How can machine learning help to predict changes in size of Atlantic herring?

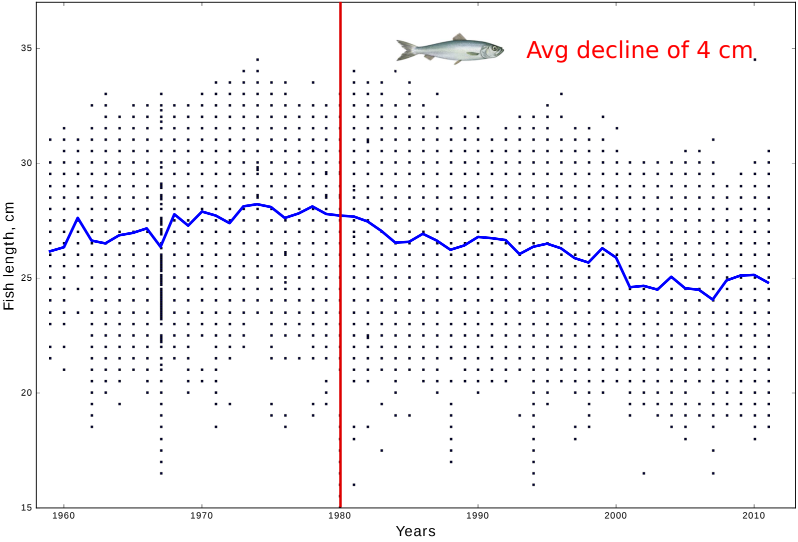
Olga Lyashevska* Clementine Harma Deirdre Brophy
Coilin Minto Maurice Clarke

* Marine and Freshwater Research Centre
Galway-Mayo Institute of Technology (GMIT)
Galway, Ireland

olga.lyashevska@gmit.ie

July, 22 2016

Background


Avg decline of 4 cm
Fish length, cm
35
30
25
20
15
1960
1970
1980
1990
2000
2010
Years

Problem

- Herring are one of the most important pelagic species exploited by fisheries;

Problem

- Herring are one of the most important pelagic species exploited by fisheries;
- Reductions in growth have consequences for stock productivity;

Problem

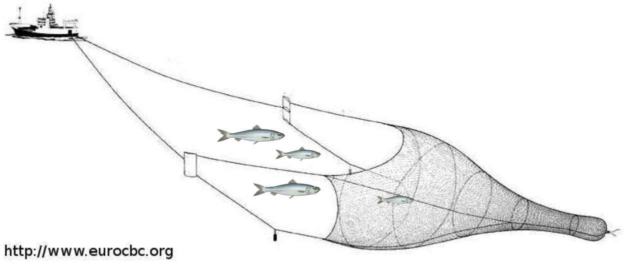
- Herring are one of the most important pelagic species exploited by fisheries;
- Reductions in growth have consequences for stock productivity;
- The cause of the decline remains largely unexplained;

Problem

- Herring are one of the most important pelagic species exploited by fisheries;
- Reductions in growth have consequences for stock productivity;
- The cause of the decline remains largely unexplained;
- Likely to be driven by the **interactive effect** of various factors:
  - sea surface temperature;
  - zooplankton abundance;
  - fish abundance;
  - fishing pressure;

Data

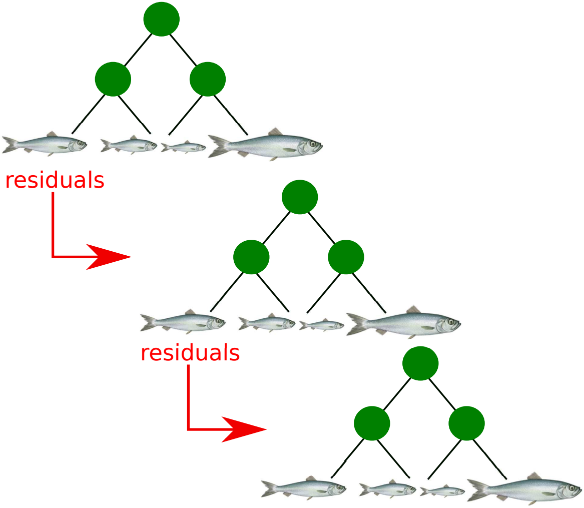
- 1959 – 2012;
- throughout the year;
- random sampling ($n = 50$ to 100) from commercial vessels;
- pelagic trawling;
- age and weight-at-length;
- total sample size 50,000;



