Applications of **machine learning** to the **forecasting** of short-term **sea lice abundances** in British Columbia

Tristan Garry

Supervisors: Marie-Josée Fortin, Martin Krkošek
Department of Ecology and Evolutionary Biology
University of Toronto, Toronto, Ontario, Canada

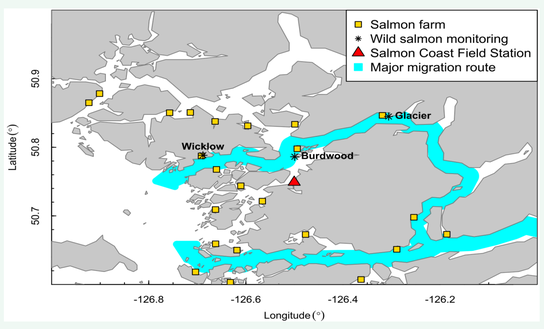
Correspondence: tristan.garry@mail.utoronto.ca

UNIVERSITY OF TORONTO

Introduction

- Two species of sea lice are common on Pacific salmon: ***Lepiophteirus salmonis*** & ***Caligus clemensi*** and cause high mortality in juvenile salmon and populations[1]
- Sea lice outbreaks in salmon farms have been found to increase parasitism in wild salmon

L. Salmonis in 3 life stages: mature female with eggstrings, mature female, immature louse.[2]



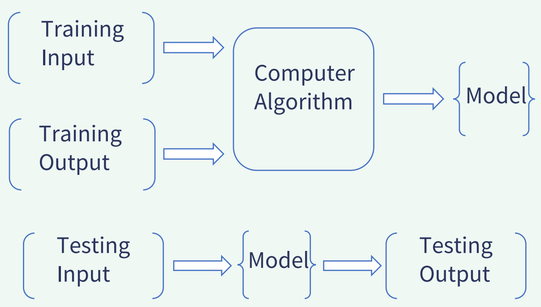
Research area in the Broughton Archipelago with local salmon farms and wild salmon monitoring efforts indicated.[2]

Objective: Using monitoring data from the Broughton Archipelago (BC, Canada), determine if machine learning can be applied to accurately **forecast sea lice abundances** during previously unobserved seasons?

- If successful, can we use these models to help in the management of these parasites and other types of outbreaks?

What is Machine Learning (ML) and why use it?

- Training inputs and output are processed by a computer algorithm to create a model
- The model consists of weights that best transform the inputs to outputs
- The model can then take previously unseen inputs to produce new output



- The models chosen in this study were selected due to their successful use in forecasting in the past[3], [4]

Applying the Models

- Models were split into training and testing data
- All the models were then trained on the training set
- These models were then tasked with forecasting the testing outputs
- The accuracy of these outputs was evaluated against the real values of that year using Mean Absolute Error (MAE)
- This process was repeated for all years in order to evaluate the models' ability to forecast previously unobserved years

2003, 2004, 2005, 2006, … , 2017

2004-2017 Inputs & Outputs → Algorithm → Models

2003 (Inputs) → Models → 2003 (Outputs)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |predicted_i - actual_i|^2$$

2003, 2004, 2005, 2006, … , 2017
⋮
2003, 2004, 2005, 2006, … , 2017

Model Inputs:

Based on wild salmon inputs and local environmental variables since 2003 and farmed salmon inputs since 2011

- Date
- Sea surface temperature
- Salinity
- Local farm sea lice abundance
- Lagged wild juvenile sea lice abundance
- Air temperature

Model Output

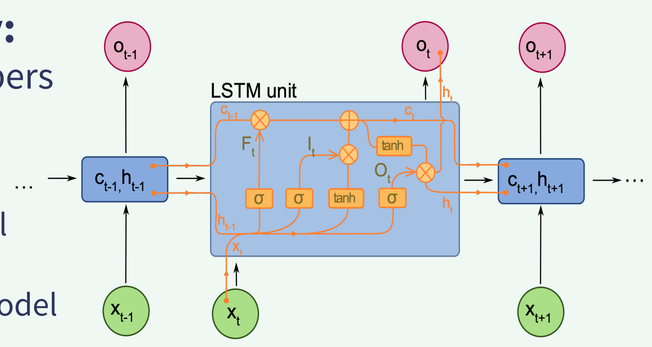
- Wild mature sea lice abundance

Models Used [3], [4]

Naïve Bayes: A simple machine learning model that given a date of the year, outputs the average of all previous years' outputs on that date. Equivalent to a null model.

Long Short-Term Memory:
A neural network that remembers recent trends

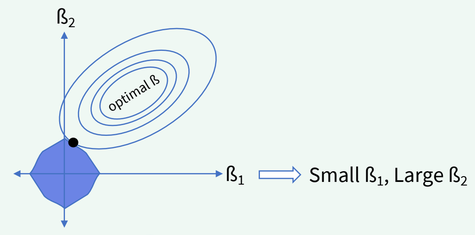
- Inputs are $\mathbf{x_t}$, outputs are $\mathbf{o_t}$
- Short-term trends $\mathbf{c_t}$ and model weights $\mathbf{h_t}$ are updated at each timestep, t and influence the model output



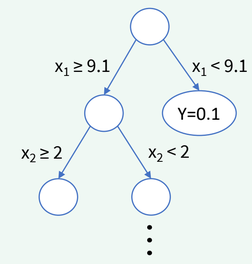
Models Used: Cont. [3], [4]

