Learning Seasonal Phytoplankton Communities with Topic Models

Arnold Kalmbach[1], Heidi M. Sosik[2], Gregory Dudek[1], and Yogesh Girdhar[3]
[1]School of Computer Science, McGill University. ([2]Biology Department, [3]Applied Ocean Physics and Eng. Department) Woods Hole Oceanographic Institution.

Abstract

We develop a probabilistic generative model for phytoplankton communities, learning the associations between taxa by co-occurrence patterns in an extensive dataset.

The community model is trained using a method designed to ensure that it is interpretable in terms of simple environmental factors.

We demonstrate that our model affords a more accurate, simple interpretation of the distribution of taxa than approaches which do not consider community structure.

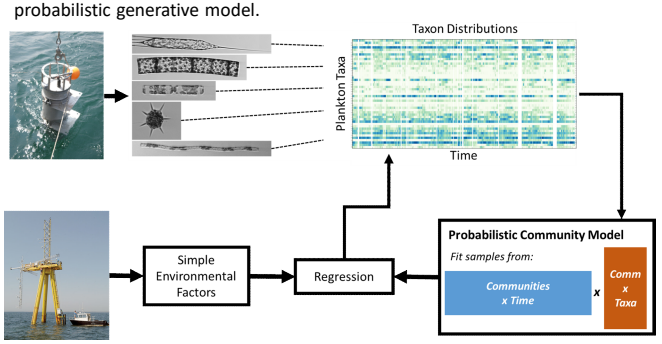
Our model indicates a remarkably strong seasonal structure in the distribution of taxa found near Martha's Vineyard, MA.

Introduction

IFCB autonomously detects and classifies phytoplankton in water samples into 47 taxa [1]. It has been deployed near Martha's Vineyard, MA, sampling continuously since Jan. 2009.

The interactions of each individual taxon with the environment require complex models to understand. Individually modelling each requires a prohibitive amount of data.

Taxa mainly co-occur with a small number of other taxa, *i.e.* they form **communities**. We can learn these communities from the IFCB dataset with a probabilistic generative model.



We select amongst the possible community models by their **interpretability**. The selected model's communities can be accurately predicted by a simplistic regression model driven by basic oceanographic data.

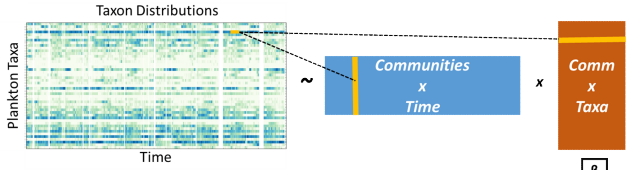
Regression Inputs	
Ocean Summary	**Meteorological Summary & Other**
Wave Height	Mean Wind Speed
Wave Period	Mean Wind Dir.
Wave Dir.	Relative Humidity
ADCP Vel. (2 depths)	Barometric Pressure
ADCP Dir. (2 depths)	IR Rad. Capt.
Tide	Solar Rad. Capt.
Water Temp	Num. Plankton Observed
Salinity	Day of Year

Models Compared	
Regression Method	**Regression Target**
Our Community Model	Distribution of Communities → Distribution of Taxa
Direct Regression	Distribution of Taxa
PCA Regression	Truncated SVD Weights → Distribution of Taxa

Interpretable Probabilistic Community Model

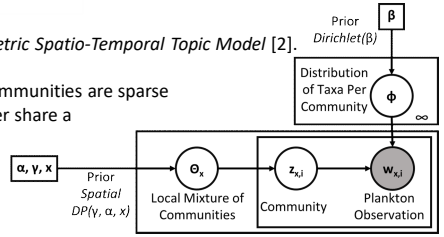
Our model represents a day's measurements as a probability distribution over taxa.

These are factored into a distribution over communities for each day and a distribution over taxa for each community.



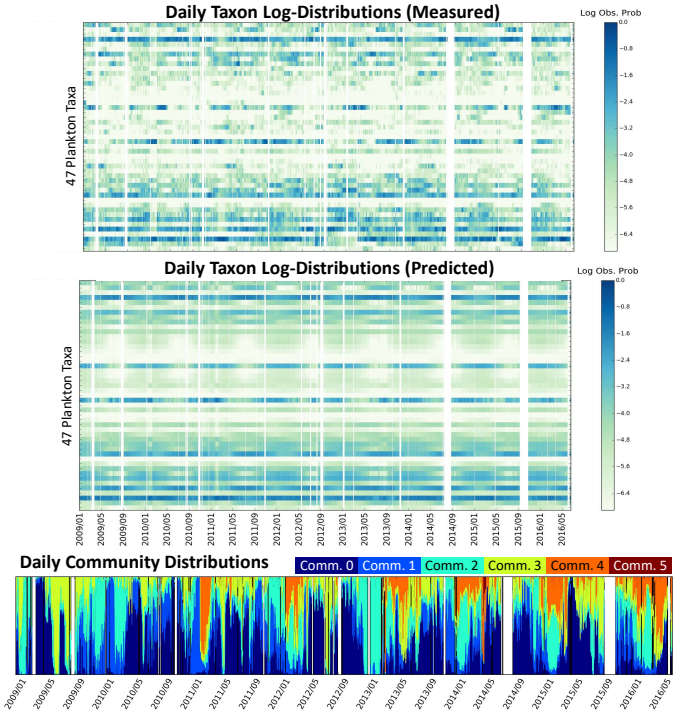
We use a *Bayesian Non-Parametric Spatio-Temporal Topic Model* [2]. With this model:

