

Using topic models to describe disturbance & quantify responses to environmental change

Gavin L. Simpson & Emma Wiik ▪ University of Regina

ISEC 2018 St Andrews ▪ July 5th 2018

Acknowledgements



**NSERC
CRSNG**

Slides: bit.ly/isectopicmodels

Copyright © (2018) Gavin L. Simpson Some Rights Reserved

Unless indicated otherwise, this slide deck is licensed under a Creative Commons Attribution 4.0 International License.



Community response to environmental change



By Mauricio Antón CC BY 2.5, via Wikimedia Commons

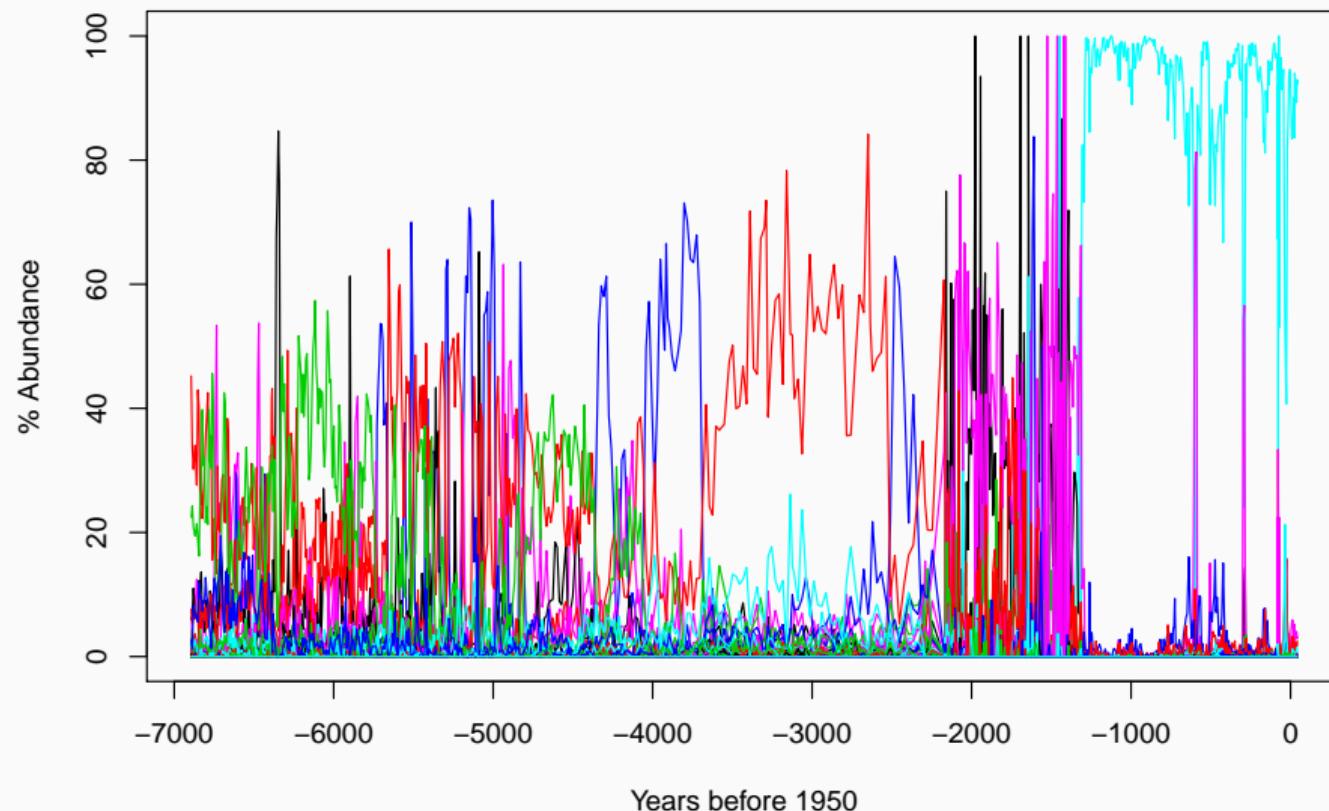
Community response to environmental change



By Credit: NOAA George E. Marsh Album [Public domain], via Wikimedia Commons

How have aquatic communities responded to these rapid changes over the last few millennial?

