Improving Maritime Shipping Emissions Estimates Using Machine Learning

April 6, 2024

Abstract

We employ machine learning to improve upon engineering estimates of carbon dioxide emissions from maritime shipping. Traditional estimates rely on engineering approximations that may not entirely capture actual fuel use. We match reported annual ship-level emissions from a European Union emissions reporting program with tracking data and technical characteristics for the global fleet of dry bulk ships. Following industry standard procedures, we calculate engineering estimates of annual ship-level emissions and include these values as a predictor. We train various machine learning algorithms on reported fuel consumption and achieve high out-of-sample prediction accuracy.

1 Introduction

The most recent official estimate of greenhouse gas (GHG) emissions from maritime shipping places the sector's contribution at around 3% of global emissions (Faber et al., 2020). Since this estimate was published, various events have significantly impacted the shipping industry, including the COVID-19 pandemic and the implementation of new measures aimed at improving ship efficiency. Such events have almost certainly had immediate effects on emissions, and quantifying their magnitude can provide insights into the effectiveness of actual policy measures and how to improve them. To do so, it is essential to have accurate and timely estimates of emissions. In this paper, we develop a prediction methodology that augments the industry-standard method based on engineering relationships with machine learning in order to provide more accurate estimates. These estimates can be updated and published in real-time, providing invaluable information for policymakers, researchers, and industry stakeholders alike.

To develop and evaluate our model of shipping emissions, we leverage publicly available emissions reporting data from the Monitoring, Reporting, and Validation (MRV) program implemented by the European Union (EU). Carbon dioxide (CO_2) emissions are directly proportional to fuel consumption,[1] and we therefore take (log) reported annual ship fuel consumption as our target variable. We evaluate various machine learning algorithms, using data on both ship characteristics and ship activity as predictor variables. Crucially, we include features derived from the engineering-based fuel consumption calculations used by the International Maritime Organization (IMO). These features use hourly ship activity observations from Automatic Identification System (AIS) data, including location, speed, and draft,[2] to construct annual aggregate features in a manner that is consistent with the physics that determine fuel use. This approach allows us to leverage physics-based formulas while also flexibly allowing for deviations from them due to factors that are unaccounted for or inaccurately measured.

Using our hybrid approach, we are able to achieve a high degree of accuracy in predicting fuel consumption. For our data set comprising the global fleet of dry bulk ships, our best-performing model achieves an out-of-sample coefficient of determination (R^2) of 0.954 and mean absolute percentage error (MAPE) of 10.9%.[3] This is a significant improvement over a purely calculation-based approach, which achieves an R^2 of 0.58 and an MAPE of 20.4%. We evaluate both linear models (ordinary least squares (OLS), lasso, and ridge) and tree-based models (random forest, gradient boosting Regression, and CatBoost). While ridge regression performs the best in terms of R^2 for our hold-out sample, we find that all models perform quite well, with even the worst-performing achieving an R^2 of 0.931 (MAPE=12.3%).

The primary contribution of this work is a highly accurate methodology for predicting fuel consumption that can be used to produce timely estimates of shipping emissions under varying ship activity, such as travel speed. To our knowledge, this is the first study to directly predict fuel

[1] For example the most common fuel, heavy fuel oil (HFO), emits 3.114 kg CO_2 per kg of fuel.

[2] Draft is the vertical distance between the waterline and the bottom of the hull. It is a function of how heavily laden the ship is and affects fuel consumption.

[3] Our training set consists of data from 2019 and 2020, and our hold-out test sample is 2021 data.

consumption using a hybrid of the industry-standard bottom-up calculation method and machine learning and apply it at a large geographic scale. The closest work to ours is that of Yan et al. (2023) and Clarke et al. (2023) who both apply machine learning models to predict a ship-level annual average CO_2 emission ratio with units of mass-per-distance-travelled. They obtain total emissions by multiplying this efficiency metric by the observed or reported distance travelled. Employing this approach for prediction relies on the assumption that fuel efficiency "averages out" over various operating conditions (e.g., speed) during a year for which fuel consumption is reported. While this may provide a satisfactory approximation when ship speeds remain similar to those used to train the model, it is in general biased due to the nonlinearity of fuel consumption with speed. As such, the error may be large when attempting to predict emissions under scenarios in which ships adjust their speed of travel, for example in response to emissions regulations.

