# Data Exploration

The dataset is comprised of a semi-contiguous observation period spanning from July 6, 2017 to August 6, 2018 (two days of data spanning July 31, 2018 to August 1, 2018 are missing). Null values seem to be represented as zero integers, will replace with NaN’s (excluding the fuel type Boolean columns).

The time series data can be broken up frames, contiguous periods of engine activity (as judged by Fuel Consumption) and other contagious periods of engine inactivity presumably while in port or otherwise not collecting data. I was initially concerned about the ‘spikes’ present in this data, but they generally contain multiple data points and are very well correlated with other readings like engine RPM and Speed through water (not shown) so I have elected not to drop or smooth them.



**Figure 1** Plot of Main Engine Fuel Consumption over the dataset. The red vertical lines denote the frame boundaries we will be using to separate valid from invalid data collection period’s.

The dataset has been hour level qualified and our target variable, Main Engine Fuel Consumption, represents the averaged burn rate over the hour. Because of the aggregate nature of the data, fuel consumption measurements at the beginning and end of voyages have been contaminated by having the engine off for an indeterminate period of time and will be excluded from analysis.

Stopping periods will similarly be excluded, as they will lead to overestimates in true model performance. These data filtering steps reduce the samples in our data set from the original 9.4K down to 6.3K points, a ~33% reduction.

# Train Test Split Design

I have elected to fit and evaluate models using separate data collection intervals to try and prevent the high degree of temporal correlation within adjacent data points from giving us over-confidence in the performance of our model.

Partitioning training and testing data by data collection makes it difficult however to ensure our model will experience the full range of possible input conditions. Will need to use cross validation to fully evaluate model performance across diverse input conditions.

# Feature Engineering

I have elected to build the following additional features:

1. Difference in compass heading between current and previous (1 hour pior) data point – presumably turning will impact vessel drag and fuel burn, and may indicate if in port or not. The heading deltas must be normalized to lie within the range +/- pi radians (turning the ship to the left or to right is ambiguous in this dataset without making assumptions about rudder position). Because left turns and right turns will be treated equally we will take the Sin transformation of the heading delta normalizing our feature between +/-1.
2. Trim as recommended by the appendix, thanks! Draft measurements for the first two trips are completely missing and the third are partially missing. We will drop data collection periods zero and one from the data and back fill the missing observations in data collection period 2 with the first valid measurement.
3. Sine transformation of the average rudder deflection angle over the hour. May also be indicative of turning or lateral environmental forces.
4. Continuous rudder deflections without change in heading may be indicative of lateral ocean currents or winds, will use the correlation value (rudder\_deflections(radians)-mean(rudder\_deflection)) /(compass\_heading\_deltas (radians) – mean(compass\_heading\_deltas)) as a feature.
5. Water depth readings of zero will be replaced by the max water depth value as indicated in the appendix.
6. A number of ​water surface temperature readings indicate absolute zero. Will replace those readings with linearly interpolated values between the nearest valid data points.
7. True wind direction relative to compass heading will be used to measure wind; the ships bow and vector of motion are not necessarily synonyms as currents or wind could be driving some lateral movement. To capture wind direction we will need to take both the sine and cosine components of this angle.
8. True wind speed is all that matters in regards to fuel burn so we will use this value as is. Again a few anomalous zero readings here, will fill them with a linear interpolation between points.
9. Some data points were missing Shaft Speed (RPM), and some were missing Shaft Power (kW). We will impute missing Shaft Power and RPM values from the known linear relationship between Shaft Speed, RPM and Power.

# Model Building and Selection

A hold out test of 10 measurement windows comprising 992 data points was used for calculating performance metrics and a GroupKFold generator object was used for cross validation.

Model building and selection was accomplished using a GridSearchCV generator to automate the testing of various combinations of preprocessing steps and learning hyper parameters.

# Findings

In my first Fuel Consumption modeling attempt I included Engine Shaft Power (a direct measure of engine rate of work) in the feature set and yielded a very well fitting model. Linear regression provides very robust fuel consumption estimation with regression coefficient on this data (R2 = 0.99) see below.



**Figure 2.** Correlation between model predicted and true fuel consumption on the holdout set. Coefficient of determination of the fit is 0.99 +/- 0.01 std as estimated from 5 fold cross validation. Redline indicating a perfect fit.

Looking at feature importance measures I noticed the engine power parameter was the essentially the only feature being used by the model. The environmental conditions like wind were under first inspection uncorrelated with fuel consumption. I have plotted Fuel consumption as a function of speed through water below (both in reality and in the modeled data).



Figure 3 Fuel Consumption as a function of speed through water on the holdout set, in truth (left), a linear model using engine performance data (right) (blue 0.3 knots --> yellow 35 knots ).

I was at first troubled by this finding, but I (eventually) thought about the problem from the perspective of a physicist, and realized my initial analysis was somewhat silly. Environmental conditions, like wind, do not materially impact the thermodynamic efficiency with which the engine transforms fuel energy into (propeller) rotational energy (assuming water maintains a constant load on the propeller). In short the rate of fuel consumption of a vessel should only depend on how hard the crew runs the engine (engine power/rpm/torque)![[1]](#footnote-1)

Environmental conditions will however impact the efficiency with which fuel energy is converted into forward motion; so I can instead treat Speed Over Ground as the dependent variable of our model.



Figure 4 Speed over Ground on the holdout set as a function of engine speed. Points are again colored by prevalent wind speed (blue 0.3 knots --> yellow 35 knots). Note that higher winds tend to correlate with lower ground speeds.

This data was best fit by a Random Forest regressor built from 10 trees and max tree depth of 4 with a R2=0.88 on the holdout set. Using this model we could then investigate the effects various environmental parameters would have on the predicted ground speed at a given engine shaft speed.

Plotted below are the effects, as predicted by our model, of setting wind speed to zero (top) and setting trim to the mean trim value over all voyages (bottom).





1. Variations in temperature would have a small impact on parameters like water density, and engine/environment temperature deferential that would in turn impact the thermodynamic

   efficiency of the propeller but I would expect these effects to be minute under realistic environmental variations. [↑](#footnote-ref-1)