Classification of Ecological States using Topological Data Analysis

Douglas Krouth[1], Casey Mckean[1], Wako Bungula[1], Danelle Larson[2]

[1]University of Wisconsin-La Crosse, La Crosse, WI [2]Long-Term Resources Monitoring for the Upper Mississippi River Restoration Program, United States Geological Survey, La Crosse, WI

Introduction

- We aimed to better understand the underlying ecological conditions present within the Upper Mississippi River over the past 28 years.
  - *Ecological condition* refers to the state of ecological systems, which includes their physical, chemical, and biological characteristics and the processes and interactions that connect them. (EPA, 2020)
- Topological data analysis (TDA) provides a 2-D graph that represents a high dimensional data by implementing noise and dimension reduction techniques while retaining topological features of the data.

Primary Questions

1. Do ecological states in the Upper Mississippi River persist over time? If so, in what timeframes can these distinct ecological states be recognized?
2. Do there exist "state variables" which can be used to model the variance present within the data?

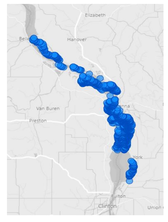
   For example, do the variables Chlorophyll, Total Phosphorous, and Turbidity represent most of the variance present within a given pool?
3. What are the different ecological states present within the Upper Mississippi River?

   Is there a primary ecological state as well as subgroups which may share similar features but have a distinct ecological characteristic?

Methods

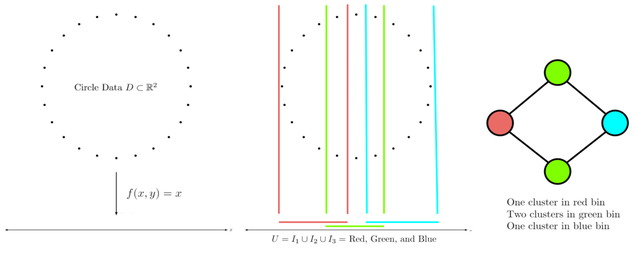
Subset of data chosen to test TDA application

- Pool 13
- Stratum : Contiguous Backwater Lakes
- Summer Season
- 28 Years of samples
- Created 3 approximately decadal periods based on major changes in water quality [8]



Steps of TDA Mapper Algorithm

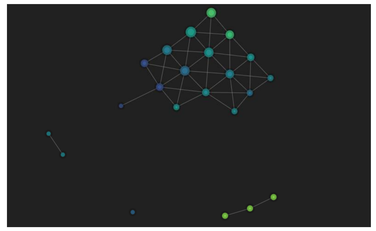
1. Let $D \subset \mathbb{R}^n$ be a data set in n-dimensional space.
2. Define $f : D \longrightarrow \mathbf{X} \subset \mathbb{R}^m$ $(m < n)$. Cover $f(D)$ with U.
3. The pre-image $f^{-1}(U)$ covers D with overlapping bins. Cluster each bin.
4. Denote each cluster by a vertex and a non-empty intersection by an edge.



State Variables

Pool 13 (1993 – 1999)

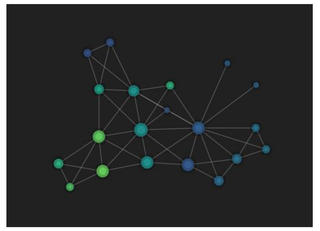
TDA Output : turbidity, total phosphorous, chlorophyll a



3 variables, 1990s
EPSILON = 30
MIN_SAMPLES = 1
N_CUBES = [5, 5]
PERC_OVERLAP = [0.35, 0.35]

	TURB	CHLcal	PredictedTP
count	486.00	486.00	486.00
mean	25.03	38.36	0.35
std	12.98	26.30	0.25
min	1.00	0.71	0.01
25%	15.25	18.96	0.20
50%	24.00	33.13	0.25
75%	34.00	47.89	0.43
max	84.00	122.85	1.35

TDA Output : All variables



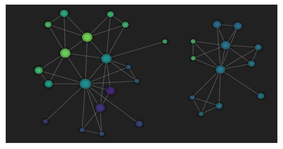
All variables, 1990s
EPSILON = 30
MIN_SAMPLES = 1
N_CUBES = [5,5]
PERC_OVERLAP = [.35,.35]

	TURB	CHLcal	PredictedTP
count	685.00	685.00	685.00
mean	29.34	31.49	0.31
std	14.21	22.97	0.20
min	1.00	1.79	0.01
25%	20.00	14.76	0.19
50%	29.00	23.93	0.24
75%	39.00	40.87	0.33
max	83.00	107.88	1.06

The structure of the mapper graph contains one primary connected component whether only the 3 variables (Turbidity, Chlorophyll, Total Phosphorous) or all quantitative variables of interest are used.