We propose an alternative prediction methodology that has the similar data requirements for prediction, but leverages engineering and physics relationships to improve prediction performance across a wider range of operating conditions. We construct a predictor variable representing annual ship-level energy use by summing theoretical hourly energy use, which is calculated with the industry-standard Admiralty formula (Faber et al., 2020, p. 64). This formula relies on ship characteristics and hourly speed and draft observations. To allow for potential inaccuracies in the functional form, we also construct slight variations on this quantity and provide these as additional predictor variables.

Our methodology balances the advantages of both calculation-based and pure machine learning approaches. Compared to other machine learning approaches, our use of engineering- and physics-based relations to engineer predictor features means we rely less on the machine learning algorithms to capture the physics of ship fuel consumption. The fact that linear models (in logs) perform similarly to non-linear, tree-based models in our analysis is consistent with this conjecture. In contrast to a purely calculation-based approach, incorporating machine learning allows the model to capture deviations from theoretical relationships, as well as incorporate further information that may have predictive power, but for which there is no theoretical foundation to suggest a functional form. One example of this is our inclusion of additional features derived from AIS tracking data that provide information on both ship behaviour and potential data errors, which is a well-known issue researchers face when using this data.

This paper is related to three strands of literature. The first of these has developed various bottom-up emissions estimation methodologies using engineering calculations applied to AIS tracking data (e.g., Jalkanen et al., 2009; Moreno-Gutiérrez et al., 2019; Olmer et al., 2017; Tvete et al., 2020). We build directly from these techniques in constructing our predictor variables. In particular, we follow the methodology used in the most recent IMO emissions report as closely as possible (Faber et al., 2020). A second strand of literature has begun to utilize annual emissions reports from the EU's MRV program to validate emissions estimates obtained with the bottom-up approach. The largest study in scope is the IMO's analysis, however only the first year of MRV reporting was available at the time (Faber et al., 2020). Subsequently, a handful of authors have provided similar analyses, albeit with very limited geographical scope (e.g., Doundoulakis and Papaefthimiou, 2022;

Hensel et al., 2020; Mannarini et al., 2020; Wu et al., 2023). In contrast, we analyse emissions from all dry bulk ships entering the EU, and provide a methodology for global estimates. Furthermore, rather than using the MRV data to simply validate predictions, we employ it to improve upon previous estimation methodologies. Lastly, a recent strand of literature applies various machine learning techniques to improve shipping emissions estimates. The majority of these use high-frequency fuel consumption data, but focus on small geographic areas and/or numbers of ships (e.g. Hu et al., 2019; Jebsen and Mathiesen, 2020; Monisha et al., 2023; Ren et al., 2022; Wang et al., 2023). Guo et al. (2022) use machine learning primarily to model the physics of weather effects on ship resistance in a computationally feasible manner, thereby extending the application of previous calculation-based model from Tvete et al. (2020). Our work employs readily available annual fuel consumption data, which allows us to study emissions on a much larger scale. As described above, our work is similar in objective to that of Yan et al. (2023) who employ gradient boosting and Clarke et al. (2023), who use random forest regression to predict ship-level fuel *efficiencies*. We directly predict fuel consumption, which allows us to leverage predictors derived from engineering calculations, and thereby obtain very high prediction accuracy for fuel consumption and emissions.

2 Data

We obtain data on the global fleet of dry bulk ships, which includes annual fuel consumption reports, hourly tracking data, and ship characteristics, jointly spanning the years 2019 to 2021.

Since 2018, fuel consumption reports are publicly available from the EU MRV program, which requires commercial ships over 5,000 gross tonnage (GT) to report their total annual fuel consumption and distance travelled for all voyages into and out of the European Economic Area (henceforth referred to as EU trips) (EU, 2015). We take the reported annual fuel consumption, in tonnes (t), for EU trips as ground truth,[4] and take the log to construct our target variable.