Study Area

Objective

To identify important variables underlying changes in growth using Gradient Boosting Regression Trees (GBRT)



GBRT

- Advantages:
  - Detection of (non-linear) feature interactions;
  - Resistance to inclusion of irrelevant features;
  - Heterogeneous data (features measured on different scale);
  - Robustness to outliers;
  - Accuracy;
  - Different loss functions

GBRT

- Advantages:
  - Detection of (non-linear) feature interactions;
  - Resistance to inclusion of irrelevant features;
  - Heterogeneous data (features measured on different scale);
  - Robustness to outliers;
  - Accuracy;
  - Different loss functions
- Disdvantages:
  - Requires careful tuning;
  - Slow to train (but fast to predict);

Formal Specification

$$F_m(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{1}$$

where γ_m is a weight and $h_m(x)$ are weak learners.

Formal Specification

$$F_m(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{1}$$

where γ_m is a weight and $h_m(x)$ are weak learners.
GBRT builds the additive model in a forward stagewise fashion:

$$F_m(x) = F_{m-1}(x) + \epsilon \gamma_m h_m(x) \tag{2}$$

where ϵ is a shrinkage.

MARINE AND FRESHWATER RESEARCH CENTRE
LÁRIONAD TAIGHDE MARA AGUS FIORUISCE

Formal Specification

$$F_m(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{1}$$

where γ_m is a weight and $h_m(x)$ are weak learners.
GBRT builds the additive model in a forward stagewise fashion:

$$F_m(x) = F_{m-1}(x) + \epsilon \gamma_m h_m(x) \tag{2}$$

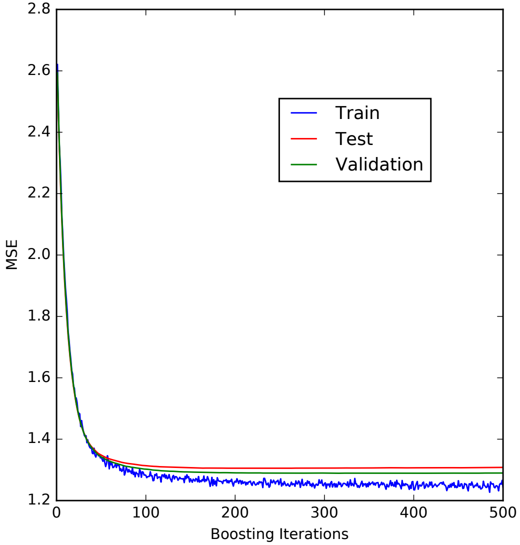
where ϵ is a shrinkage.
At each stage the weak learner $h_m(x)$ is chosen to minimize the loss function L given the current model F_{m-1} and its fit $F_{m-1}(x_i)$

$$F_m(x) = F_{m-1}(x) + \arg\min_h \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - h(x)) \tag{3}$$

GBRT hyperparameters

- number of iterations = 500;
- shrinkage (learning rate) = 0.05;
- max tree depth = 6;
- subsample = 0.75;
- loss function = Least Squares;

