Machine learning identifies fish communities from environmental DNA (eDNA)

Henry Sun[1] (henry.sun@duke.edu), Josh Kohut[2], Jason Adolf[3]

[1]Department of Marine Science and Conservation, Duke University, [2]Department of Marine and Coastal Sciences, Rutgers University, [3]Department of Biology, Monmouth University

Background

eDNA is genetic material shed by organisms in the water column

- eDNA can cheaply measure species richness and abundance from water samples
- eDNA offers a new way to monitor fishery status and biodiversity without invasive trawl surveys

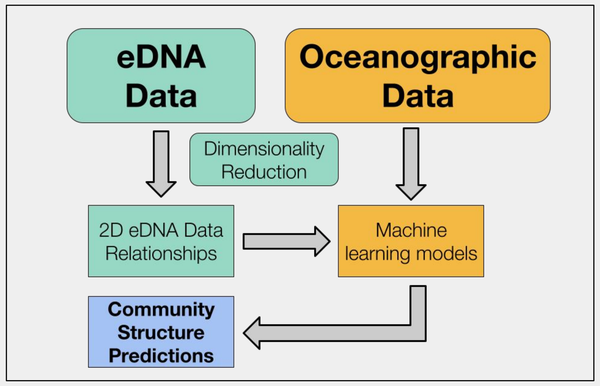

Figure 1: Project workflow. Machine learning visualizes and analyzes eDNA data, then generates predictions of fish communities from corresponding oceanographic data

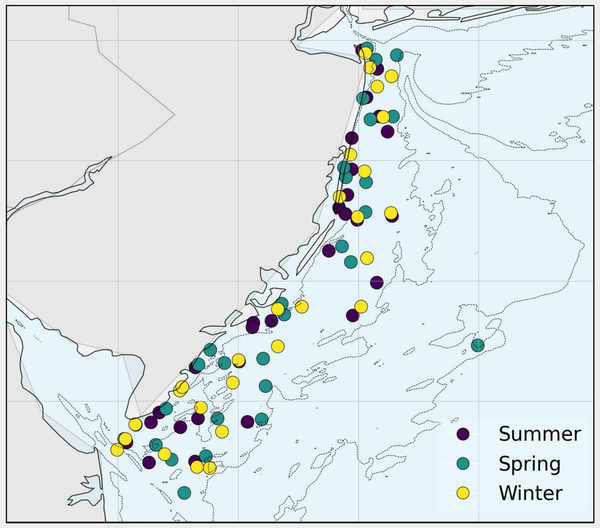

Figure 2: Data availability. The eDNA dataset used contains % abundance values for 75 species across 89 sites sampled in **winter**, **spring**, and **summer**

References & Acknowledgements

All datasets, code, and references used for this project can be found on GitHub via the QR code at the top right of the poster.

Thanks to all RIOS interns, PhD students, and staff for your unwavering support. Jason, thanks for providing all the eDNA data. Josh, thanks for your flexibility, hospitality and encouragement. This research was funded by NSF Grant OCE-1831625.

Visualizing Relationships in eDNA data

Key Takeaways

- **Dimensionality reduction (DR)** reduces complex, unvisualizable datasets into two or three dimensions
- Overall, **tSNE** performed the best at clustering and retaining variation from eDNA out of four DR methods
- **PCA** performed better on lower variability presence-absence data, while the **VAE** was consistently poor

Tested four DR methods: PCA, tSNE, UMAP, and a VAE neural network

- PCA is a **linear** method, while the others are all **nonlinear**
- Grouped stations into clusters on 2D eDNA data for method evaluation
- Expected three eDNA clusters for different sampling seasons