Ship characteristics are taken from the World Fleet Register (WFR), purchased from Clarksons Research, which includes a wide range of ship characteristics, such as ship type, size, engine power, etc. We match these characteristics to the EU MRV data using the IMO number, a unique identifier for each ship. We retain only dry bulk carriers, after which between 3,600 and 3,900 ship observations remain per year, comprising roughly 30% of the global dry bulk fleet.[5]

Ship tracking data consists of messages transmitted by Automatic Identification Systems (AIS) fitted to each ship, which rely on global positioning systems and radio transmitters to track and report the ship's activity. Our dataset is purchased from Spire and contains hourly transmissions from all dry bulk ships, recorded by both land- and satellite-based receivers. Draft observations from static AIS messages are merged to dynamic messages by hour and Maritime Mobile Service Identity (MMSI), which identifies a ship's transceiver. Draft values are input manually by the ship's

[4]The MRV program requires third-party validation of all reports to ensure accuracy. Furthermore, during the period that we study there were no binding emissions regulations and therefore no clear incentive to misreport fuel consumption.

[5]We consider dry bulk to include all ships over 10,000 DWT categorized in the WFR under Ore Carrier, Bulk Carrier, Chip Carrier, Open Hatch Carrier, Forest Product Carrier, Aggregates Carrier, Cement Carrier, Nickel Carrier, and Slurry Carrier.

crew and therefore many observations are missing. We follow a commonly-used strategy and replace each missing draft value with the last valid observation.

AIS data is subject to various sources of error, and we largely follow standard procedures to clean it, which are described in greater detail in **??**. We employ an intentionally conservative cleaning strategy, and rely on a final validation step with the MRV (described at the end of this section) to ensure that valid data is used to train our model. Throughout, we employ the haversine formula to calculate distances between location observations in an accurate and computationally feasible manner.

One common error we observe are physically infeasible changes in a ship's trajectory and/or location. Since many of these are single, anomalous observations, we first drop single data points that represent an abrupt change of direction that is infeasible for the speed at which the ship is travelling. In other cases, a ship (as identified by the MMSI), appears to jump to a new location, pursue a feasible trajectory for some time, and then jump back to its previous location. To correct for this type of error, we split each ship's trajectory whenever both the distance between observations exceeds 140 nautical miles (nm) and the implied speed (calculated as distance divided by time between observations) exceeds 25 knots. We retain either the even or odd trajectory segments depending on which set comprises the greater number of observations. The resulting data set contains trajectories for 12,716 unique MMSI.

A second important error is missing dynamic data, i.e. hours for which there is no observation for a ship. We interpolate these missing data points as per the IMO's strategy (Faber et al., 2020), creating an observation for each missing hour, assuming a constant speed and using the haversine formula to interpolate location coordinates. Speed is interpolated as the distance between consecutive locations divided by the time difference, and the previous draft value is assigned. On average, 37% of observations for each ship-year are interpolated, although this figure varies widely, with a standard deviation of 21%. This average decreases year-on-year, from 46% in 2019 to 30% in 2021.

To match the cleaned AIS tracking data to the MRV reports, it is necessary to identify which observations to attribute to EU trips, as only activity related to those trips is reported in the MRV data. To do so, we first identify port calls based on observed speed and proximity to land.[6][7] A trip is defined as a ship's movement between two port calls. As per the EU MRV regulation, any trip with at least one of its two port calls within the jurisdiction of an EU member state is considered an EU trip (EU, 2015).[8] To obtain the various annual values we use as predictors (see Section 3), we aggregate over all observations assigned to an EU trip in each year. These annual aggregate values are then matched to the MRV and WFR data by MMSI and year. We successfully match 1989 ships for 2019, 2084 for 2020, and 2349 for 2021.

Finally, given the potential for errors stemming from the raw AIS data and the trip detection

[6]We identify a port call when a ship does not exceed a speed of one knot for at least 12 hours and is within a nation's economic exclusion zone (200 nm from the coast).

[7]This strategy will tend to detect more than the correct number of port calls, as these are defined in the regulation as occurring only when cargo is loaded or unloaded, however this is not directly observable from the tracking data (EU, 2015).

[8]Ports in the United Kingdom ceased to be included as of the beginning of 2021.

