Additionally, a small number of military vessels became identifiable as such in 2018 before which they were typically listed as 'public' or 'other' AIS types.

Table 1. AIS types in defined functional vessel classes, and the number of unique vessels in each class by year.

Class	AIS Vessel Type	2015	2016	2017	2018	2019
Recreational	36,37,1019	3011	3595	3858	5953	6596
Cargo	70-79,1003,1004,1016	1263	1306	1266	1189	1129
Tug	21,22,31,32,52,1023,1025	342	373	395	404	373
Tanker	80-89, 1017, 1024	303	262	244	218	212
Passenger	60-69, 1012-1015	171	212	245	260	263
Fishing	30,1001,1002	51	1025	158	211	224
Supply	1010	28	34	42	0	0
Research	1020	24	22	24	0	0
Enforcement	35,50,53,55	0	2	3	39	55

It was useful to define the set of all AIS reports (A) such that L, B, D, and Y are positive, real-valued numbers. That is, the set $A = \{k: L_k, B_k, D_k, Y_k > 0\}$, where k indexes the reports. Further, subsets of A for a particular class c, $S_c = \{A: Y \in c\}$ and its complement $S_c' = \{A: Y \notin c\}$ were defined.

Table 2. Total numbers by year: Number of unique MMSI, number with only zero or missing values for the indicated static variable, number with multiple D, number with multiple D including at least one zero value, number with all hull dimensions but undefined type.

	2015	**2016**	**2017**	**2018**	**2019**
# Unique Vessels	6728	7561	8428	9052	9838
# all L=0	1449	1928	2843	2220	2263
# all D=0	4310	5327	6401	6924	7827
# all B=0	3178	3931	4808	4017	3899
# all Y=0	1378	581	1994	487	683
# Multiple D	147	883	523	118	99
# Multiple w/D=0	9	846	491	25	10
# LBD>0 & Y=0	42	6	28	10	11

Table 3. Same as Table 2 but restricted to L>30 m.

	2015	**2016**	**2017**	**2018**	**2019**
# Unique Vessels	2472	2520	2468	2422	2371
# all D=0	244	451	562	464	508
# all B=0	80	181	185	177	180
# all Y=0	51	3	24	16	17
# Multiple D	136	804	474	93	91
# Multiple w/D=0	4	768	443	5	3
# LBD>0 & Y=0	5	1	4	4	6

2.3 Replacement Methods for Static AIS

The 2018 change in some AIS types suggested a simple method for improving the accuracy of static descriptors for a vessel. If a static AIS variable is accepted as valid during one time period, but provides a different, invalid or missing value during another time, then the valid value can be used to replace the values in question. This was the first method assessed in this study.

Table 4. Quadratic fitting for each class (Table 1) beam and draft, based on 2015-2019 AIS records. Shown are the class name, maximum AIS vessel length value in class (L_{max}), the extrema vessel length (L_{ex}), fitting coefficients (1), number of unique vessels used in the fit (N), the root-mean-square difference between estimated and actual values in the fit (RMSD), and the mean relative absolute difference (MRAD) of the fit.

Class	L_{max} (m)	L_{ex} (m)	**c**$_2$ (10^{-4} m^{-1})	**c**$_1$	**c**$_0$ (m)	**N**	**RMSD** (m)	**MRAD**
Beam								
Cargo	200	-46.9	4.15	0.0389	8.16	2198	1.906	0.058
Tanker	200	-159.3	3.03	0.0965	3.35	576	1.697	0.047
Passenger	199	188.1	-6.80	0.2570	0.75	67	3.052	0.141
Tug	180	197.9	-4.60	0.1808	5.03	379	2.783	0.101
Fishing	40	58.3	-20.5	0.2386	1.90	36	1.059	0.136
Recreational	163	-707.7	0.84	0.1187	3.33	667	1.335	0.089
Research	126	18.1	21.4	-0.0775	9.64	35	4.012	0.142
Supply	130	30.2	12.4	-0.0746	15.48	46	4.608	0.153
Draft								
Cargo	367	366.4	-1.10	0.0812	-1.21	3048	1.408	0.125
Tanker	337	390.2	-1.40	0.1069	-3.27	718	1.405	0.101
Passenger	362	498.4	-0.35	0.0353	0.94	182	0.593	0.094
Tug	180	118.0	7.00	0.1651	-0.29	379	0.996	0.148
Fishing	40	14.4	-2.70	0.0079	2.60	36	0.616	0.191
Recreational	163	-6.1	2.31	0.0028	2.13	667	0.870	0.201
Research	126	145.9	-4.20	0.1225	-1.36	35	0.706	0.164
Supply	130	145.4	-5.20	0.1519	-2.97	46	0.633	0.110

The second method was developed to fill missing B and D values when no such replacement value is available, and to potentially detect faulty values of these variables. Hull aspect ratios such as D/L are often selected by marine engineers to maximize operational performance (Bertram and Schneekluth, 1998; Papanikolaou, 2014; Zhang et al., 2008), and therefore often vary in a consistent way within a functional class. The dependence of beam $B(L)$ and draft $D(L)$ on length for each class were represented using n-degree polynomials with independent variable L as

$$\phi_n(L) = c_0 + \sum_{i=1}^{n} c_i L^i \tag{1}$$

where the constants c_i were determined through standard least-squares (Table 4). A minimum of 10 independent (L, S) pairs for each class were required for the estimate, where S represented the static value B or D being modeled. Changes in vessel draft due to changes in deadweight tonnage were not represented by (1). Bulk measures of the accuracy of (1) compared to values from AIS were root mean square difference (RMSD)

$$\sqrt{\frac{1}{N_c}\sum_{k=1}^{N_c}(\phi_n(L_k) - S_k)^2} \tag{2}$$

and mean relative absolute difference (MRAD)

$$\frac{1}{N_c}\sum_{k=1}^{N_c}\frac{|\phi_n(L_k) - S_k|}{S_k} \tag{3}$$

where L_k is the k-th AIS length value in class c, S_k is the matching static value, and k=1,…,N_c. The relation between $(\phi_n(L_k) - S_k)/S_k$ and L_k was also examined to further evaluate this method of estimating static values.