Complex multivariate species data



Dimension reduction

Typically we can't model all 100+ taxa in data sets like this

- (M)ARSS-like models don't like the large n

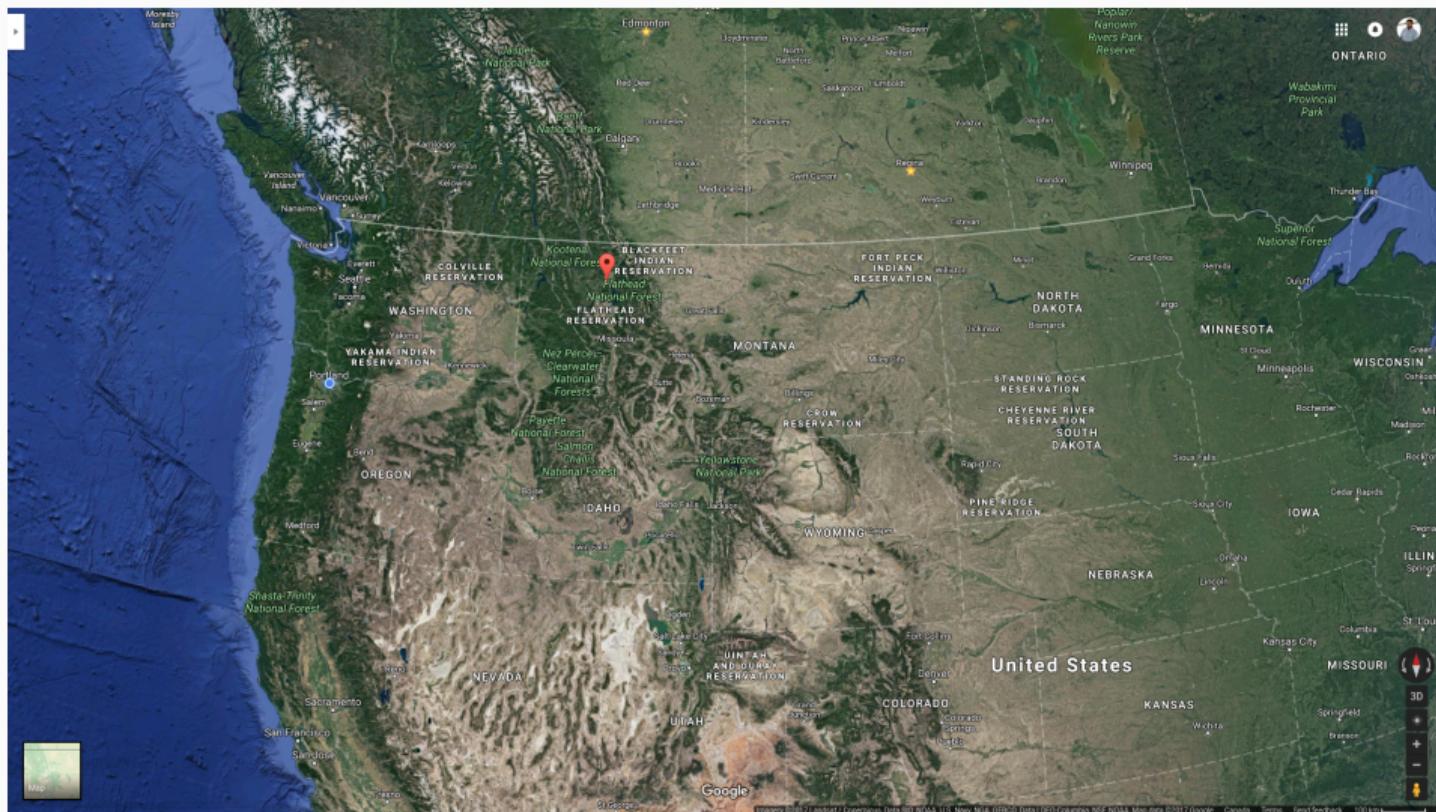
Seek a reduced dimensionality of the data that preserves the signal

Existing dimension reduction methods aren't appropriate for questions we want to ask

- Interpretation of latent factors is complex (PCA, CA, Principal Curves)

Can we group species into J **associations** and soft cluster samples as compositions of these associations?

Foy Lake — Montana



Foy Lake — Montana



Topic models

Machine learning approach for organizing text documents

- Latent Dirichlet Allocation (LDA) — (Blei, Ng, & Jordan, *J. Mach. Learn. Res.* 2003)

Generative model for word occurrences in documents

- Valle, Baiser, Woodall, & Chazdon, R. (2014) *Ecology Letters* **17**
- Christensen, Harris, & Ernest. (2018) *Ecology* doi:10.1002/ecy.2373

Community of Skittles



Individual skittles from one of four flavoured packs



What are the proportion of flavours in each pack?

How many of each pack comprise the skittle community?

Latent Dirichlet Allocation

Aim is to infer the

- distribution of skittle **flavours** within each pack β_j , and
- distribution of skittle **packs** within each community (*sample*)

Achieve a soft clustering of samples — mixed membership model

Achieve a soft clustering of **species** (flavours) into **associations** of taxa (packs)

User supplies J — the number of associations *a priori* — $j = \{1, 2, \dots, J\}$

J chosen using AIC, *perplexity*, CV, ...

Latent Dirichlet Allocation

1. Flavour distribution for j th type of Skittle

$$\beta_j \sim \text{Dirichlet}(\delta)$$

2. Proportions of each type in the Skittle community

$$\theta \sim \text{Dirichlet}(\alpha)$$

3. For each skittle s_i ,

- Choose a pack in proportion

$$z_i \sim \text{Multinomial}(\theta)$$

- Choose a flavour from chosen pack with probability

$$p(s_i | z_i, \beta_j) \sim \text{Multinomial}(\delta)$$

Correlated Topic Model

LDA assumes associations of species are **uncorrelated**

Potentially more *parsimonious & realistic* if associations were correlated

2. Proportions of each type in the Skittle community — draw

$$\eta \sim N(\mu, \Sigma)$$

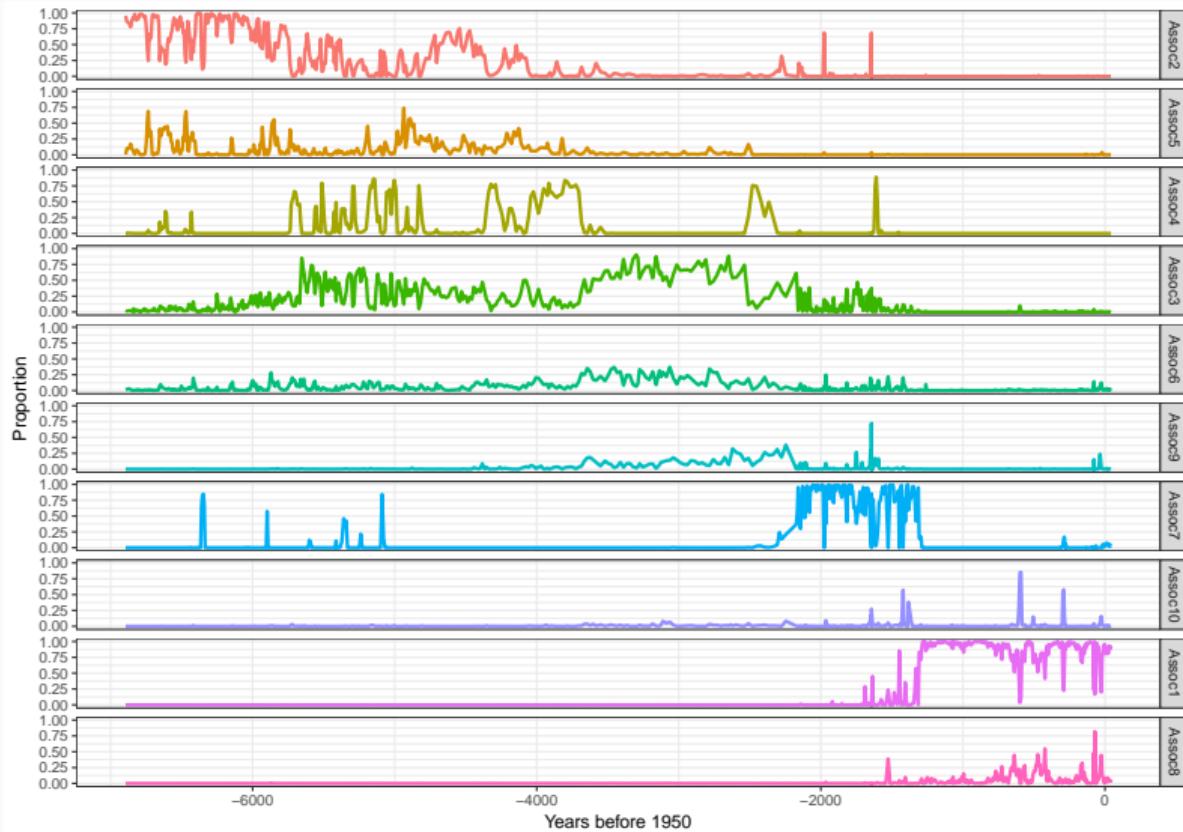
with $\eta \in \mathbb{R}^{J-1}$ and $\Sigma \in \mathbb{R}^{(J-1) \times (J-1)}$