Table 1: Reported fuel consumption summary statistics

Year	count	mean	sd
2019	585	1389	932
2020	671	1311	951
2021	692	1334	932

procedure, we develop our model using a subset of data for which the observed distance travelled on EU trips agrees well with the reported distance. Specifically, we select only those ship-year observations for which the calculated and reported distances agree within 500 nm.[9] Summary statistics of the resulting data set are provided in Table 1. We pool the observations from 2019 and 2020 to construct a training set of 1281 observations, while the observations from 2021 are set aside for out-of-sample testing.

3 Methodology

We train a model of fuel consumption in two steps: First, we calculate theoretical annual fuel consumption based on the bottom-up methodology used by the IMO. Second, we train machine learning models to predict fuel consumption, using as predictors both the calculated fuel consumption and additional ship characteristics. We evaluate the performance of these models using a cross-validation procedure with the training data as well as on a separate test set.

3.1 Engineering Calculation

We follow closely the IMO procedure described in Faber et al. (2020) to calculate theoretical fuel consumption, with only a few minor simplifications. The key steps are outlined below, and further details are provided in **??**.

A ship's total instantaneous fuel consumption, FC_i, is the sum of the fuel consumed by each of the main engine (ME), auxiliary engine (AE), and boiler (BO), each of which is the product of the demanded power W_i and the specific fuel consumption SFC_i for that component:

$$FC_i = W_{\mathrm{ME},i} \cdot SFC_{\mathrm{ME},i} + W_{\mathrm{AE},i} \cdot SFC_{\mathrm{AE},i} + W_{\mathrm{BO},i} \cdot SFC_{\mathrm{BO},i}.$$

Demanded power is the power required to move the ship through the water (and air). For the auxiliary engine and boiler, power values are assigned based on the ship's operating phase, as determined by its speed and distance from shore. For the main engine, demanded power is given by the Admiralty formula, which is nonlinear in speed v and draft t:

$$W_{\mathrm{ME},i} = C \cdot W_{\mathrm{ME},ref} \cdot \left(\frac{t_i}{t_{ref}}\right)^{0.66} \cdot \left(\frac{v_i}{v_{ref}}\right)^{3}, \tag{1}$$

where C is a constant applied to correct for factors such as weather and hull fouling, and the *ref*

[9] As a robustness check, we also apply an alternative criterion of ±10%.

subscript denotes reference values that are maximum ratings taken from ship specifications in the WFR.

Specific fuel consumption describes the efficiency of each engine. For the auxiliary engine and boiler this is taken to be constant and equal to SFC_{base}, while efficiency of the main engine is taken to be quadratic in engine load (the ratio of demanded power to reference power):

$$SFC_{\mathrm{ME},i} = SFC_{\mathrm{base}} \cdot \left(0.455 \cdot \left(\frac{W_{\mathrm{ME},i}}{W_{\mathrm{ME},ref}} \right)^2 - 0.710 \cdot \frac{W_{\mathrm{ME},i}}{W_{\mathrm{ME},ref}} + 1.280 \right).$$

SFC_{base} is assigned based on the ship's engine type, fuel type, and year built.

To aggregate, we assume operating conditions are constant over each hour (the frequency of the interpolated observational data) so that hourly fuel consumption is simply obtained by multiplying by the duration of each observation, t, equal to one hour. Annual fuel consumption FC is then the sum over all observations j in a year:

$$FC = \sum_j FC_{i,j} \cdot t. \tag{2}$$

3.2 Machine Learning

We evaluate the performance of several machine learning algorithms, both linear and tree-based. These include linear regression, lasso, ridge regression, gradient boosting regression, random forest regression, and CatBoost.[10]

We consider as predictor variables both ship characteristics from the WFR and activity-dependent variables derived from AIS tracking data. Given the hundreds of characteristics available and the potentially infinite variations on aggregating tracking data, we select a set of features based on a combination of domain knowledge and feature importance results from preliminary analyses. These are described in Table 2. The features we derive from tracking data can be split into three categories. The first includes straightforward aggregation of observational behaviour, such as total distance travelled, number of trips taken, and fraction of time spent at ports. The second category includes the fuel consumption calculated as per Section 3.1, as well as components of the Admiralty formula (1). Lastly, we include variables that represent aspects of the quality of the tracking data, including the number of interpolated observations when the ship is at sea (when most fuel is consumed) and the longest distance between observations.