2.4 Multiclass Classification

Logistic regression (LR) is widely used to represent a dichotomous (2-valued) variable (y) that has a single transition between one value and the other (generally 0 and 1), dependent upon predictor variables $\boldsymbol{X}$ (Hilbe, 2016; Hosmer Jr et al., 2013). Here LR was used to identify vessels according to their functional class. Basic LR models the odds ratio of probability $0 \leq \pi \leq 1$ for y=1 as

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{i=1}^{N_v}\beta_i X_i = \boldsymbol{\beta}\cdot\boldsymbol{X} \tag{4}$$

where $\boldsymbol{X}$ is a set of N_v independent variables (alternatively called covariates or predictors), and $\boldsymbol{\beta}$ is a vector of coefficients. In this application, the predictors were the difference between the AIS-reported values of draft and beam and those predicted from (1). Inverting (4) yields the probability

$$\pi(y = 1|\boldsymbol{X}) = \frac{\exp(\boldsymbol{\beta}\cdot\boldsymbol{X})}{1 + \exp(\boldsymbol{\beta}\cdot\boldsymbol{X})} \tag{5}$$

In practice, a set of data $\mathcal{D} = \{\boldsymbol{X}, y\}$ of index $k = 1,\ldots,n$, is divided according to the value of y into two sets of size n_0 and n_1, respectively. The $\boldsymbol{\beta}$ are then determined, usually by maximizing the log-likelihood function

$$\arg\max_{\beta} \sum_{i=1}^{n} [y_i \log \pi_i + (1 - y_i)(1 - \log \pi_i)] \tag{6}$$

where the π_i carry the $\boldsymbol{\beta}$-dependence. A common issue that must often be addressed is unbalanced data, when $n_0 \gg n_1$, or the reverse, which can bias (6), resulting in poor estimates of the coefficients and degrade the fidelity of the model. See King and Zeng (2001) and Salas-Eljatib et al. (2018) for additional details. A similar issue arises when $\mathcal{D}$ contains clusters around one or more points in the data space (Merlo et al., 2006). Defining a subset of $\mathcal{D}$ using random subsampling is often employed in the case of unbalanced data, whereas Tomek Link, Synthetic Minority Oversampling, and Neighborhood Cleaning are common solutions to clustered data (Elhassan and Aljurf, 2016; Guo and Wei, 2019). In this study, random subsampling was used to address the data imbalance as there was little clustering in the data.

LR can also be used to represent the probabilistic choice between two distinct quantities based on the same independent variables. Here we examined the probability of vessels being in class c compared to the probability of the vessel belonging to any other class c',

$$\ln\left(\frac{\pi(c|\,\delta,\gamma)}{\pi(c'|\,\delta,\gamma)}\right) = \boldsymbol{\beta}_c \cdot \boldsymbol{X} \tag{7}$$

given the parameters δ and γ related to the residuals of (1), defined below. Similar “one-vs-rest” classification schemes (Bisong, 2019) have been applied to a variety of labels, including cancer diagnosis (Zhu and Hastie, 2004), handwriting analysis (Klimaszewski, 2015), and astronomical redshift (Stivaktakis et al., 2019).

The result of LR (5) is a real value in the range [0,1]. A threshold probability value is typically defined such that if $\pi < \pi_0$ then y is considered to equal 0, and y=1 when $\pi \geq \pi_0$. The most common selection for this threshold is π_0=0.5, but this is somewhat arbitrary. In this study π_0 was allowed to vary, and the resulting changes in the rate of true positive (TPR) vessel classifications, and the rate of false Positive (FPR) classifications were found for each class, assuming the AIS-reported vessel type was correct. These were used to construct Receiver Operating Characteristic (ROC) curves, defined as TPR vs. FPR on the unit square, and the Area Under Curve (AUC) of the ROC (Fawcett, 2006; Huang and Ling, 2005). ROC curves in proximity to the upper-left corner of the domain (high TPR, low FPR) are have higher fidelity. Values of AUC range from 0 to 1, with the higher values generally considered an indication of an accurate classification scheme. An AUC value of 0.5 indicates even probability of TP and FP, essentially random classification.

3. Results

The vessel class with the highest number of unique vessels was the Recreational class (Table 1). From 2015 to 2019 the total number of Recreational vessels roughly doubled after Marine Cadastre started reporting class-B AIS in 2018. The number of reported Fishing vessels spiked in 2016. This is also likely to again be due to changes in reporting. During that same time period the number of Tanker vessels decreased by almost 1/3, but this was likely due to a change in operations, not reporting. Overall, the total number of vessels roughly doubled (Table 2), with most of that due to an increase in the number of small (L<30 m) vessels. The total number of larger vessels showed a weak trend, decreasing from 2520 in 2016 to 2371 in 2019.

3.1 Hull Dimensions

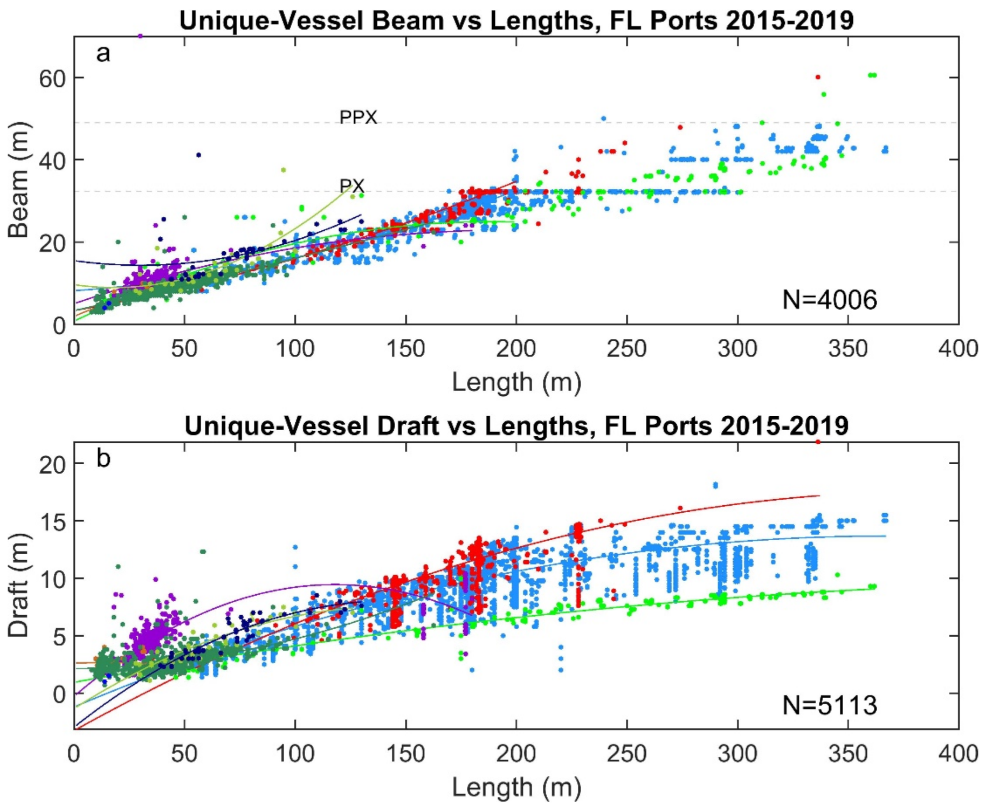

Figure 2. (a) Unique-vessel beam vs length, by functional class (Table 1). Dashed lines indicate Panamax beam (PX) and Post-Panamax (PPX) beam sizes. Number of vessels (N) with both L, Y>0 and 0<B≤200 m is indicated. (b) Unique-vessel draft vs length, coded by functional class. Solid lines are quadratic fits for each class. Number of vessels with L, D, B, Y>0 is indicated.