Then transform η_J to proportional scale

Σ controls the correlation between *associations*

Blei & Lafferty (2007) A correlated topic model of Science. *Ann. Appl. Stat.* 1, 17–35.

Correlated Topic Model



Latent Dirichlet Allocation — Trend estimation

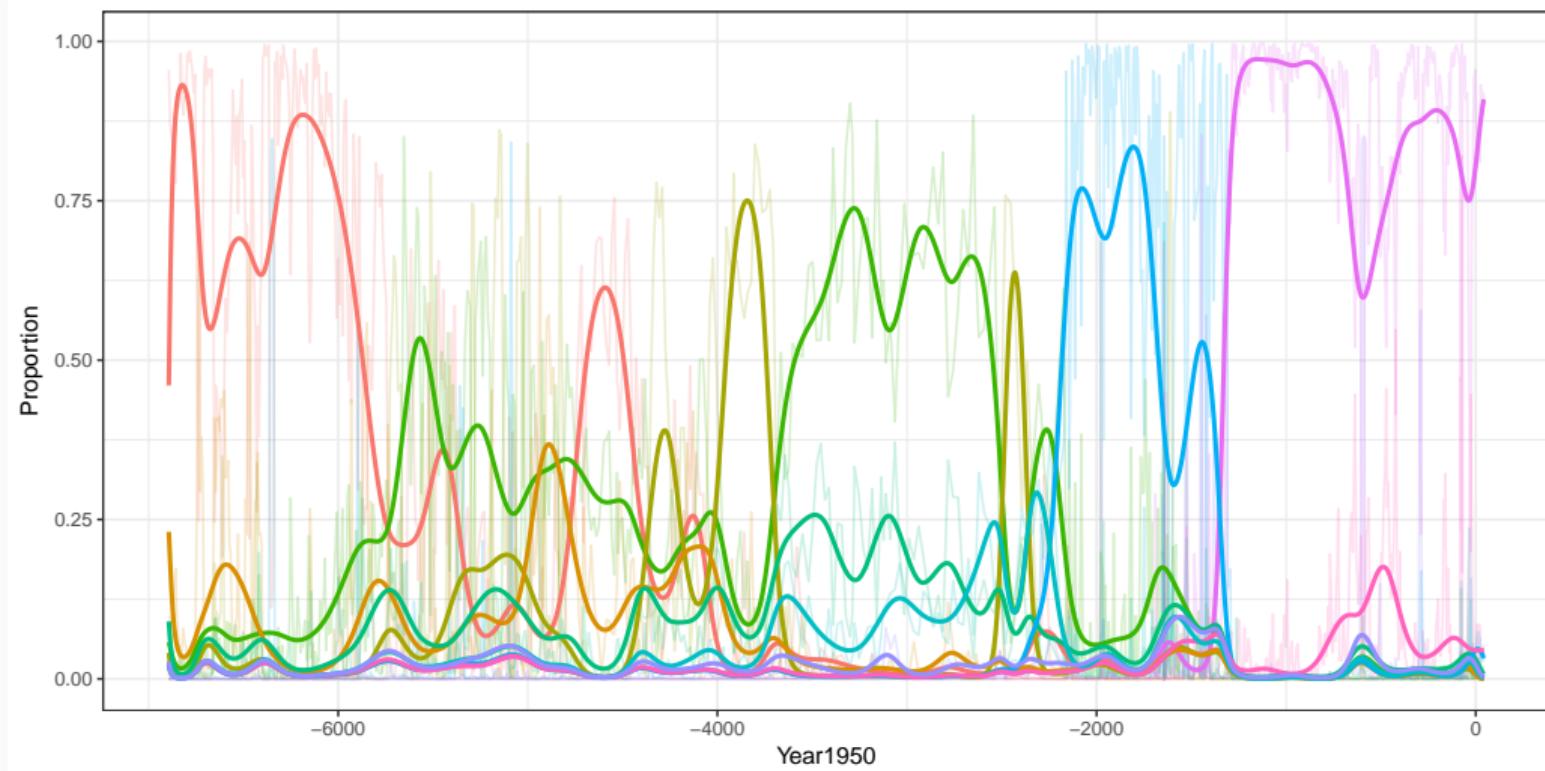
LDA knows nothing about the temporal ordering of the samples

Estimate trends in proportions of species associations using a GAM

- use adaptive spline to allow for rapid adaptation to changing data
- model each association as $\sim \text{Beta}(\mu, \phi)$

Other methods would also be appropriate: eg. Bayesian change point model or Dirichlet regression

Correlated Topic Model — Trend estimation (Dirichlet Regression)



Summary

LDA & CTM proved well-capable of summarizing the complex community dynamics of Foy Lake

- Reduced 113 taxa to 10 associations of species
- Species associations match closely the expert interpretation of the record
 - make autecological sense also
- The CTM was more parsimonious — removed one rare association
- Estimated trends in proportions of species associations capture mixture of
 - smooth, slowly varying trends, and
 - rapid (regime shift?) state change ~ 1.3 ky BP

Future directions

Choosing J is inconvenient

Address this via **Hierarchical Dirichlet Processes** and Bayesian Nonparametrics