We perform final data preprocessing before training the models. We drop observations for which the discrepancy between log calculated fuel consumption and log reported consumption is greater than three standard deviations from the mean. We use median imputation for missing values of ship characteristics *tonnage per centimeter (TPC)* (15% missing) and *net tonnage (NT)* (0.2% missing). All numeric variables are log-transformed, using $log(1 + x)$. The categorical variable *size category* is coded as an ordinal variable. Finally, for lasso and ridge regression only, we normalize the log-transformed variables to have zero mean and unit variance.

[10]We use Python's Scikit-learn library for all but CatBoost, for which we use the CatBoost library.

Table 2: Variable definitions

Variable	Description
Ship Characteristics	
Age	Age at observation (years)
Beam Moulded	Maximum width (meters)
Deadweight Tonnage	Capacity (tonnes)
Draft	Maximum draft (meters)
Length Between Perpendiculars	Length between the forward and aft perpendiculars (meters)
Length Overall	Total length (meters)
Main Engine Power	Power rating of main engine (kilowatts)
Net Tonnage	Dimensionless index based on the total volume of cargo spaces
Size Category	e.g., Handymax, Panamax, etc.
Speed	Representative speed from WFR (service or average observed) (knots)
Tonnage Per Centimeter	Load change required to change draft by one centimeter (tonnes/cm)
Ship Activity	
At-Port Fraction	Fraction of hourly observations with operational phase 'at-port'
Distance Travelled	Distance calculated from AIS data (nautical miles)
Number Of Trips	Number of trips detected from AIS data
Calculated	
Admiralty Draft Term	Aggregation of draft term of (1), calculated as $\sum(t/t_{ref})^{0.66}$
Admiralty Instantaneous Component	Aggregation of instantaneous component of (1), calculated as $\sum t^{0.66} \cdot v^3$
Admiralty Speed Term	Aggregation of speed term of (1), calculated as $\sum(v/v_{ref})^3$
Admiralty Static Component	Static component of (1), calculated as $C/(t_{ref}^{0.66} \cdot v_{ref}^3)$
Calculated Fuel Consumption	Theoretical fuel consumption, calculated as per (2)
Relative Draft	Aggregation of relative draft, calculated as $\sum(t/t_{ref})$
Relative Speed	Aggregation of relative speed, calculated as $\sum(v/v_{ref})$
Data Quality	
Interpolated Fraction At-Sea	Fraction of hourly observations with operational phase 'at-sea' that are interpolated
Longest Distance	Longest distance between consecutive locations after interpolation (nautical miles)

All derived values aggregated annually over EU trips.

Table 3: Calculated fuel consumption summary statistics

Year	count	mean	sd
2019	585	1440	1219
2020	671	1402	1292
2021	692	1383	1135

Hyperparameters are tuned using grid search with k-fold cross-validation, with k=5. The parameter grid values are provided in **??**. The best performing set of parameters for each algorithm is selected on the basis of the mean R^2 value across the splits. We assess the generalizability of the optimally tuned models in two ways. First, we perform 3x repeated 10-fold cross-validation and take the mean performance scores from the validation splits. Secondly, we fit each model on the entire training set (data from 2019 and 2020) and predict fuel consumption on the test set (2021 data).

4 Results

We first describe the calculated fuel consumption, before presenting the results of our hybrid machine learning approach, including a comparison of different machine learning algorithms using both the training set and the test set.

Summary statistics for the annual fuel consumption, calculated as per Section 3.1 are presented in Table 3. Over all three years of data, the mean calculated fuel consumption of 1407 tonnes is slightly higher than the reported average of 1343 tonnes. Figure 1 presents the calculated versus reported values for both the training and test sets. For the training set, the R^2 is only 0.275, which highlights the opportunity for improving the accuracy of fuel consumption and emissions estimates from the calculation-based approach.

The cross-validation training set performance of each optimally-tuned model is summarized in Table 4, and the optimal parameters are provided in Table 6 in **??**. The linear estimators perform slightly better than the tree-based models, both in terms of better average accuracy and less variation. Within the linear models, performance is almost identical, despite the regularization parameter being equal to one for ridge regression. The overall best performance is achieved with ridge regression, giving an R^2 of 0.956 and an MAPE of 9.8%.