Scatter plots of the hull dimensions illustrate how the dependence of vessel beam $B(L)$ and draft $D(L)$ varied by class (Figure 2), with both generally increasing with L. There was little class difference apparent for $B(L)$. For $L <$ ~200 m, B increased roughly linearly with L for all classes. Tug and Supply class vessels had the largest beam for L < 50 m, and 50 m<L< 100 m, respectively. Larger vessels (L > ~200 m) often had limited B by design. Many of these ships have been in operation for years and were built to pass through the Panama Canal, so had B capped at the "Panamax" limit of 32.31 m, in place since the opening of the canal in 1914. Vessels at or just below this beam size were found for, roughly, 170 m <L< 300 m. In 2016 the Panama Canal expanded the maximum permissible vessel beam to 51.25 m ("PostPanamax"). Ships with B > 32 m were exclusively Passenger, Tanker, and

Cargo class with L>200 m (Figure 2), though their voyage may not have necessarily included passage through the Panama Canal.

In contrast, $D(L)$ showed more separation by class (Figure 2). Tugs had the highest nominal rate of increasing D with L, and Passenger class the lowest, though Tugs were generally limited to L<~60 m. The Cargo class included the largest L reported. Tankers often had the highest D for a given L in their range, and Cargo class generally had drafts between those of Tankers and Passenger classes for $L \gtrsim$100 m. There was less apparent distinction between the classes in the range $D \lesssim 3$ m and $L \lesssim 60$ m.

3.2 Static Errors

The quality of the static data was measured by the number of vessels with missing or conflicting static values. The unique MMSIs in the study region each year were first identified. Then the reported values for the static variables of every vessel were determined each year. All vessels examined reported a single value for L, B, or Y. About 1-10% of all vessels, depending on the year, had multiple D values (Table 2), with up to 24 unique values for a single vessel in one year. A high percentage of vessels reported zero (or were missing) values for L, B, Y, or D, with D having the highest rate of zero, reaching ~80% in 2019. The number of vessels reporting at least one $D = 0$ and at least one $D > 0$ over the same year fluctuated, peaking in 2016 at just under 12% of vessels, and declining to ~1% in 2019. These rapid changes in quality may be indicative of changes to the handling of the AIS data, rather than changes in the raw AIS data themselves. The static error rates were lower for vessels with $L > 30$ m (Table 3). For example, only about 10-20% of vessels failed to report any D value in a given year.

Individual AIS reports with a missing or zero static value, and a nonzero value for the same vessel in another report, can be easily corrected by filling the missing value with the nonzero value. Most static values were unchanging, so a single non-zero value would be sufficient. However, in the case where multiple D are available, the choice needs to be judicious, or some level of acceptable error needs to be determined based on the application.

Those vessels entirely missing a static variable, or those without an historical record on which to draw, require another method for correction. A simple method for estimating $D(L)$ was therefore tested. The first step was to identify those MMSI with a complete set of static variables, and then implement (1) with n=2 for each class of ships with at least 10 unique (L, D) value pairs per class. All classes except Enforcement class met these qualifications. The minimum count of ten was somewhat arbitrary, but helped avoid fitting too sparsely represented classes.

3.3 Polynomial Correction

Beam size could only reasonably be represented by a polynomial for $L <$ ~200 m, above which Panamax restrictions dominated the distribution of vessel beam sizes (Figure 2). Just over 4000 total vessels with complete static AIS data were partitioned by functional class and their beam estimated using (1). The most abundant vessel class was Cargo, with about 2200 unique vessels identified (Table 4). Tanker, Passenger, and Tug classes all had several hundred unique vessels; all other classes contained a few dozen unique vessels.

Differences between the estimated beam (B_2) and the beam from AIS (B) were found for each year, and were generally small. For example, in 2017, 66% of the residual values $\gamma = |B_2 - B| <$ 1 m, and 89% were < 2 m (Figure 3). A smaller number of much larger γ were found in all classes. The relative difference $r_B = \gamma/B$ was usually higher for smaller (L<~75 m) vessels. With the exception of a few outliers, the highest r_B was ~0.8-1.0, found near L~10 m. Overall, about 63% of the values had γ/B<0.1, and about 90% had γ/B<0.25. These percentages decreased in 2018 and 2019 to about 40% and 75%, respectively, with the increased number of smaller Recreational vessels in the database.

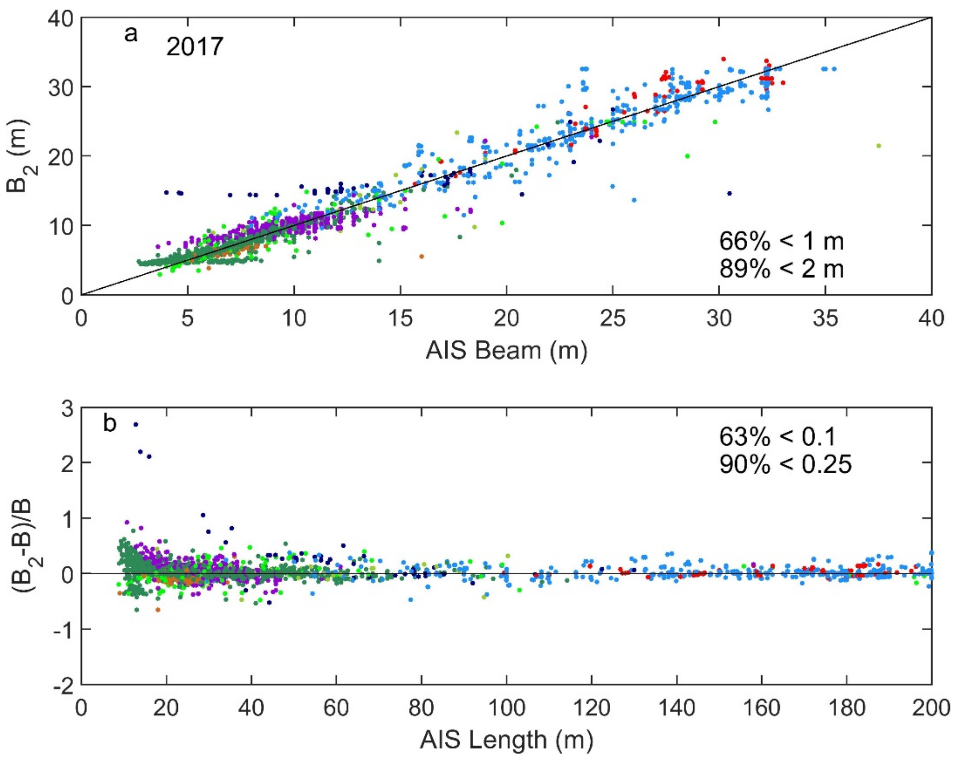

Figure 3. (a) Polynomial predicted draft (B_2) vs AIS (from 2017) reported draft. Black line indicates the identify; (b) relative difference of estimated and reported beam vs vessel length from AIS.

The resulting beam RMSD for all years was highest (4.6 m) for Supply class, with a MRAD 0.15 (Table 4). The smallest RMSD was slightly above 1 m, found for the Fishing class, though because these vessels are smaller (maximum L~40 m), their MRAD was 0.136. The smallest MRAD was found for the Tanker class at just under 0.06.