- assume J is infinite & put a prior distribution over J

Associations in LDA & CTM are static — distributions are fixed for all samples

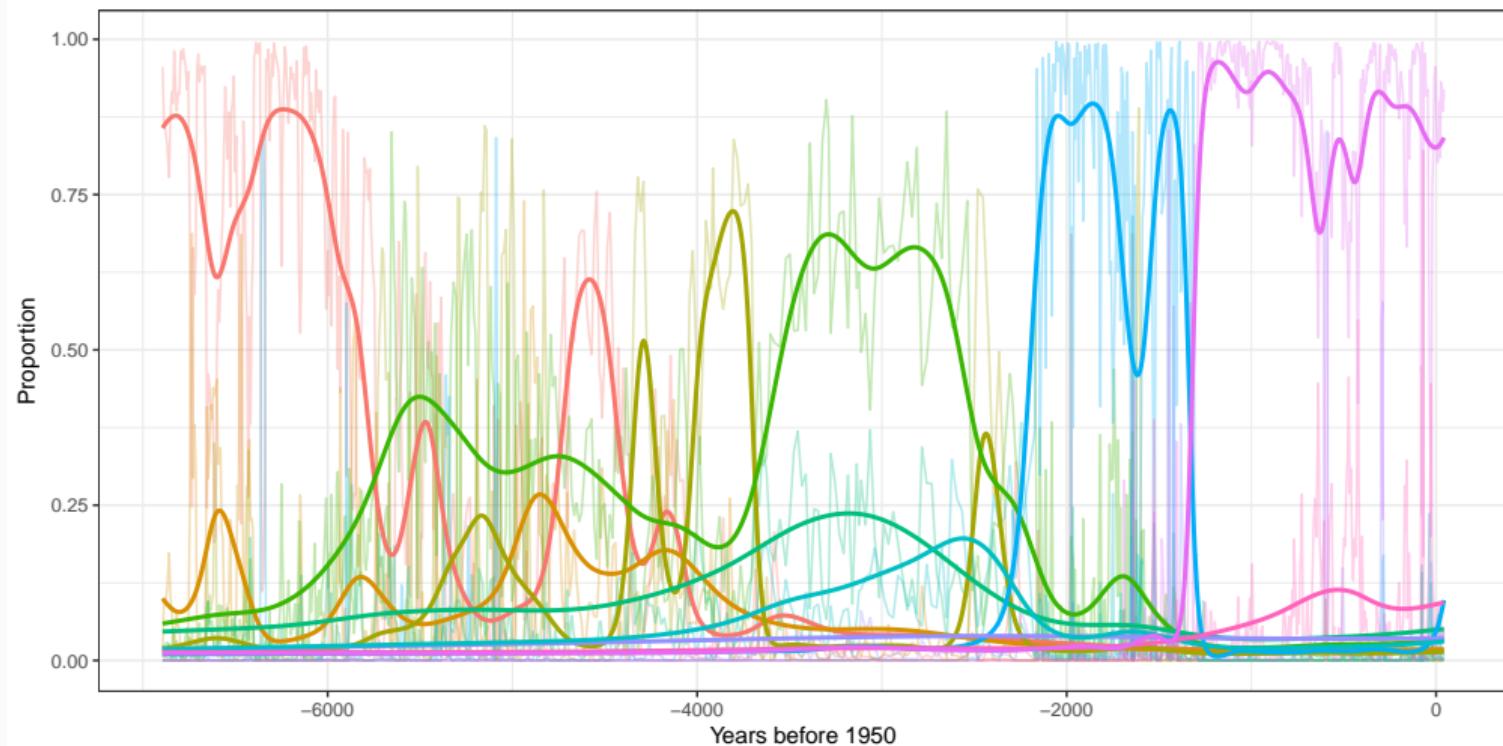
- dynamic & structural topic models allow distributions to vary smoothly with time

Many developments in this field:

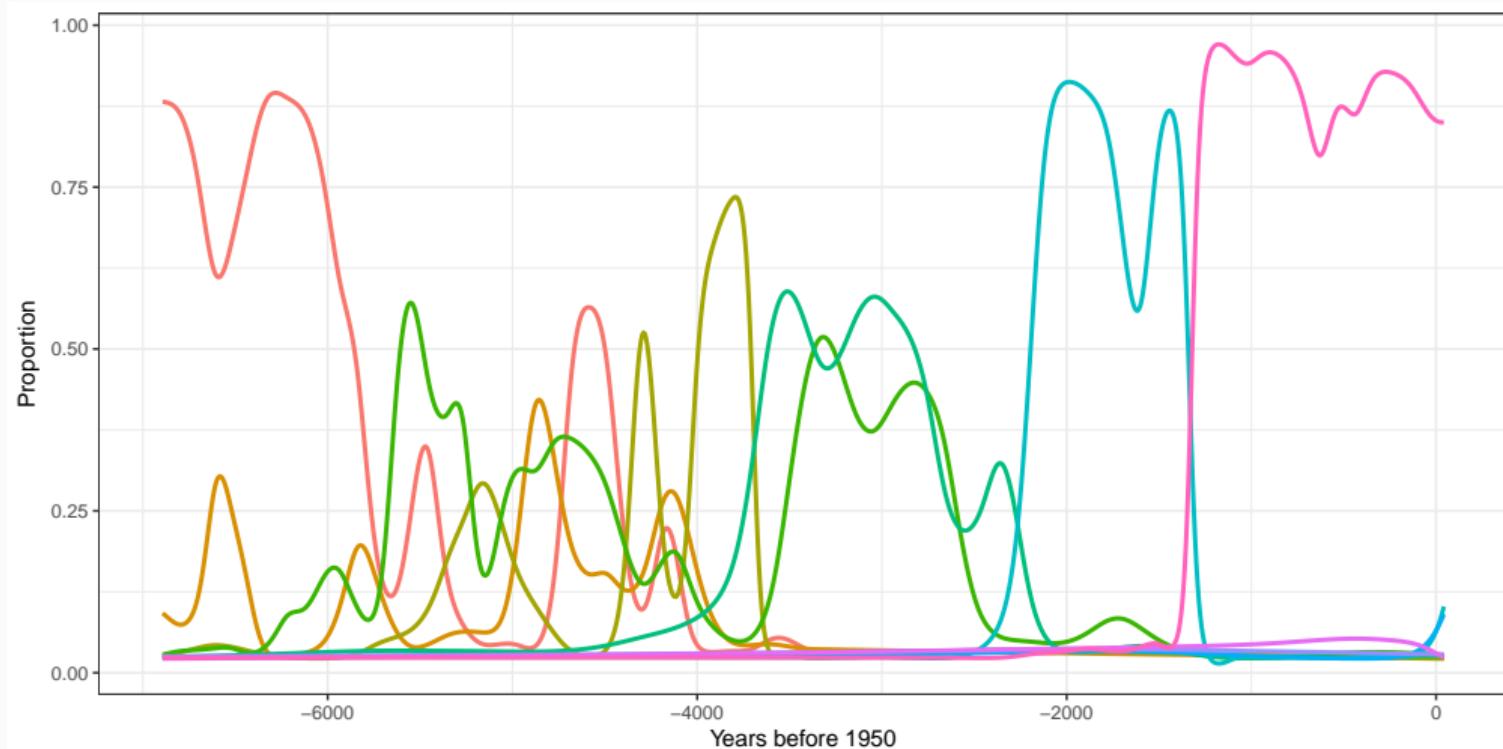
- *Chinese Restaurant Process*,
- *Indian Buffet Process*, &
- ...