The true out-of-sample validation on the test set of 2021 data is very similar, as shown in Table 5. The linear models again out-perform the tree-based models in general, with ridge regression again achieving the highest R^2, at 0.954, corresponding to a MAPE of 10.9%. Interestingly, CatBoost very narrowly achieves the best MAPE at 10.8%. For comparison, the predictions of these two models are shown in Figure 2.[11] CatBoost appears to perform slightly worse at high fuel consumption values. All models perform substantially better than the calculation-based approach, which achieves an R^2 of only 0.58. For robustness, we repeat our analyses with alternative criteria for inclusion in the data set, both doubling the distance discrepancy and applying a relative difference of 10%. The

[11] See Figure 3 in **??** for the plot in levels.

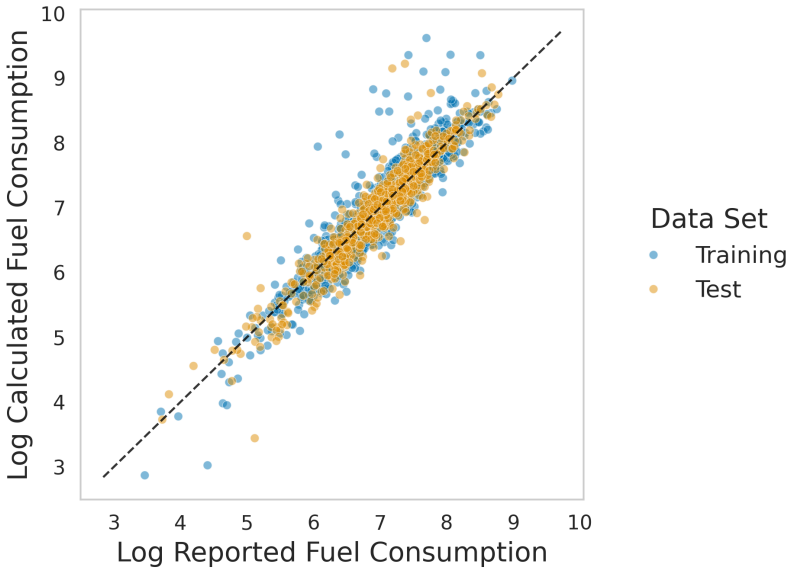

Figure 1: Engineering calculations vs. reported fuel consumption

Table 4: Training set cross-validation scores

	R^2		MAE (t)		MAPE (%)	
Model	mean	sd	mean	sd	mean	sd
Ridge Regression	0.956	0.019	122	13	9.8	0.7
Linear Regression	0.956	0.020	122	14	9.8	0.7
Lasso	0.954	0.020	125	14	10.0	0.7
Gradient Boosting Regressor	0.942	0.039	130	16	10.6	0.9
CatBoost Regressor	0.939	0.053	127	17	10.3	0.9
Random Forest Regressor	0.914	0.036	157	19	12.3	1.0

Note: Statistics computed from three repetitions of 10-fold validation.

Table 5: Test set scores

Model	R^2	MAE (t)	MAPE (%)
Ridge Regression	0.954	131	10.9
Linear Regression	0.953	132	10.9
Lasso	0.951	132	10.9
CatBoost Regressor	0.941	133	10.8
Gradient Boosting Regressor	0.939	143	11.6
Random Forest Regressor	0.931	151	12.3
Calculation	0.580	260	20.4