Differences between D_2 and the AIS-reported D, followed a similar pattern. About 70% of residuals $\delta = |D_2 - D|$ values were < 1 m and 90% were < 2 m (Figure 4). The majority (~61%) of the relative differences δ/D were < 0.1. This was fairly consistent for the other years. The draft RMSD for all years was largest for Cargo and Tanker ships, at ~1.4 m. The higher number of Cargo, Tanker, and Passenger vessels in the draft error analysis than that for beam was due to the inclusion of L>200 m vessels in the former. Passengers ships had the lowest RMSD, just under 0.6 m. Most of the draft MRAD were about 0.1-0.2, for all classes.

The polynomials (1) by definition yielded values of vessel length (L_{ex}) that defined extrema values of B or D, where the rate of change of the modeled variable changes sign. This was an acceptable feature for L_{ex} outside the range of reported L values, or when L_{ex} was near the range endpoints. Most instances of L_{ex} were acceptable, but there were some exceptions. The most obvious exception being the D_2 estimate for the Tug class, where L_{ex} ~118 m, with Tug lengths ranging 20 <L< 180 m. This condition was associated with a gap in the Tug class between ~70 <L<150 m, with tugs of both larger and smaller L. Tugs with L above this gap may be more appropriately placed into a different class (e.g., Cargo), as they were generally pusher or articulated tug-barge vessels. Future studies involving vessel classification should carefully consider both vessel type and function.

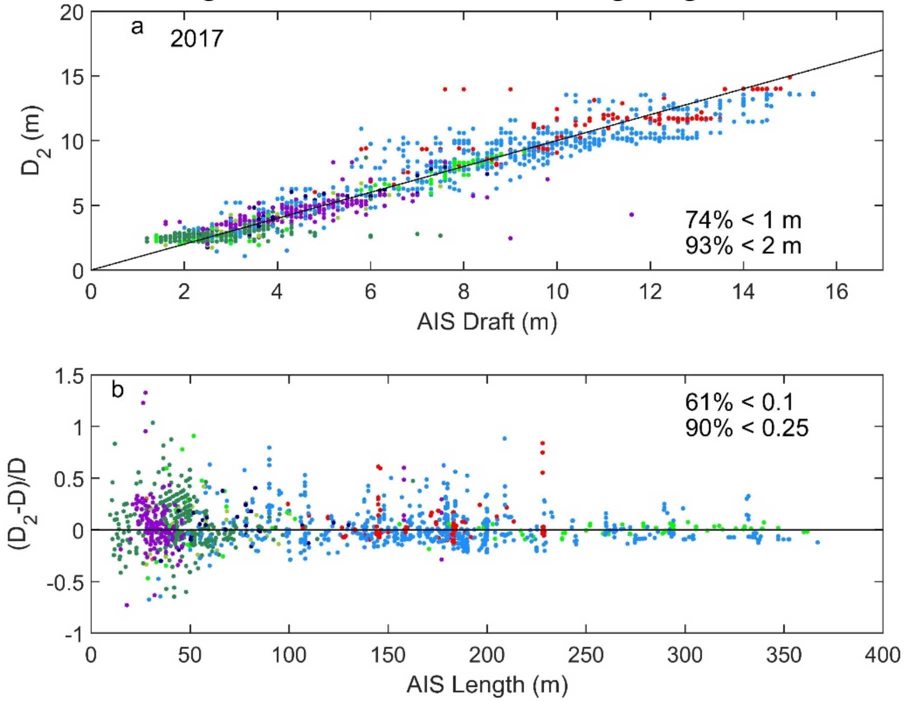

Figure 4. Same as Figure 3 but for vessel draft.

3.4 Classification

LR was applied as a tool for predicting the class c based on each set of (L, B, D) from AIS. Each class was treated separately, and the c' (7) was then the set of all reports not belonging to c. The polynomial models (Table 4) for B (with L< 200 m) and D (1) for the particular c were used to calculate residuals γ and δ for all the AIS reports. The hypothesis being that vessels in c will be distinguished by lower residuals compared to those from c', and therefore could be usefully modeled with LR. Reports in c were assigned y=1, and the rest y=0. The change in the distribution of vessel beam at L~200 m motivated the LR models be developed in 4 cases: Case 1 included all AIS reports (0 < L< 400 m); case 2 was for 200 < L< 400 m; cases 3 and 4 were for 0 < L< 200 m. Cases 1-3 used only δ as a predictor, whereas case 4 used both δ and γ as predictors.

Initial attempts to build the LR models from these data frequently yielded p-values for the $\boldsymbol{\beta}$ coefficients well above 0.05, and were therefore not considered useful. This was attributed to the unbalanced nature of the data, that is, when the ratio of the number of vessel reports in the two sets $n_c/n_{c'}$ was very large or very small. To eliminate this effect, the larger of the two sets were randomly subsampled (without replacement) so that $n_c = n_{c'}$ and the LR recalculated. Rebalancing consistently yielded $p<0.05$ for the $\boldsymbol{\beta}$ values. Independent subsampling of the original data was repeated 200 times, which was sufficient for the mean coefficient values, denoted $\bar{\beta}_c$, to converge (e.g., Figures 5, 6). The coefficients of all the iterations were stored, from which 95% confidence intervals were computed directly from the distribution of the β_c. The probability of a vessel being correctly identified to be in the "one" class versus "the rest" was then defined as when $\pi(c|\delta,\gamma) \geq \pi_0(\bar{\beta}_c)$.

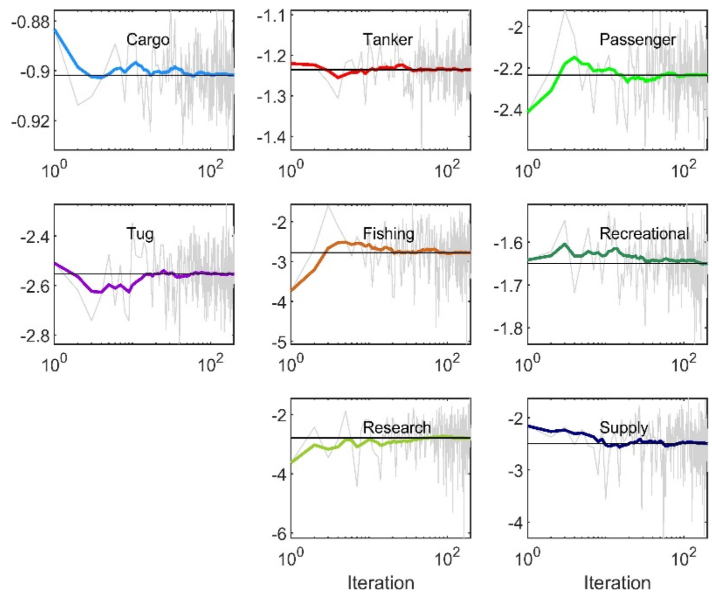

Figure 5. Case 1 constant LR coefficient for each iteration (grey), the mean value (black) and the cumulative average, for each vessel class indicated.