Extra slides...

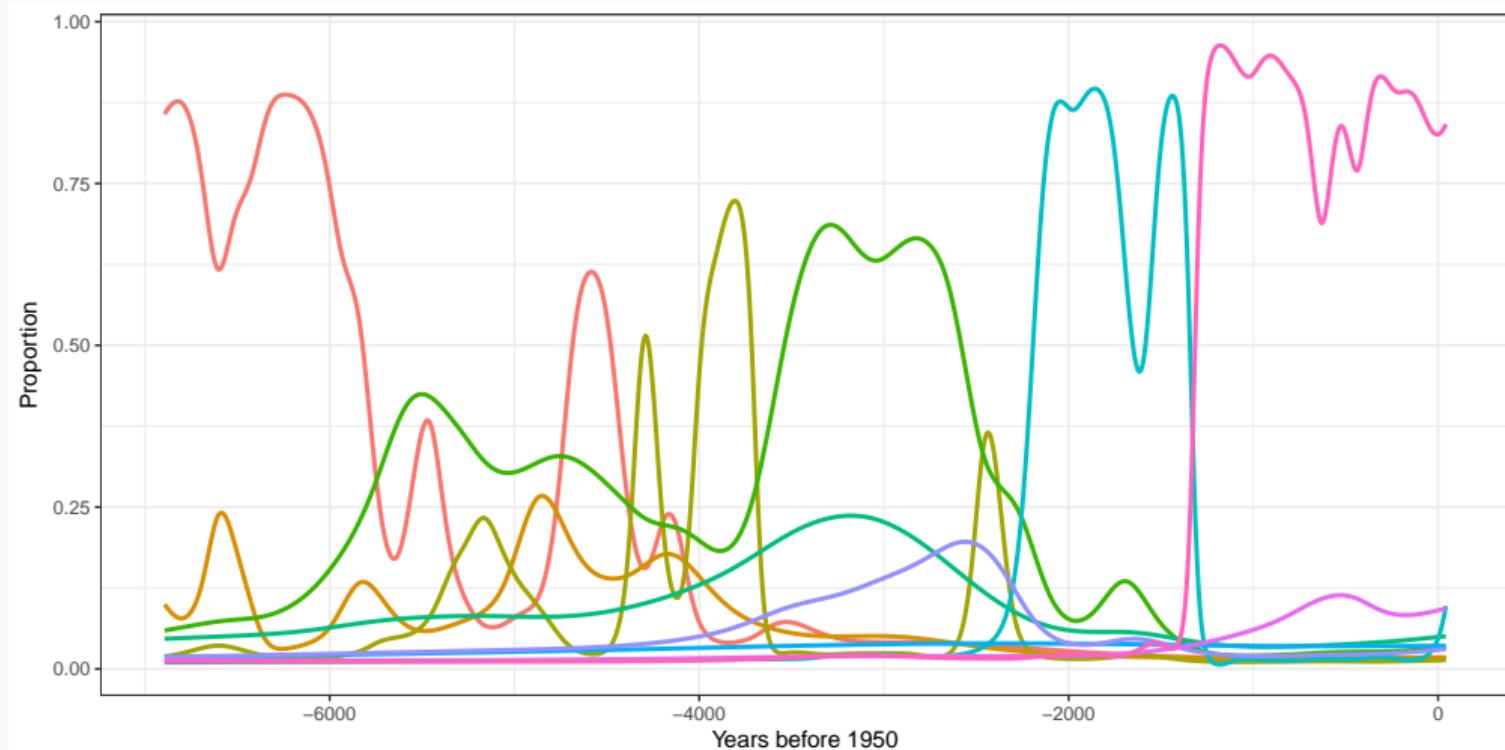
Correlated Topic Model — Trend estimation (Adaptive GAM)



Latent Dirichlet Allocation — Trend estimation (AdaptiveGAM)



Correlated Topic Model — Trend estimation (Adaptive GAM)