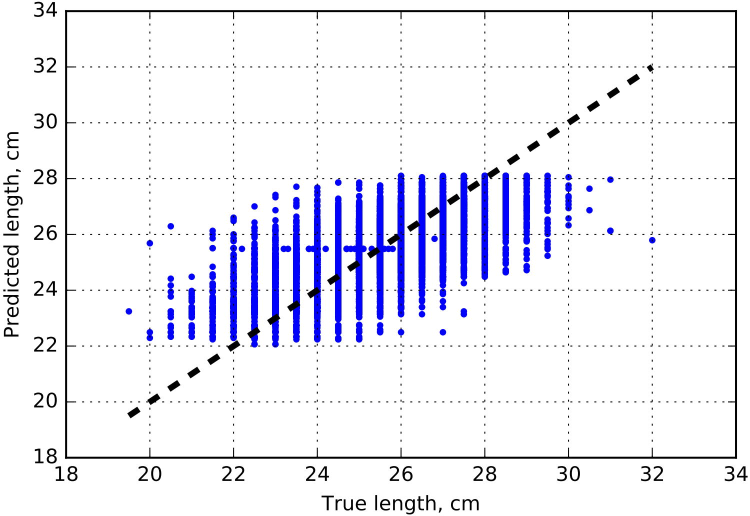
Model estimation



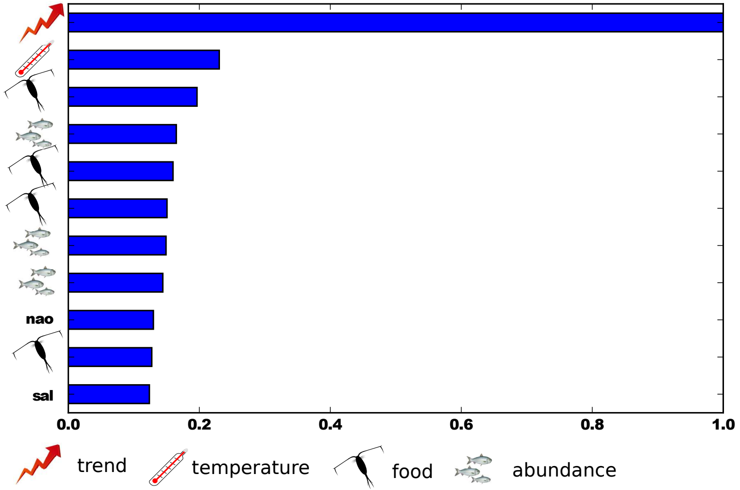
Low R^2 due to individual variability

- MSE: 1.31
- R^2 train: 54.5%
- R^2 test: 51.7%
- R^2 val: 52.6%

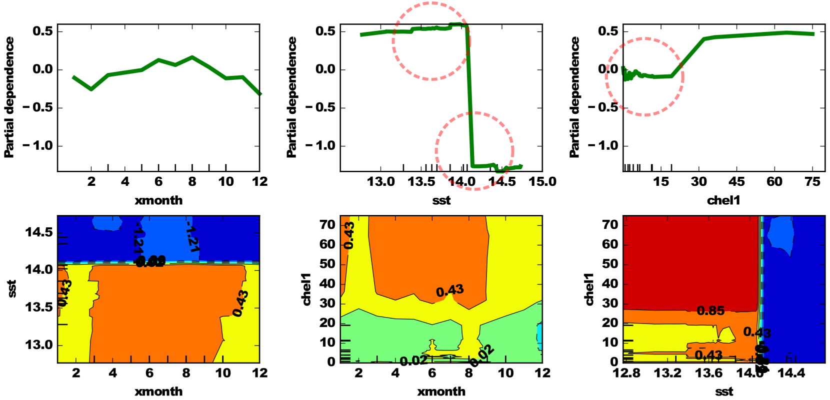
True vs Predicted



Variable Importance Plot



Partial Dependence Plots



Conclusions

- trend, sea surface temperature and food availability are three most importans features;

Conclusions

- trend, sea surface temperature and food availability are three most importans features;
- sea surface temperature above 14 degrees negatively relates to fish length, whereas food availability is invariant;

Conclusions

- trend, sea surface temperature and food availability are three most importans features;
- sea surface temperature above 14 degrees negatively relates to fish length, whereas food availability is invariant;
- there is a high degree of interaction between all features;

Conclusions

- trend, sea surface temperature and food availability are three most importans features;
- sea surface temperature above 14 degrees negatively relates to fish length, whereas food availability is invariant;
- there is a high degree of interaction between all features;
- not a cause-effect relationship, but a relative importance of the variables;

lyashevska
linked.in/lyashevska

Acknowledgements:
This research was carried out with the support of the Irish Environmental Protection Agency grant (Ecosystem tipping points: learning from the past to manage for the future, project code 2015-NC-MS-3) and the support of the Marine Institute under the Marine Research Sub-programme funded by the Irish Government.