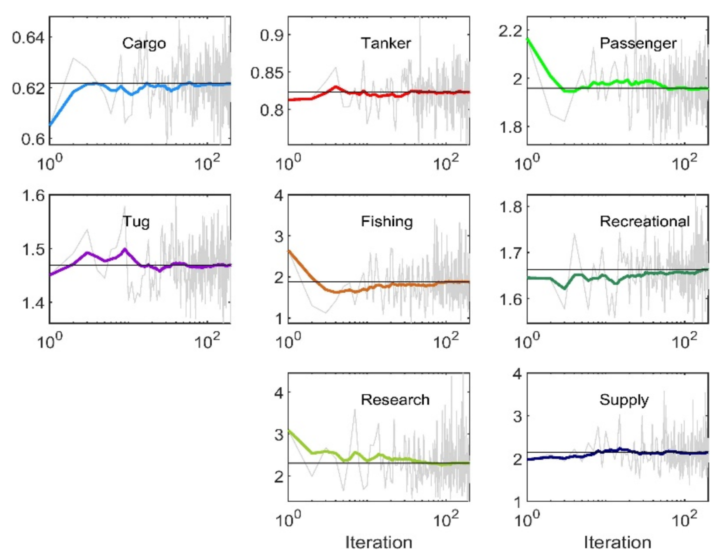

Figure 6. Same as Fig 5 but for the LR coefficient associated with the Draft variable.

The model was tested using a limited version of k-order cross-validation methods (Aly, 2020; Pala and Atici, 2019). The data was divided into k=10 sections of equal length. For each class in each case, the indices within c and those within c' were divided separately due to the imbalance of the data. The 62 mean coefficients computed from the k subsets were generally close to those computed using all the data. Relative differences between the full-data coefficients and the mean of the k data coefficients were almost all small. For 57 coefficients, the relative difference was <5%, with the majority being <1%. The largest exceptions to this all occurred in Case 4, where the mean coefficient for B was about twice that obtained in the full-data case. The second largest deviation was for Fishing vessels, where the coefficient for D differed from the full-data case by 10%. The relative difference of coefficients for Research vessels'

L, D, B were 6%, 7%, and 6%, respectively. There were a small number of instances where the maximum likelihood coefficient calculation converged to a value very different from those obtained in almost all other calculations for the same case and class. Coefficient values more than 10 times the value obtained using all the data were discarded.

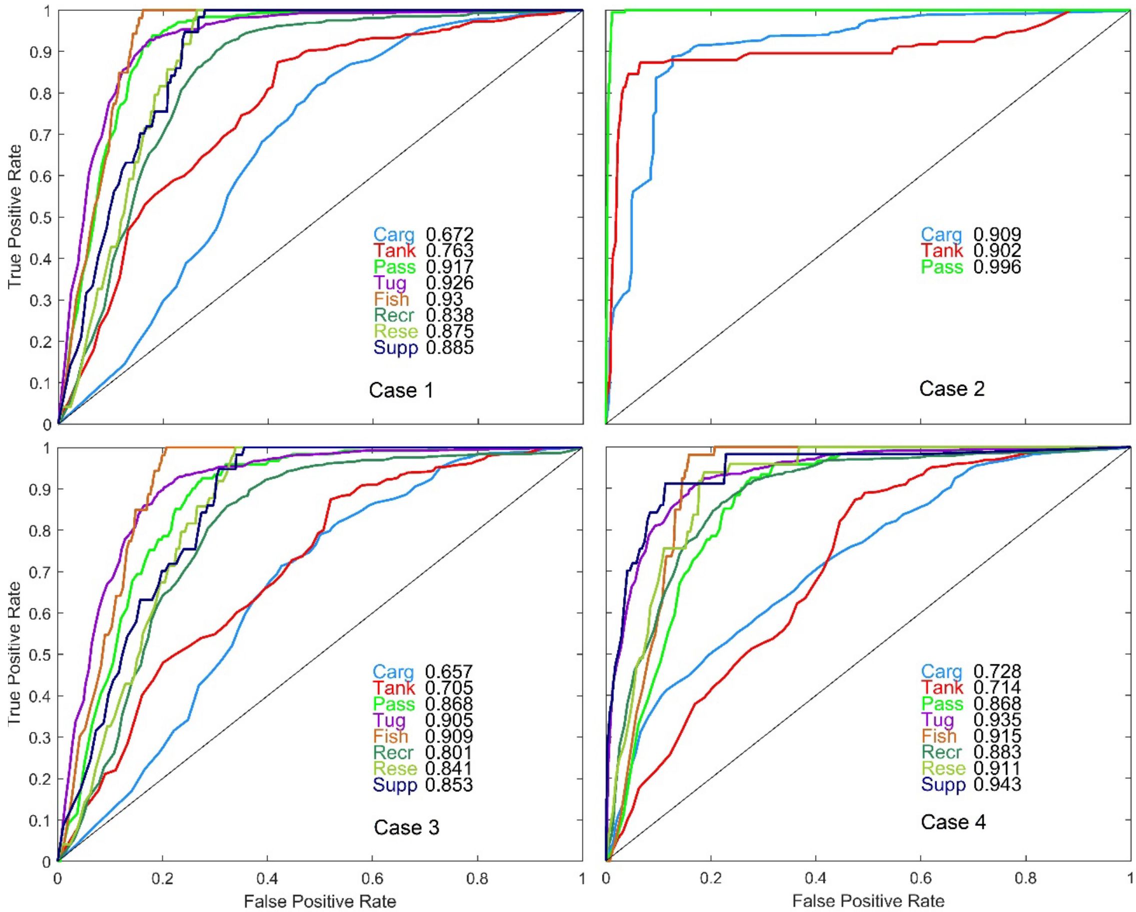

Figure 7. ROC curves and their AUC values for the classes (Table 1) and cases indicated. The diagonal indicates the random classification case.

For all classes and cases ROC curves (Figure 7) were above the random diagonal, indicating the results of the classification scheme was better than random. Case 1 (all vessels) had the highest ROC curves and AUC values for Fishing, Tug, and Passenger classes, all which had an AUC > 0.9. Overall, Case 2 (large vessels) had the best results, with steeply rising curves at low FP, and AUC values above 0.9. Case 3 yielded the lowest AUC scores for all classes, with Cargo and Tanker classes being the worst performing with AUC of 0.657 and 0.705, respectively. All other classes in this case had AUC > 0.8. The inclusion of a second predictor variable (γ) in Case 4 raised all AUC scores compared to case 3, with Supply class rising by 0.09. Relatively large increases also occurred in the Cargo, Recreational, and Research classes. The lowest AUC in Case 4 was 0.714 for the Tanker class. The regression model developed for Case 1 can be applied to any AIS transmission, assuming sufficient statics are available. Application of the other Cases would depend on the static values (Figure 8).

