A data mining approach for predicting main-engine rotational speed from vessel-data measurements

Dimitrios Kaklis[1,2,3]

joint work with:
G. Giannakopoulos[2], I. Varlamis[1], C. Spyropoulos[2], T. Varelas[3]

(1): Harokopio University of Athens - HUA
(2): NCSR Demokritos - NCSRD
(3): DANAOS Shipping Co.

dkaklis@iit.demokritos.gr

June 10, 2019

The general problem

Optimal vessel routing

- Given a vessel, its starting and destination ports and an initial route,
- find alternative routes that minimize the vessel's fuel-oil consumption (*foc*)

The solutions

optimization alternatives

- **environmental optimization**: optimize the route by taking into account environmental data, e.g., wind/currents (speed, direction), wave (height, frequency, direction).
- **vessel-based optimization**: optimize the route *wrt* vessel characteristics, e.g., vessel speed, engine-rpm, trim, roll, heave and pitch motions.

Multi-objective optimization problem

Combine both approaches in a multi-objective optimization setting and compute the *foc*-best route for a given vessel

Motivation I

The facts

- The *foc* prediction problem employs a rich and composite feature set
- Before training any multi-parameter prediction model it is important to study the effect of each parameter separately
- the main-engine revolution speed influences fuel consumption

Exploratory analysis on a real dataset

feature importance on estimating foc		
feature	**importance**	**description**
RPM	0.98353	main-engine revolutions per minute
STW	0.00365	vessel speed through water
overground speed	0.00266	vessel speed with respect to the ground
apparent wind speed	0.00133	the relative speed, i.e., the speed experienced by an observer or a measuring instrument on the ship

Table: **the top-4 ranked features by importance, using Random Forest regression** (*contribution of each feature in decreasing the mean variance from the actual mean value of the population*)

Motivation II

The pertinent literature shows that:

- no robust, low complexity, analytical relation between RPM and V exists
- power absorbed by the ship's propulsion system is analogous to the product $RPM \cdot Q$, where Q (torque) depends exclusively from the ratio V/RPM,
- velocity V is a feature that can be easily measured and does not require further installments (e.g. sensors) on board

Relation between RPM and V

very similar rate of change between the two variables

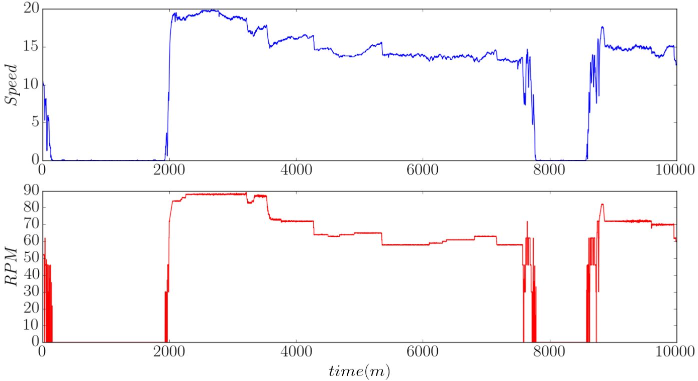


Figure: plot of main-engine's rotational speed (*RPM*) and observed speed *V* during vessel's route

Relation between RPM and V

changes in velocity (over small time windows) are highly correlated with RPM

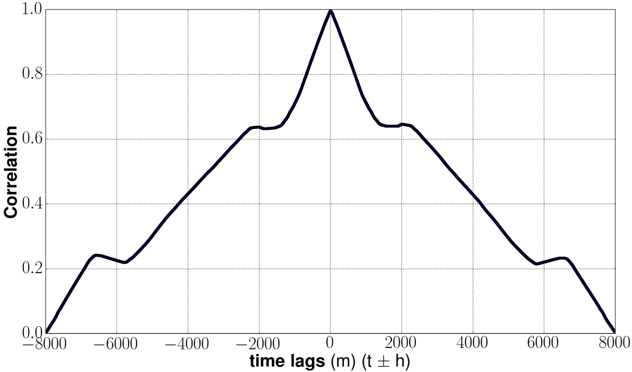


Figure: The correlogram of *RPM* and *V* during a vessel's route

Conclusions from feature study

The exploratory analysis shows that:

- *RPM* plays a pivotal role in the prediction of *foc*
- *RPM* and *V* have a high linear relation (**PPMCC** ≈ 0.95)
- Since there is a strong linear relationship on certain time - windows we can use lag variables as extra features on our estimators