- Community mixtures and communities are sparse
- Observations near each other share a community prior
- The number of communities is set automatically.

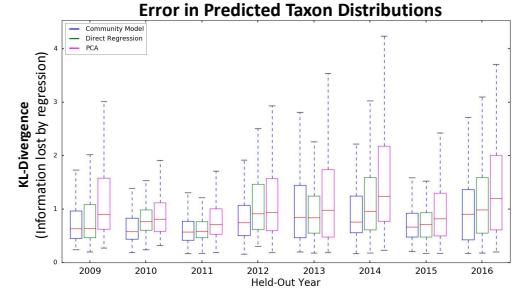


Results

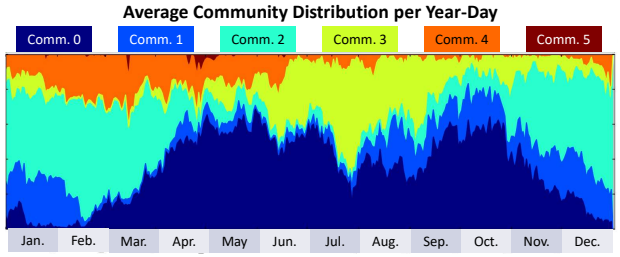
Despite the simplicity of the regression model, prediction via communities captures most of the low frequency variation in the taxon distributions



Taxon distribution regression via our community model shows less information lost than regression via PCA or direct regression.

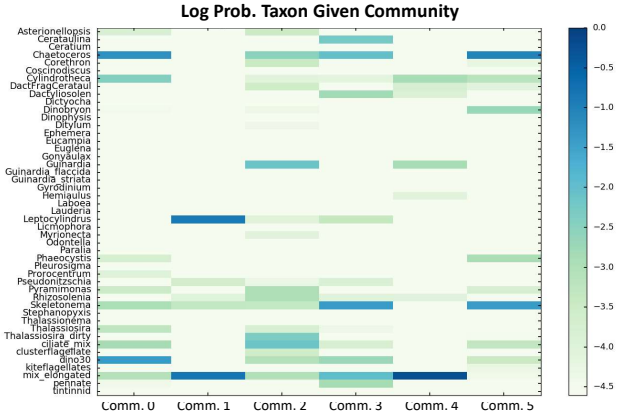


The averaged community distribution for each year-day in the best performing model shows strong seasonal structure. The five most common communities can be identified by their active and dormant seasons.



Discussion

The communities learned by our model provide an interpretable view of complex, shifting phytoplankton populations.



Our related and ongoing work includes more sophisticated regression models, applications such as search for rare taxa [3], and sample-efficient community model training techniques.

Contact Arnold Kalmbach
Mobile Robotics Lab, McGill University
akalmbach@cim.mcgill.ca
www.cim.mcgill.ca/~akalmbach

References [1] H. M. Sosik and R. J. Olson, "Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry." *Limnology and Oceanography: Methods*. 2007

[2] Girdhar, Y., Walter Cho, Campbell, M., Pineda, J., Clarke, E., & Singh, "Anomaly detection in unstructured environments using Bayesian nonparametric scene modeling." *ICRA*. 2016

[3] A. Kalmbach, Y. Girdhar, H. M. Sosik, & G. Dudek, "Phytoplankton Hotspot Prediction with an Unsupervised Spatial Community Model." *ICRA*. 2017

Acknowledgements This work was supported in part by awards to YG from NOAA through its Cooperative Institute for the North Atlantic Region (CINAR) program, and from WHOI; and to HMS from NASA's Ocean Biology and Biogeochemistry Program, and from NOAA through CINAR. We also thank the captain and crew of the Research Vessel Pisces and scientists from NOAA's Northeast Fisheries Science Center for enabling our participation in EcoMon surveys. We gratefully acknowledge the support via grant to GD of the Natural Sciences and Engineering Research Council of Canada (NSERC).