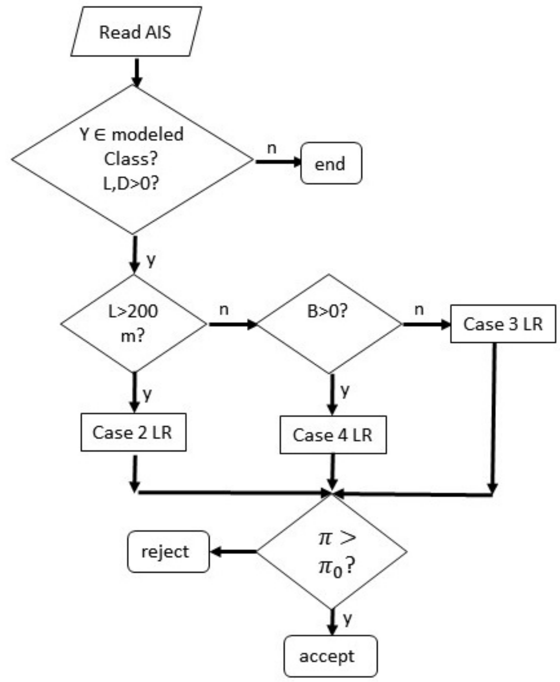

Figure 8. Schematic of vessel classification algorithm for different Cases of vessel dimensions as described in the text.

One way to explore the reliability of a classification scheme is to examine the differing characteristics of its least- and most-confident predictions. Here, the True Positives in Case 1 (all vessels) were examined. Vessel reports classified as a TP for a high π_0 were more likely to be correctly classified, and those satisfying low π_0 – but not moderate or high π_0 – were more likely to be incorrectly classified. There were two primary reasons a vessel report might have been included in the low confidence group: 1) the vessel was misclassified in the AIS report, so as expected the algorithm rated it with low probability of being a TP, and 2) a deficiency in the classification scheme, such as in the development of the classes or misapplication of the algorithm. Examining the characteristics of the two groups helped identify limitations of both the data set and the classification scheme.

The two sets of AIS reports were identified such that they exclusively define a TPR > 0.95 or < 0.05 (Figure 7), indicating low and high confidence in their classification, respectively. The π_0 at which these occurred varied by class. Summing over all classes, there were 487 reports in the low confidence group, and 210 in the high confidence group. Static variables for these vessels were then scraped from a third-party vessel traffic website, and the classification obtained was compared to that provided in each AIS report. In the low confidence group, 53 (12%) classifications did not match. In the high confidence group, 5 (2%) classification inconsistencies were found. A null hypothesis that these two ratios are the same was rejected based on both chi-squared and Fischer's exact test well above the 99% confidence level. This further demonstrated the method ability to detect misclassified vessels. However, the majority of reports in the low confidence scheme were not misclassifications but large difference δ between predicted and reported draft.

The low confidence group had an average $\delta / D_2 = 0.42$, compared to 0.003 for the high confidence group, indicating vessels in the former group departed from the polynomial estimated values much more than those in the latter group. The majority of the low confidence group was comprised of a total of 368 entries from Cargo and Tanker vessels, which, as noted above, can have a wide variation in draft during their course of operations. The LR algorithm flagged these with low confidence, and can be used to identify vessels operating near their extreme drafts.

Future development should account for such normal variations of draft. The low confidence group also contained 60 Recreational and 36 Tug entries, neither of which undergo significant changes in draft during normal operations. Four of the draft values reported by the Recreational ships were roughly a factor of 3 larger than the value obtained from the third-party website, but with equal L values, suggesting these draft entries may have been in feet instead of meters. All but three of the Recreational reports had L<60 m, putting them in the area of high draft variation within their class (Figures 3 and 4). For the Tugs, 18 reported relatively small length (L<50 m), of which 13 were deeply drafted (6–10 m) pusher tugs that generally operate coupled to much longer vessels or barges. The remaining 18 Tugs reports were also deeply draft pusher or articulated tugs reporting L >150 m.

4. Discussion

Erroneous or missing AIS static values are not unusual. For example, in 2019 about 21% of vessels with length>30 m operating near large Florida ports did not transmit their draft through AIS, and about 7.5% did not transmit their beam (Table 3), introducing errors in any analysis, algorithm, or operation based on the presumption the values are accurate. Here novel schemes for detecting and potentially correcting vessel beam, draft, and classification have been explored that rely on the partition of AIS types into 9 vessel classes, though not all vessels fit into the defined classes, and some vessels may better fit a class different than one indicated by their AIS type. Examples of the latter were articulated tug-barge vessels that might be more accurately classified as Tanker or Cargo vessels as their function and design is very different than the more typical (and smaller) tugs that are used to support the maneuvering of other, larger vessels. The LR classification scheme in this study demonstrated skill in verifying AIS-transmitted classification, detecting incorrectly classified vessels, and flagging those with incorrect draft or operating near an extreme draft.

The cornerstone of the methods presented here was the creation of independent, low-order polynomial relations between vessel length and the beam and draft for each vessel class. For both B and D, over 60% of the relative differences between predicted (1) and AIS-reported values were less than 0.1, and over 90% had relative errors < 0.25 (Figures 3 and 4). For many classes, these differences were due to intra-class variations in hull design, particularly for smaller Tugs and Recreational vessels. For Cargo and Tanker classes, changing deadweight was also a contributing factor to these differences. To compensate for these variations, it would be useful to create bands of values rather than simple polynomial relations. Varying the coefficients in (1) within their 95% confidence intervals would be one method to quickly develop these ranges. Using a band of acceptable values for B and D would also likely result in increased π_0 of the True Positive rates (Figure 7).

Improvement of the classification scheme might also be achieved by the addition of dynamic variables such as speed, location, and turning rate, as predictor variables. For instance, it is likely a petroleum tanker will have lower draft immediately following a port call in Florida, which is not a significant petroleum producing state. Similarly, Fishing vessels are more likely to visit and remain within certain offshore areas than, say, large Cargo vessels. These examples of distinguishing vessel behavior are not sufficient to make a class determination by themselves, but could be useful in conjunction with other variables.

The ongoing development of corrective schemes for AIS variables suggests that these data can be treated much like some other large observational data sets, with varying levels of quality analysis and control (QA/QC). NOAA has an extensive procedure for QA/QC of real-time oceanographic measurements (Hofmann and Healy, 2017), with older instrument types such as tide gauges having more robust protocols than newer instruments such as chemical sensors. Possible levels of QA/QC for AIS are outlined as follows:

Level 0: raw, decoded AIS data, directly readable in the form of text, csv, or similar formats. No correction applied.

Level 1: Vessels would be identified using their reported MMSI, and possibly their IMO number, name, and other identifying information (Winkler, 2012). Missing or suspect static variables would be replaced with values taken from the historical records of the identified vessel. The existence of such records is assumed, so this would be best applied to vessels of sufficient age to generate the proper database. This level could also include removal and correction of isolated anomalous dynamic values such as large spikes in velocity or position. Precautions would need to be implemented in cases of erroneous MMSI, when the same MMSI is reported for different vessels, or when a vessel changes its MMSI as sometimes occurs when coming under new ownership.

Level 2: Interpolative schemes would be used to fill missing static values for vessels without records sufficient to permit application of Level 1 corrections. The schemes would be developed using sets of related vessel types. The polynomial relations developed here provide an example, where vessels were organized into functional classes and the (presumably correct) length and class were used to estimate beam and draft. It would be instructive to develop these relations on much larger sets of vessels as it is possible some bias was introduced in the selection of Florida as a test bed. With a sufficient number of vessels, it may be possible to create interpolative methods for each AIS type. Other groupings of vessels might yield different results, but constraints of nautical design necessitate a limited ranges of hull geometries (Figure 2). Multi-hull designs such as catamarans and trimarans would likely need to be treated separately.