Intuition behind LDA

Latent Dirichlet allocation represents a trade-off between two goals

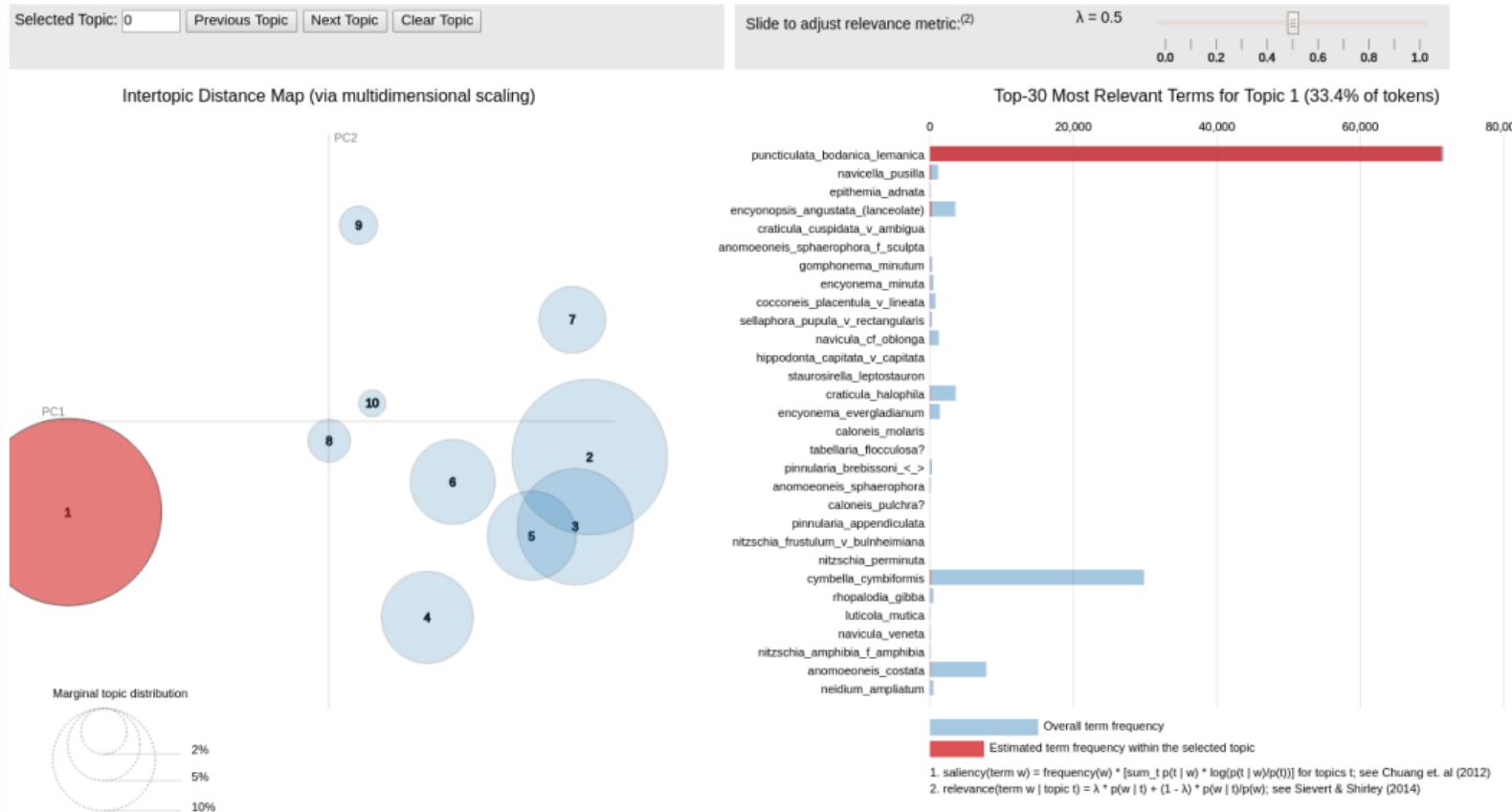
1. for each sample, allocate its individuals to a **few associations of species**
2. in each association, assign high probability to a **few species**

These are in opposition

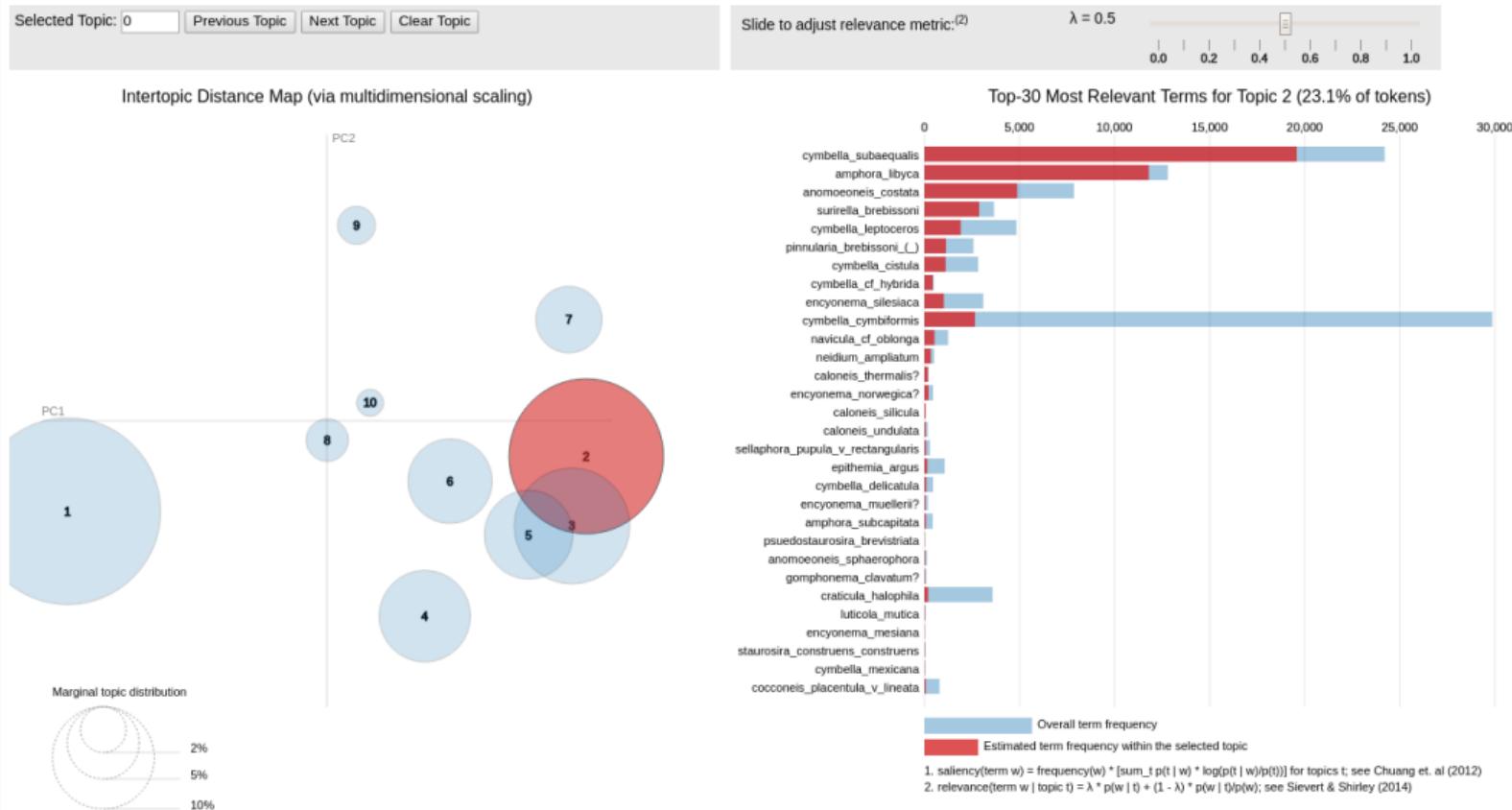
- assigning a sample to a single association makes 2 **hard** — all its species must have high probability under that one topic
- putting very few species in each association makes 1 **hard** — to cover all individuals in a sample must assign sample to many associations