$*$*PPMCC : measure of the linear correlation between two variables*

Problem formulation

Given a vessel's speed for n consecutive moments, find a function

$$f(V_1, ..., V_n) : R^n \rightarrow R^1, \quad V_i = V(t_i), \; i = 1, \ldots n,$$

which estimates the engine's RPM at moment t_{n+1}

Our approach

- RPM - V $\rightarrow$ a partially linear function with non linear segments (very similar V values correspond to different values of RPM)
- we use the **Linear Mixed Model (LMM)** to model both the fixed and random behavior of this relationship:

$$rpm = D \cdot V + \Gamma \cdot \bar{V}_N + \epsilon,$$

  - ✓ V : fixed effect vector,
  - ✓ $\bar{V}_N$: random effects vector for clustering data in areas of similar velocity variations
  - ✓ D, Γ: vectors/matrices of the fixed-/random-effects regression coefficients,
  - ✓ ϵ : error term (the part of the response variable not explained by the model)
- we use **Smoothing Splines** to evaluate the regression coefficients D and Γ which have been shown to perform well on MME-formulated problems

Algorithm overview

The algorithm

- Cluster the input space D of $(V, \bar{V}_N) \rightarrow RPM$ values to form $D_1 \ldots D_k$ different clusters
- Train k different models M_i, one for each cluster D_i , $i \in [1, k]$
- At the evaluation stage classify each instance $(V, \bar{V}_N)$ to the most similar cluster and predict the corresponding RPM with the specified model trained on this particular group of data.

Partitioning the input space

clustering of the velocity space

- Clustering is performed on a vector of the form: $(V(t_i), \bar{V}_N(t_i))^T$ with $\bar{V}_N(t_i) = mean[V(t_{i-N}, ..., V(t_i)]$, and the corresponding $RPM(t_i)$ value.
- Clusters correspond to vessel sub-trajectories in which the vessel has the same velocity change pattern (e.g. accelerates, decelerates or keeps stable velocity).

Evaluating various Regression & Clustering method alternatives

Regression methods

- **Spline regression (SR)**
- Linear regression (LR)
- Random Forest regression (RF)
- A baseline Neural Network (NN)

Clustering methods

- K-means clustering
- **Delaunauy triangulation clustering (DTC)**

Datasets

description

- real time-series dataset consisting of $3 \cdot 10^6$ obs. (3 month trip of a vessel) of 1 min. granularity, provided by Danaos shipping co.
- feature vector consist of : **power , speed , trim , draft and weather based features (wind speed/angle ,wave height)**
- The techniques presented in this work have been tested in datasets of size $\approx 10^q$ $(q = 3, 4, 5)$.
- In all the experiments that follow we apply the two-sample **Kolmogorov-Smirnov (K-S) test.**

The effect of clustering

In the next 3 slides initial results are visualized from testing our method in ≈10^3 obs. from one vessel trip. Smaller errors (MAE) are better!

With K-Means

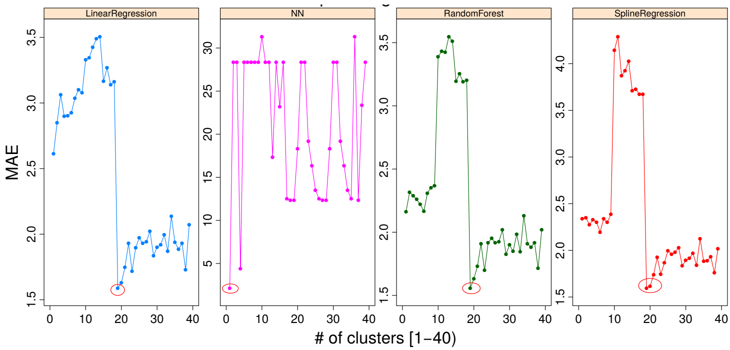
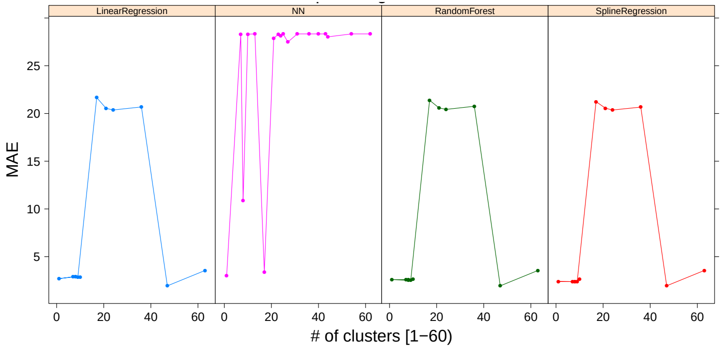