Level 3: AI/ML methods would synthesize the full AIS record, including both static and dynamic variables, of the individual vessel and other vessels, to detect and correct errors and omissions in AIS reports. Some initial steps towards developing such a set have been taken using corrected AIS position records (Masek et al., 2021). Level 3 might also include use of data beyond the AIS, such as Synthetic Aperture Radar (SAR) and optical imaging from low-orbiting satellites to determine ship class, size and speed (Purivigraipong, 2018; Riveiro et al., 2018), stationary mounted cameras, local radar, or similar instruments placed onto aircraft (Eaton et al., 2018). The addition of B to the predictor set increased the AUC values of some classes by ~0.1 (Figure 7), suggesting the addition of other predictors could further increase the accuracy of the classification scheme. The number of useful predictors is likely to be limited by the “curse of dimensionality” (Geenens, 2011) where the calculation of model parameters (e.g., β) fails to converge due to a sample space made sparse by the inclusions of too many independent variables.

The AIS provides essential information for the management and control of maritime operations, is widely used in retrospective studies of vessel activities, and in the ongoing transformation of the maritime industry by artificial intelligence and related technologies (Artikis and Zissis, 2021; de la Peña Zarzuelo et al., 2020; Plaza-Hernández et al., 2020). The methods described here provide a new method for detecting and potentially updating some static AIS variables, supporting these efforts.

Acknowledgments

This research was made possible in part through a cooperative agreement between NOAA’s Office of Coast Survey and the University of South Florida through the Center for Ocean Mapping and Innovative Technologies (COMIT) (Award #NA20NOS4000227), the NOAA IOOS Program Office as sub-awards through the Southeast Coastal Ocean Observing Regional Association (Award #NA16NOS0120028, Sub-Award #IOOS.16(028)USF.ML.OBS.1 and IOOS.16(028).USF.ML.GAPS MINIPROPOSAL.5), and the Gulf of Mexico Coastal Ocean Observing System (Award #NA16NOS0120018, Sub-Award #02-S160275), and from the Greater Tampa Bay Marine Advisory Council-PORTS, Inc. (Award #2500-1066-00).

LITERATURE CITED

Aly, H.H., 2020. A novel approach for harmonic tidal currents constitutions forecasting using hybrid intelligent models based on clustering methodologies. Renewable Energy 147, 1554-1564.

Artikis, A., Zissis, D., 2021. Guide to Maritime Informatics. Springer Nature.

Bertram, V., Schneekluth, H., 1998. Ship design for efficiency and economy. Elsevier

Bisong, E., 2019. Logistic Regression, Building Machine Learning and Deep Learning Models on Google Cloud Platform. Springer, pp. 243-250.

Bošnjak, R., Šimunović, L., Kavran, Z., 2012. Automatic identification system in maritime traffic and error analysis. Transactions on maritime science 1 (02), 77-84.

Chen, P., Shi, G., Liu, S., Gao, M., 2018. Pattern Knowledge Discovery of Ship Collision Avoidance based on AIS Data Analysis. International Journal of Performability Engineering 14 (10).

de la Peña Zarzuelo, I., Soeane, M.J.F., Bermúdez, B.L., 2020. Industry 4.0 in the port and maritime industry: A literature review. Journal of Industrial Information Integration, 100173.
Demšar, U., Virrantaus, K., 2010. Space–time density of trajectories: exploring spatio-temporal patterns in movement data. International Journal of Geographical Information Science 24 (10), 1527-1542.
Dobrkovic, A., Iacob, M.-E., van Hillegersberg, J., Mes, M.R., Glandrup, M., 2016. Towards an approach for long term AIS-based prediction of vessel arrival times, Logistics and Supply Chain Innovation. Springer, pp. 281-294.
Eaton, R.S., German, S., Balasuriya, A., 2018. Maritime Border Security using Sensors, Processing, and Platforms to Detect Dark Vessels, 2018 IEEE International Symposium on Technologies for Homeland Security (HST). IEEE, pp. 1-5.
Elhassan, T., Aljurf, M., 2016. Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. Global J Technol Optim S 1.
Emmens, T., Amrit, C., Abdi, A., Ghosh, M., 2021. The promises and perils of Automatic Identification System data. Expert Systems with Applications 178, 114975.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8), 861-874.
Geenens, G., 2011. Curse of dimensionality and related issues in nonparametric functional regression. Statistics Surveys 5, 30-43.
Guo, H., Wei, T., 2019. Logistic regression for imbalanced learning based on clustering. International Journal of Computational Science and Engineering 18 (1), 54-64.
Guo, S., Mou, J., Chen, L., Chen, P., 2021. An Anomaly Detection Method for AIS Trajectory Based on Kinematic Interpolation. Journal of Marine Science and Engineering 9 (6), 609.
Harati-Mokhtari, A., Wall, A., Brooks, P., Wang, J., 2007. Automatic Identification System (AIS): Data Reliability and Human Error Implications. Journal of Navigation 60 (3), 373-389.
Harre, I., 2000. AIS adding new quality to VTS systems. The Journal of Navigation 53 (3), 527-539.
Hilbe, J.M., 2016. Practical guide to logistic regression. CRC Press.
Hofmann, C., Healy, J., 2017. Real-time quality control experiences using QARTOD in Australian ports. Australasian Coasts & Ports 2017: Working with Nature, 612.
Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied logistic regression. John Wiley & Sons.
Huang, J., Ling, C.X., 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering 17 (3), 299-310.
Jahn, C., Scheidweiler, T., 2018. Port Call Optimization by Estimating Ships' Time of Arrival, International Conference on Dynamics in Logistics. Springer, pp. 172-177.
King, G., Zeng, L., 2001. Logistic regression in rare events data. Political analysis 9 (2), 137-163.
Klimaszewski, J., 2015. A comparison of regularization techniques in the classification of handwritten digits. Journal of Theoretical and Applied Computer Science 9 (4), 3-7.
Lim, G.J., Cho, J., Bora, S., Biobaku, T., Parsaei, H., 2018. Models and computational algorithms for maritime risk analysis: a review. Annals of Operations Research, 1-22.
Liu, B., 2015. Maritime Traffic Anomaly Detection from Ais Satellite Data in Near Port Regions, Computer Science. Dalhousie University, p. 91.
Masek, M., Lam, C.P., Rybicki, T., Snell, J., Wheat, D., Kelly, L., Smith-Gander, C., 2021. The open maritime traffic analysis dataset.
Merlo, J., Chaix, B., Ohlsson, H., Beckman, A., Johnell, K., Hjerpe, P., Råstam, L., Larsen, K., 2006. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. Journal of Epidemiology & Community Health 60 (4), 290-297.
Meyers, S.D., Luther, M.E., Ringuet, S., Raulerson, G., Sherwood, E., Conrad, K., Basili, G., 2020. Characterizing Vessel Traffic Using the AIS: A Case Study in Florida's Largest Estuary. Journal of Waterway, Port, Coastal, and Ocean Engineering 146 (5), 05020005.
Mitchell, K.N., Scully, B., 2014. Waterway performance monitoring with automatic identification system data. Transportation Research Record 2426 (1), 20-26.