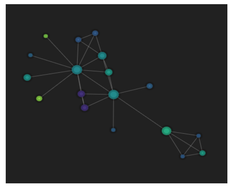
- Variance present in this subset can potentially be traced to three potential state variables

Ecological states over time



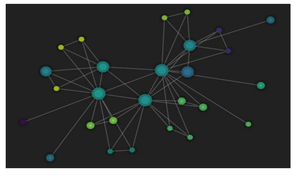
1993 – 1999

Two Primary connected components, one of which is significantly larger



2000 – 2009

One primary connected component



2010 – 2019

One primary connected component

The persistence of the primary connected component in the mapper graph across three separate decade groups indicates that there is potentially a single primary ecological state for Pool 13, Contiguous Backwater Lakes, Summer.

* Graphing Parameters: All variables, N_cubes=[5,5], perc_overlap=[5,5] DBSCAN (eps=20, min_samples=1)

Preliminary Results

- Chlorophyll a, total phosphorous, and turbidity account for 94% of the variance present in Pool 13 (all years)
- Recognition of at least one persistent ecological condition in Pool 13, Contiguous Backwater Lakes, Summer
- Majority of the variance within Pool 13 subset data can be represented as a single connected component
- Additional ecological states may exist and be defined with closer examination of statistical outliers (shown in Figure 2), as well as the second, smaller primary simplicial complex from 1993-1999 (Figure 3)

Discussion

- Further analysis is required to determine ecological characteristics of primary state present in Pool 13
- It's possible that combinations of variables other than phosphorus, chlorophyll, and turbidity are better at capturing the underlying shape of the data for other study pools or in future, long-term data.
- Our team will continue to investigate the primary questions for other pools of the Upper Mississippi River.

Acknowledgements

This work was funded as part of the U.S. Army Corps of Engineers' Upper Mississippi River Restoration Program. We thank the many Long-Term Resource Monitoring affiliated scientists and biologists that have developed and maintained a high-quality monitoring program over the past two decades. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Data Statement: All data are publicly available through the Upper Mississippi River Restoration Program's Long-term Resource Monitoring (https://umesc.usgs.gov/ltrm-home.html).

Citations

[1] Singh, Gurjeet; M´emoli, Facundo; Carlsson, Gunnar, Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, Eurographics Symposium on Point-Based Graphics, 2007.
[2] P. Y. Lum; G. Singh; A. Lehman; T. Ishkanov; M. Vejdemo-Johansson; M. Alagappan; J. Carlsson; G. Carlsson, Extracting insights from the shape of comp lex data using topology, Sci Rep 3, 1236 (2013).https://doi.org/10.1038/srep01236
[3] Ouellette, J.; Quanta Magazine; The Mathematical Shape of Big Science Data. Retrieved October 13, 2020, fromhttps://www.quantamagazine.org/the-mathematical-shape-of-big-science-data-20131004/
[4] Müllner, D. (2011, September 12). Modern hierarchical, agglomerative clustering algorithms. Retrieved October16, 2020, from https://arxiv.org/abs/1109.2378[5] Bishop, Christopher M. Pattern Recognition and Machine Learning. 2006. Vol. 1, Singapore, Springer, 2006.
[5] E. (2020, September 16). Ecological condition. Retrieved April 11, 2021, from https://www.epa.gov/report-environment/ecological-condition#:~:text=%E2%80%9CEcological%20condition%E2%80%9D%20refers%20to%20the,and%20interactions%20that%20connect%20them.
[6] van Veen et al., (2019). Kepler Mapper: A flexible Python implementation of the Mapper algorithm. Journal of Open Source Software, 4(42), 1315, https://doi.org/10.21105/joss.01315
[7] Hendrik Jacob van Veen, Nathaniel Saul, David Eargle, & Sam W. Mangham. (2019, October 14). Kepler Mapper: A flexible Python implementation of the Mapper algorithm (Version 1.4.1). Zenodo. http://doi.org/10.5281/zenodo.4077395
[8] Houser 2021, Status and Trends 3 report