Figure: MAE minimized for apprx. 18 clusters with K-Means (circled) except for NN

The effect of clustering II

With Delaunay Triangulation clustering



The effect of clustering III

With Delaunay Triangulation clustering (scaled error axis)

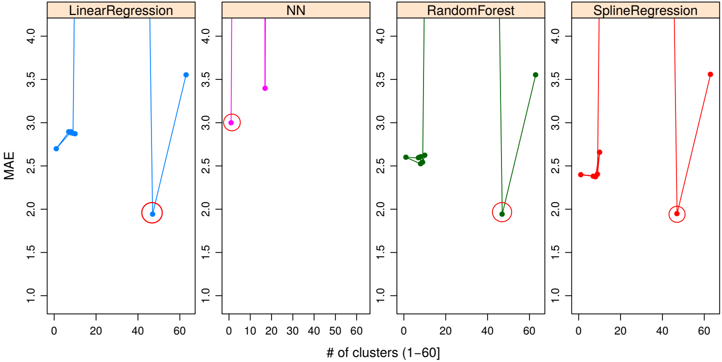


Figure: MAE minimized for apprx. 45 clusters with DTC (circled) except for NN

Findings

From the results above we conclude that:

- clustering improves the regression algorithms performance especially for LR, RF, SR
- NN performs better on single cluster than on many clusters
- There appears to be an optimal number of clusters for which they achieve the highest accuracy

Splines vs others

Splines perform substantially better

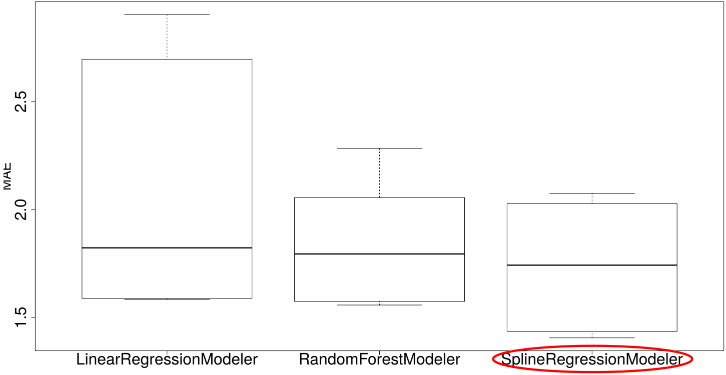


Figure: The distribution of MAE values for 10 different trips of the same vessel

Combining splines with clustering

Splines perform substantially better when combined with DTC

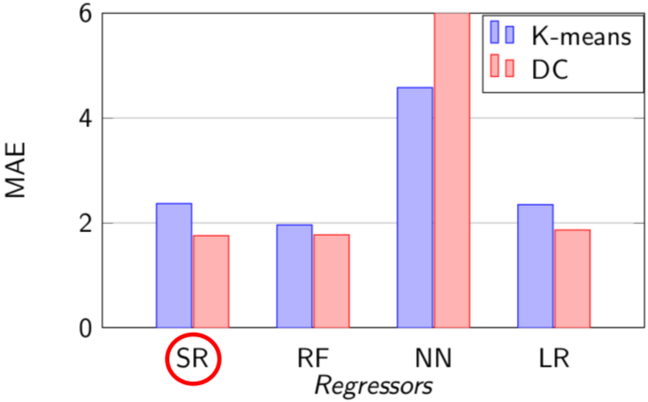


Figure: MAE values for 10 different trips of the same vessel for K-means and DTC separately

Conclusions

- **Spline (SR)** and **RF (Random-Forrest)** regression alongside with clustering perform better than LR and NN.
- Enhancing our feature vector with **the mean of velocity at** N **previous time - steps** we managed to improve the accuracy of our predictive scheme.
- **Splines** combined with **DTC** perform slightly better than the other three regression methods tested
- NN performs better when trained to a single cluster than to many clusters

Future work / next steps

Further research needs to be conducted as far as :

- finding the **optimal number of clusters** for both clustering techniques
- the number and placement of **knots in SR** used to approximate the underlying function on each partition
- the hyper-parameter N that controls the **previous time steps**
- finding a way to **incorporate NN** to the setting of our proposed method

Thank you