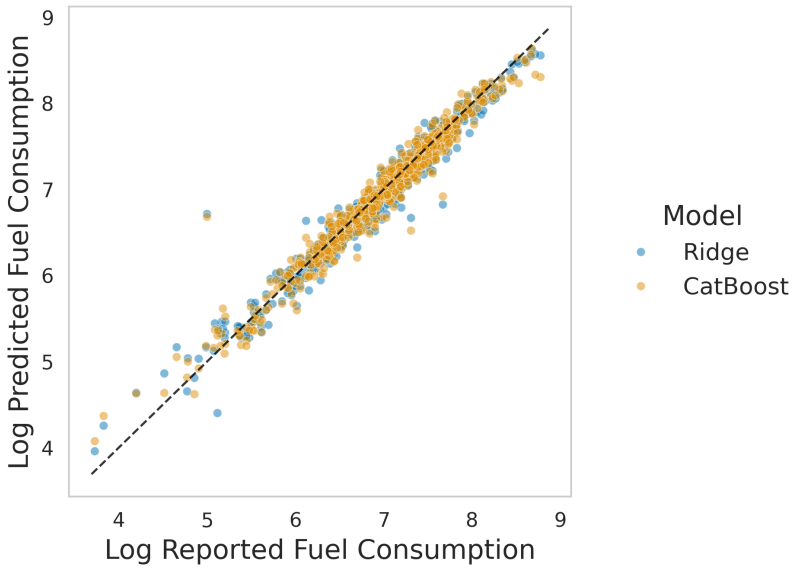

Figure 2: Test set prediction accuracy (log-log) for ridge regression and CatBoost

results are qualitatively similar, with slightly lower predictive accuracy with noisier data.[12]

5 Conclusion

Accurate and timely estimates of fuel consumption are key to understanding the drivers of GHG emissions from maritime shipping and to assess the effectiveness of emissions policies. In this paper we have developed a highly accurate methodology for predicting CO_2 emissions from readily-available tracking ship tracking data, which can be employed at an arbitrary frequency or with any subset of ships that may be of interest. To carefully capture the nonlinear effects of travel speed, we take a hybrid approach in which we calculate predictor variables from industry-standard fuel consumption formulas. We demonstrate a substantial increase in predictive accuracy over the pure calculation-based approach. Our results are not particularly sensitive to the specific machine learning model that is employed, and in fact computationally-efficient linear models perform best. It is our hope that this approach will be employed to provide more accurate and up-to-date estimates of shipping

[12]See Appendix E for detailed results.

emissions.

A Data Cleaning

B Summary Statistics

C Trip Detection

D Machine Learning Details

The optimal tuning parameters are provided in Table 6.

Table 6: Optimal tuning parameters

Model	Parameter	Value
CatBoost Regressor	depth	6
	l2 leaf reg	1
	learning rate	0.05
Gradient Boosting Regressor	learning rate	0.05
	max depth	6
Lasso	alpha	0.005
Linear Regression	n/a	n/a
Random Forest Regressor	max depth	30
	n estimators	200
Ridge Regression	alpha	1

The inferior performance of the CatBoost model versus ridge regression is shown in Figure 3.

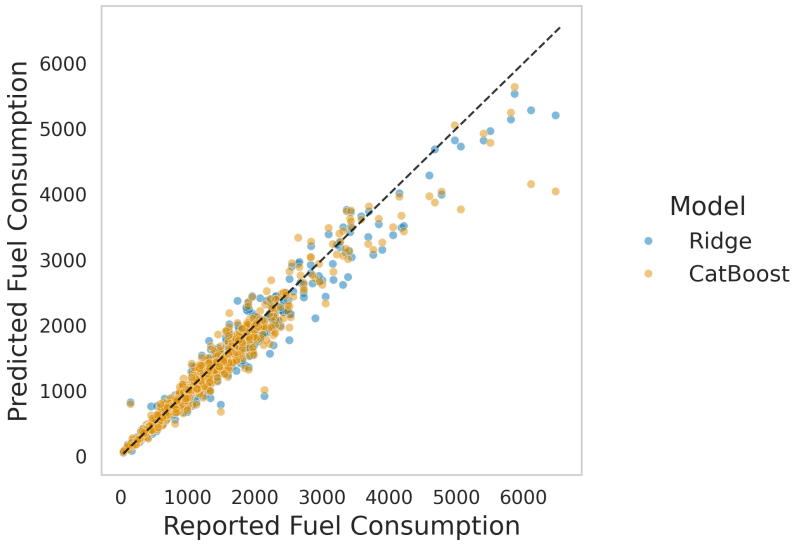

Figure 3: Test set prediction accuracy (levels) for ridge regression and CatBoost

E Alternative Data Inclusion Criteria