Murk, D.W., 1999. Vessel traffic management: a new philosophy, Proceedings of the Marine Safety Council, Washington, DC.

Oh, J.-Y., Kim, H.-J., Park, S.-K., 2018. Detection of ship movement anomaly using AIS data: a study. Journal of Navigation and Port Research 42 (4), 277-282.

Pala, Z., Atici, R., 2019. Forecasting Sunspot Time Series Using Deep Learning Methods. Solar Physics 294 (5), 50.

Papanikolaou, A., 2014. Ship design: methodologies of preliminary design. Springer.

Plaza-Hernández, M., Gil-González, A.B., Rodríguez-González, S., Prieto-Tejedor, J., Corchado-Rodríguez, J.M., 2020. Integration of IoT Technologies in the Maritime Industry, International Symposium on Distributed Computing and Artificial Intelligence. Springer, pp. 107-115.

Purivigraipong, S., 2018. Review of Satellite-Based AIS for Monitoring Maritime Fisheries. ENGINEERING TRANSACTIONS 21 (1), 44.

Riveiro, M., Pallotta, G., Vespe, M., 2018. Maritime anomaly detection: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8 (5), e1266.

Robards, M., Silber, G., Adams, J., Arroyo, J., Lorenzini, D., Schwehr, K., Amos, J., 2016. Conservation science and policy applications of the marine vessel Automatic Identification System (AIS)—a review. Bulletin of Marine Science 92 (1), 75-103.

Rong, H., Teixeira, A., Soares, C.G., 2019. Ship trajectory uncertainty prediction based on a Gaussian Process model. Ocean Engineering 182, 499-511.

Salas-Eljatib, C., Fuentes-Ramirez, A., Gregoire, T.G., Altamirano, A., Yaitul, V., 2018. A study on the effects of unbalanced data when fitting logistic regression models in ecology. Ecological Indicators 85, 502-508.

Shelmerdine, R.L., 2015. Teasing out the detail: How our understanding of marine AIS data can better inform industries, developments, and planning. Marine Policy 54, 17-25.

Sheng, K., Liu, Z., Zhou, D., He, A., Feng, C., 2018. Research on Ship Classification Based on Trajectory Features. Journal of Navigation 71 (1), 100-116.

Sidibé, A., Shu, G., 2017. Study of automatic anomalous behaviour detection techniques for maritime vessels. The Journal of Navigation 70 (4), 847-858.

Silveira, P., Teixeira, A., Soares, C.G., 2013. Use of AIS data to characterise marine traffic patterns and ship collision risk off the coast of Portugal. The Journal of Navigation 66 (6), 879-898.

Smestad, B.B., Asbjørnslett, B.E., Rødseth, Ø.J., 2017. Expanding the possibilities of ais data with heuristics.

Son, G.M., Choi, W.J., Baek, J.E., Shin, D.W., Yang, C.S., 2022. Approach to Classifying Ship Types from AIS Data Using DNN and CNN, ISRS 2022 (International Symposium on Remote Sensing 2022). ISRS, pp. 242-244.

Steidel, M., Lamm, A., Feuerstack, S., Hahn, A., 2019. Correcting the Destination Information in Automatic Identification System Messages. Springer International Publishing, Cham, pp. 496-507.

Stivaktakis, R., Tsagkatakis, G., Moraes, B., Abdalla, F., Starck, J.-L., Tsakalides, P., 2019. Convolutional neural networks for spectroscopic redshift estimation on euclid data. IEEE Transactions on Big Data 6 (3), 460-476.

Sun, Y., Chen, X., Jun, L., Zhao, J., Hu, Q., Fang, X., Yan, Y., 2021. Ship trajectory cleansing and prediction with historical ais data using an ensemble ann framework. Int. J. Innov. Comput. Inf. Control 17, 443-459.

Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., Huang, G.-B., 2017. Exploiting AIS data for intelligent maritime navigation: A comprehensive survey from data to methodology. IEEE Transactions on Intelligent Transportation Systems 19 (5), 1559-1582.

Wang, Y., Yang, L., Song, X., Li, X., 2021. Ship classification based on random forest using static information from AIS data, Journal of Physics: Conference Series. IOP Publishing, p. 012072.

Wang, Y., Zhang, J., Chen, X., Chu, X., Yan, X., 2013. A spatial–temporal forensic analysis for inland–water ship collisions using AIS data. Safety Science 57, 187-202.

Winkler, D., 2012. AIS Data Quality and the Authoritative Vessel Identification Service (AVIS). National GMDSS Implementation Task Force, Arlington, VA.

Xin, X., Liu, K., Yang, X., Yuan, Z., Zhang, J., 2019. A simulation model for ship navigation in the "Xiazhimen" waterway based on statistical analysis of AIS data. Ocean Engineering 180, 279-289.

Yang, D., Wu, L., Wang, S., Jia, H., Li, K.X., 2019. How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications. Transport Reviews, 1-19.

Zhang, P., Zhu, D.-x., Leng, W.-h., 2008. Parametric approach to design of hull forms. Journal of Hydrodynamics, Ser. B 20 (6), 804-810.

Zhao, L., Shi, G., Yang, J., 2018. Ship trajectories pre-processing based on AIS data. The Journal of Navigation 71 (5), 1210-1230.

Zhou, Y., Daamen, W., Vellinga, T., Hoogendoorn, S., 2019. Review of maritime traffic models from vessel behavior modeling perspective. Transportation Research Part C: Emerging Technologies 105, 323-345.

Zhu, J., Hastie, T., 2004. Classification of gene microarrays by penalized logistic regression. Biostatistics 5 (3), 427-443.

Figure Captions

Figure 1. Map of peninsular Florida. The 5 largest ports are indicated.

Figure 2. (a) Unique-vessel beam vs length, by functional class (Table 1). Dashed lines indicate Panamax beam (PX) and Post-Panamax (PPX) beam sizes. Number of vessels (N) with both L, Y>0 and 0< B ≤200 m is indicated. (b) Unique-vessel draft vs length, coded by functional class. Solid lines are quadratic fits for each class. Number of vessels with L, D, B, Y>0 is indicated.

Figure 3. (a) Polynomial predicted draft (B_2) vs AIS (from 2017) reported draft. Black line indicates the identify; (b) relative difference of estimated and reported beam vs vessel length from AIS.

Figure 4. Same as Figure 3 but for vessel draft.

Figure 5. Case 1 constant LR coefficient for each iteration (grey), the mean value (black) and the cumulative average, for each vessel class indicated.

Figure 6. Same as Fig 6 but for the LR coefficient associated with the Draft variable.

Figure 7. ROC curves and their AUC values for the classes (Table 1) and cases indicated. The diagonal indicates the random classification case.

Figure 8. Schematic of vessel classification algorithm for different Cases of vessel dimensions as described in the text.