Trading off these two goals therefore results in LDA finding tightly co-occurring species

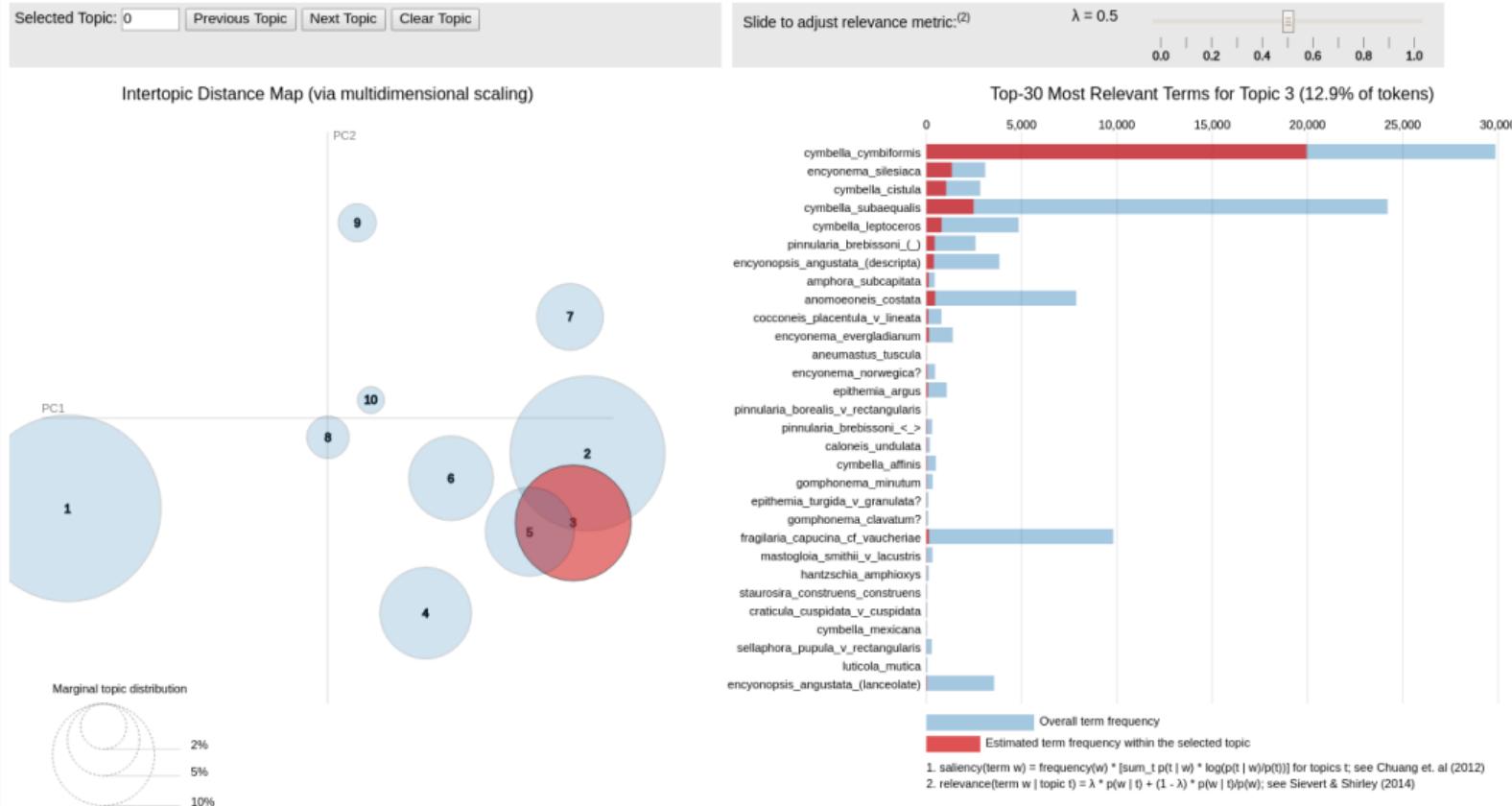
Latent Dirichlet Allocation — Association 1



Latent Dirichlet Allocation — Association 2



Latent Dirichlet Allocation — Association 3



Latent Dirichlet Allocation — Association 4

Selected Topic: 0

Slide to adjust relevance metric:⁽²⁾

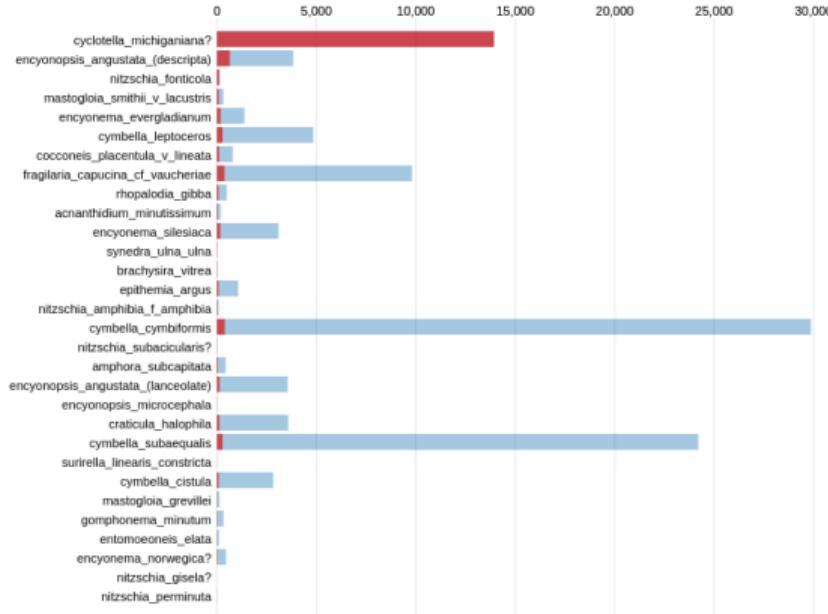
$\lambda = 0.5$



Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 4 (8.1% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_i p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t) / p(w)$; see Sievert & Shirley (2014)