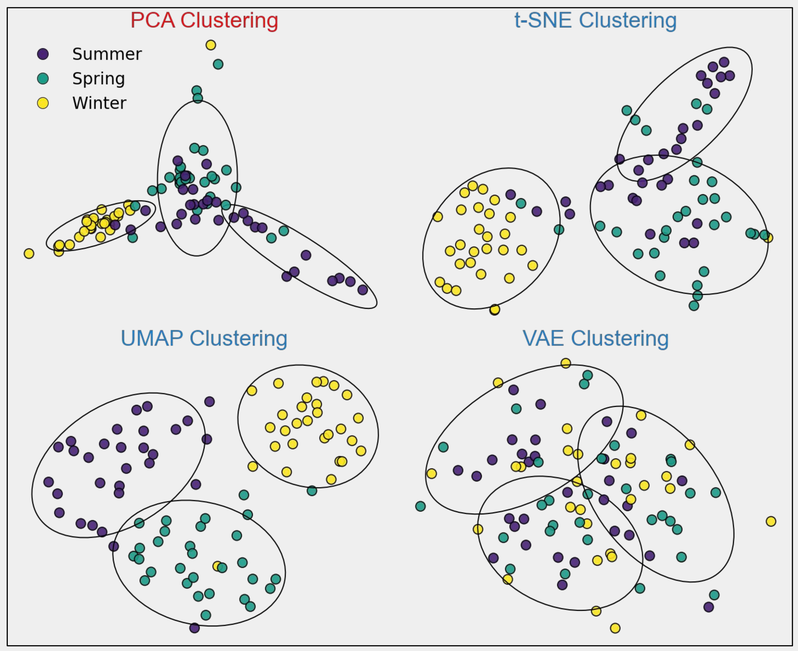

Figure 3: K-means clustering maps (K=3) of seasonal 2D encodings, circle indicates cluster region. Strong clustering by season in tSNE, PCA, and UMAP, but not with the VAE

Table 1: Comparison of linear/nonlinear DR using multiple regression (left) and K-means analysis (right). Bold = best values, ** = significant ($\alpha < 0.05$)

	Abundance	Presence	K-Means Analysis	
	R^2	R^2	CH Index	Trustworthiness Score
PCA	0.198**	**0.463****	115.297	0.583
tSNE	**0.273****	0.351**	**190.510**	0.655
UMAP	0.230**	0.370**	150.551	**0.695**
VAE	0.005	0.009	62.247	0.530

Associating eDNA with Oceanography

Key Takeaways

- **Machine learning (ML)** models can predict both fish species and community presence from oceanographic data
- ML most accurately predicted both species and community assignments in **summer**
- **39%** of the **variance** in fish communities was explained by oceanography in summer, versus just **23%** in winter

Methodology

- Built random forest ML models for winter mixed, summer stratified, and all seasons using temperature and salinity data
- First predicted if each **species** found in >20% of stations was present, then predicted each station's **community cluster assignment**

Table 2: Overall model accuracies for species and community models. ^Community model was not ran for All Seasons due to temporal cluster overlaps

	Winter	Summer	All Seasons
Species	67.7%	80.6%	82.4%
Community	16.7%	66.7%	N/A^

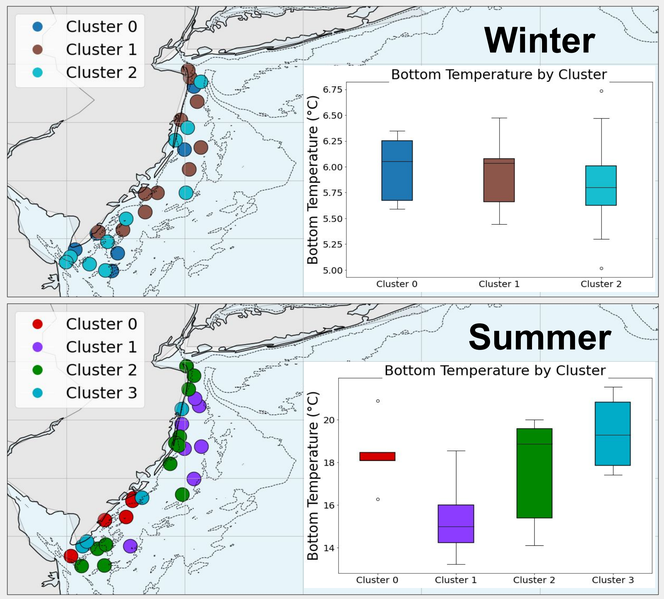

Figure 4: Community cluster locations by season and boxplots of **bottom temperatures** demonstrating ocean variability between clusters

- **Temperature** was the most important predictor for both the combined species model and summer community model
- **Dominant taxa:** Clupeiformes were the most common taxa in all winter communities, whereas the dominant taxon varied between spot (Cluster 0), Northern searobin (C1, C2), and Atlantic menhaden (C3) in summer