Forecasting Phytoplankton Abundance Using Remotely Sensed Data and Machine Learning

Cheenar Banerjee (cheenarb@), Sierra Kaplan-Nelson (sierrakn@)

Overview

- Marine satellite and in-situ monitoring can help managers understand where key productivity hot spots are in the Arctic that form the basis for areas rich in fisheries and marine resources
- Understanding where key ecosystem components are in the Arctic will help inform the regulation of expanding human activities, such as oil and gas development and shipping

Models

- **Baseline Linear Regression**
- **Locally Weighted Linear Regression**
  - Weight by location (inverse Euclidean distance from query point)
  - Weight by time (inverse number of days from query point)
  - Weight by location and time (sum of above)

- **Linear Regression** $J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})^2$

- **Lasso Regression** $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})^2 + \alpha||\theta||_1$

- **Ridge Regression** $J(\theta) = \sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})^2 + \alpha||\theta||_2^2$

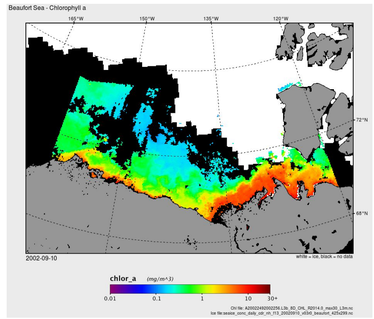
- **Decision Tree**
- **Random Forest**

$$Q_{left}(\theta) = (x, y)|x_j <= t_m$$
$$Q_{right}(\theta) = Q \setminus Q_{left}(\theta)$$

$$G(Q, \theta) = \frac{n_{left}}{N_m}H(Q_{left}(\theta)) + \frac{n_{right}}{N_m}H(Q_{right}(\theta))$$

Data

- 2002-2014: Train (1080268 samples)
- 2015: Test (46298 samples)
- Remotely sensed
- Early July to late September
- Cholorophyll a levels as proxy for biological productivity measure

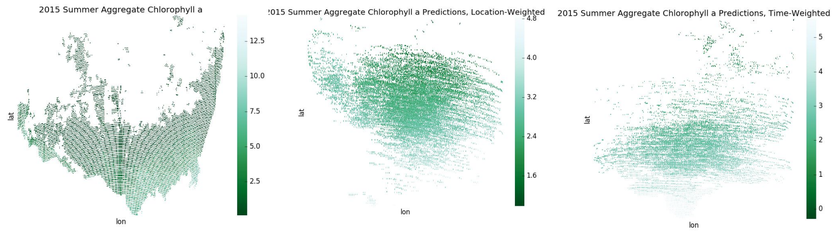


Features

- Latitude (0.421)
- Longitude (0.115)
- Depth (0.102)
- Mean photosynthetically active radiation (0.068)
- Sea surface temperature (0.062)
- Cloud cover (0.058)
- Max photosynthetically active radiation (0.046)
- Distance to land (0.033)
- Hours of daylight (0.013)
- Sea ice (0.012)
- doy.start (0.005)
- doy.end (0.004)

Feature importance coefficients from random forest baseline model

Discussion



- Best model: random forests, no weighting
  - Best model only accounted for ~48% of variance
- Locally weighted regression by location and day of year improves basic regression but hurts random forest, which is better at not overfitting.
- Day of year is a noisy measure of temporal aspect
- Predictions for concentrations in future years difficult due to weather variance year to year.

Results

Unweighted	Train MSE	Train R2	Test MSE	Test R2
Linear Regression	3.056	0.339	2.862	0.412
Lasso	3.815	0.175	3.958	0.187
Ridge	3.055	0.339	2.862	0.412
Decision Tree	1.757e-09	0.999	5.474	-0.124
*Random Forest**	*0.047*	*0.990*	*2.548*	*0.477*
Location-Weighted	**Train MSE**	**Train R2**	**Test MSE**	**Test R2**
Linear Regression	2.722	0.323	2.612	0.403
Ridge	2.722	0.323	2.612	0.403
Decision Tree	1.6933-09	0.999	6.323	-0.691
Random Forest	0.047	0.988	2.974	0.213
Time-Weighted	**Train MSE**	**Train R2**	**Test MSE**	**Test R2**
Linear Regression	2.891	0.357	2.612	0.403
Ridge	2.891	0.357	2.613	0.403
Decision Tree	1.776e-09	0.999	5.978	-0.389
Random Forest	0.047	0.989	2.995	0.319
Location and Time-Weighted	**Train MSE**	**Train R2**	**Test MSE**	**Test R2**
Linear Regression	2.852	0.278	2.726	0.162
Ridge	2.852	0.278	2.727	0.163
Decision Tree	1.716	0.999	5.842	-0.676
Random Forest	0.048	0.987	2.935	0.218

Future Steps

- Forecasting within shorter timeframe (predict late season from early season data)
- Test model on data from different location and compare feature weights
- PCA/ICA to reduce dimensionality and learn more about feature importance

References

- Kotta, J., Kutser, T., Teeveer, K., Vahtmäe, E. and Pärnoja, M. (2013). Predicting Species Cover of Marine Macrophyte and Invertebrate Species Combining Hyperspectral Remote Sensing, Machine Learning and Regression Techniques. PLOS.
- Hall, M.A. (2000). Correlation-based feature selection of discrete and numeric class machine learning. (Working paper 00/08). Hamilton, New Zealand: University of Waikato, Department of Computer Science.
- Cleveland, W. and Devlin, S. (2012). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. [online] Taylor & Francis. Available at: http://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478639