AutoML: A framework used to **sequentially test multiple machine learning models**. The best fitting model is then chosen to make predictions. The model classes assessed were:

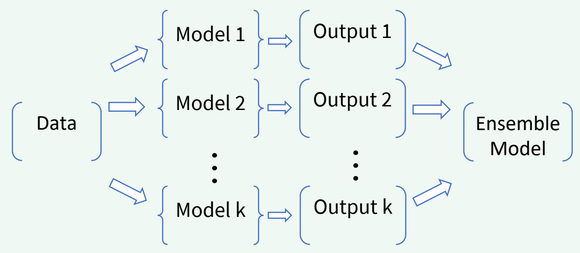
- **Elastic Net**: A regression method where the parameters, $ß_i$, are constrained to only be large if they contribute sufficiently to the fit of the model.



- **Tree-based algorithms**: Split the data on predictor threshold values in order to best explain it. Algorithms used were decision trees, random forest, and gradient boosting.



- **Ensemble models**: Use ensembling methods on the previously fit models' predictions to create a meta-model that is better than the individual models. Algorithms used were voting and stacking.

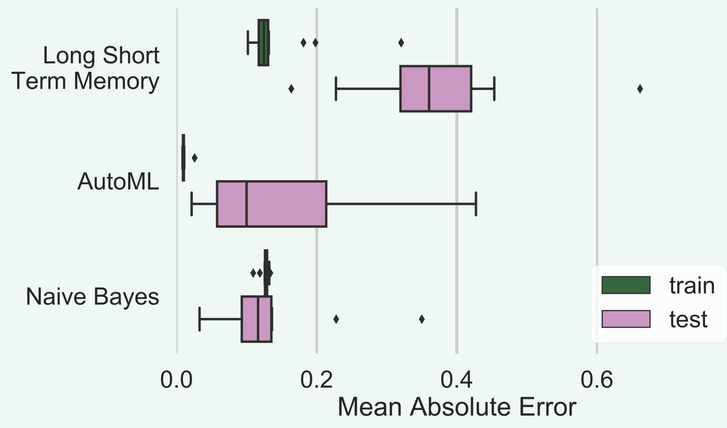


The final AutoML model used was a **Stack Ensemble** model of all other models' outputs.

Results

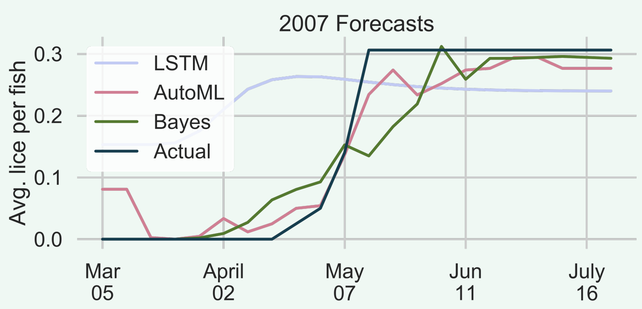
Model Performance

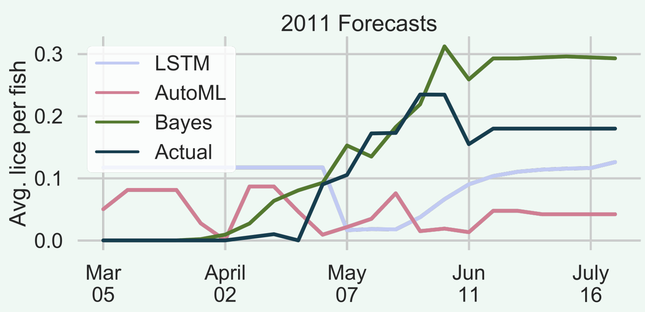
- Models described observed (training) years very well
- Models could not adequately predict unobserved (testing) years



Forecasts

- High variance in unobserved years' forecasting accuracy
- Unable to determine what makes a year forecastable or not
- E.g. Forecasts of 2007 had high accuracy in all models but 2011 had high variance in model forecasts





Key Takeaways

- Sea lice parasitism of salmon in BC is a problem
- The treatment of these sea lice would benefit from accurate forecasts of sea lice abundance
- Several machine learning models were investigated for their suitability in producing forecasts
- The fit models **described observed years well** but **could not forecast unobserved years**

Future Directions

- Topics to further investigate in forecasting this system:
  - Scale: Are we forecasting at the right scale?
  - Variables: Are we collecting the right variables?
  - Forecastability: Are these within-season forecasts inherently forecastable?
  - Model choice: Is a mathematical process model a better choice?

Acknowledgements

I would like to thank my supervisors M.J. Fortin and M. Krkošek for assistance with this project, as well as all of those involved in the collection of the large amounts of data used in these analyses

References

[1] Peacock, S. J., et al. 2013. Cessation of a salmon decline with control of parasites. Ecological Applications 23:606–620. ; [2] Thomas Bjørkan / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0); [3] James, Gareth, et al. An introduction to statistical learning. 2013. Vol. 112. Springer, New York, USA.; [4] Chollet, F. 2017 Deep Learning with Python. Manning, New York, USA

The source code of this project, more detailed references on background work, along with any accompanying writeups can be found by following the QR code to the right.