Foy Lake — Montana

Foy Lake

- Deep, freshwater lake
 - Drought-sensitive Flathead River Basin
 - Diatom assemblages sensitive to lake depth variation
 - Related to variability in effective moisture

Regime shift ~1.3ka BP

Spanbauer et al PLOS ONE 9(10) e108936

OPEN ACCESS Freely available online

PLOS ONE

Prolonged Instability Prior to a Regime Shift

Trishia L. Spenhouse^{1*}, Craig R. Beier², David G. Angeler³, Tarsha Eason⁴, Sherlyn C. Fritz⁵, Ahjond S. Garmestani⁶, Kirtly L. Nash⁷, Jeffrey R. Steine⁸

¹ Department of Earth and Atmospheric Sciences, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America, ² Biological Survey, Nebraska Cooperative Fish and Wildlife Research Unit, School of Natural Resources, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America, ³ Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden, ⁴ Office of Research and Development, National Marine Fisheries Service Laboratory, Seattle, Washington, United States of America, ⁵ National Oceanic and Atmospheric Administration, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, United States of America, ⁶ Department of Earth and Environmental Systems, Indiana State University, Terre Haute, Indiana, United States of America

Abstract

Regime shifts are generally defined as the point of "abrupt" change in the state of a system. However, a seemingly abrupt transition can be the product of a system reorganization that has been ongoing much longer than is evident in a statistical analysis of a single component of the system. Under both univariate and multivariate time series modeling, we found a long-term regime shift in the Lake Erie fish community prior to the well-known 2002 regime shift. Analysis of the time series with Fisher information and multivariate time series modeling showed that there was a ~20-year period of instability prior to the regime shift. The period of instability in the ecological regime shift coincided with regional climate change, suggesting that the system was in an unstable, forcing. Palaeoecological records also provided an opportunity to test tools for the detection of thresholds and state-stationarity, and thus to examine the long-term stability of ecosystems over periods of multiple millennia.

Citation: Spenhouse TL, Beier CR, Angeler DG, Eason T, Fritz SC, et al. (2014) Prolonged Instability Prior to a Regime Shift. PLoS ONE 9(10): e109386. doi:10.1371/journal.pone.0109386

Editor: John A. D. Isaacs, University of Cambridge, United Kingdom

Received: February 1, 2014 Accepted: September 1, 2014 Published: October 3, 2014

This is an open access article under a Creative Commons Attribution License, which permits unrestricted reproduction and distribution, provided the original author and source are credited.

Funding: The Nebraska Cooperative Fish and Wildlife Research Unit is jointly supported by a cooperative agreement between the U.S. Geological Survey, the Nebraska Game and Parks Commission, the City of Lincoln, the State of Nebraska, and the Wildlife Management Institute. This work was also funded by grants from the National Sea Grant College Program (NSGCP) (NCS-13-001), the National Sea Grant Office, the National Science Foundation (NSF) (DEB-1025670) and the Sedimentary Geology & Palaeobiology program (NSF DEB-1261670). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* Email: t.spenhouse@unl.edu

Introduction

Ecosystems can undergo regime shifts and reorganize into an alternative state when a critical threshold is exceeded [1–3]. Most quantitative regime shift research has focused on abrupt shifts that have occurred during a period of human observation; this has resulted in a better understanding of what can cause regime shifts in managed ecosystems, but little is known about how slow variables (e.g. long-term changes in climate) can alter ecosystem state. Palaeoecological records can provide insight on the frequency and duration of transitions between alternative states in systems that have been sampled for thousands of years, but are not accessible to the observed recent.

To test for regime shifts in the palaeoecological record, we used a long-term high-resolution archaeological record from Lake Erie (Michigan, USA) to determine if the lake's fish community structure changed at ~1.5 to 3 thousand years before present, with profound effects at AD 1000. For Lake Erie (10°16'N, 80°13'W, 100 m elevation) is one of the Great Lakes and is located in the eastern United States. Rivers flowing into Lake Erie include the Niagara River [4,5], the Allegheny River [6] and the New York Mountains [3,6]. Diatom assemblages in this system are sensitive to changes in lake depth driven by changes in effective volume [6] and represent one metric of ecological resilience. The percent abundances of 108 diatom species were collected from a lake

sediment core that was sampled continuously at an interval of every ~10 years, yielding a ~7 kyr record of 100 m core-samples.

To determine if regime shifts could be anticipated in this palaeoecological record we (i) plotted the abundance of each diatom species, (ii) calculated the mean species abundance (measuring variance, skewed response, kurtosis, and the autocorrelation at lag 1) [7] against time, (iii) collapsed the 108 species into 10 groups based on their abundance (e.g. rare, common, 10% and 90% most abundant) and (iv) performed a principal ordination [8]. Many of these new statistical early-warning signals have been developed based on bifurcation theory, and they have recently been applied to regime shifts in lakes [9,10], but not in AD 1000. In this paper, we first describe the observed regime shifts and kurtosis in time-series data that may indicate of triggering, the rapid alternating between two different states prior to a regime shift. Then, with some simplification, we show that increasing kurtosis in time-series data can be explained by either (i) a system that is able to return from minor disturbances as it approaches a critical condition [11]—these univariate metrics can be tested in their ability to predict regime shifts in the lake, or (ii) a system that is highly sensitive to small fluctuations, which is generally not able to implement effective management actions [12]. Hence, we sought methods (I) and (ii) to detect regime shifts and regime shifts more effectively integrate the dynamics of complex multi-state systems. Finally, an integrated index based on information theory, defines as it approaches a

PLOS ONE | www.plosone.org

1

October 2014 | Volume 9 | Issue 10 | e109386