USING TOPIC MODELS TO DESCRIBE DISTURBANCE & QUANTIFY RESPONSES TO ENVIRONMENTAL CHANGE

Gavin L. Simpson & Emma Wiik · University of Regina

ESA Portland 2017 · August 9th 2017

Slides: **bit.ly/esatopicmodels**

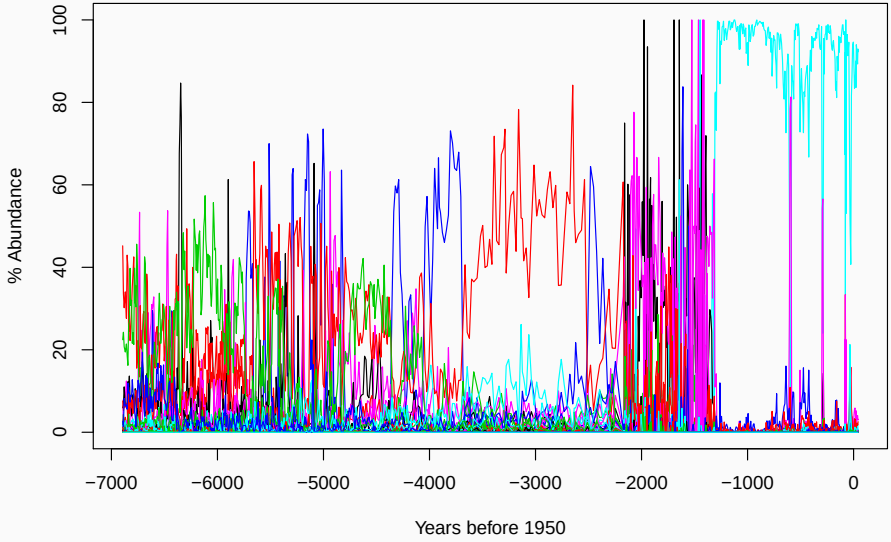
By Mauricio Antón [CC BY 2.5 (http://creativecommons.org/licenses/by/2.5)], via Wikimedia Commons

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

By Michael Knall [Public domain], via Wikimedia Commons

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

COMPLEX MULTIVARIATE SPECIES DATA



Data provided by Jeffery Stone (Indiana State University)

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

- (M)ARSS-like models don't like the large n
- Most methods assume regular spacing in time

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)
- Impose hard thresholding on the data (hierarchical clustering)
- No grouping of species or samples, or group only one

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

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

Trisha L. Spanbauer[1*], Craig R. Allen[2], David G. Angeler[3], Tarsha Eason[4], Sherilyn C. Fritz[1], Ahjond S. Garmestani[5], Kirsty L. Nash[3], Jeffery R. Stone[6]

1 Department of Earth and Atmospheric Sciences and School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America, 2 U.S. Geological 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 Risk Management Research Laboratory, U.S. Environmental Protection Agency, Cincinnati, Ohio, United States of America, 5 Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Queensland, Australia, 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 statistical analysis of a single component of the system. Using both univariate and multivariate statistical methods, we tested a long-term high-resolution paleoecological dataset with a known change in species assemblage for a regime shift. Analysis of this dataset with Fisher information and multivariate time series modeling showed that there was a ~2000 year period of instability prior to the regime shift. This period of instability and the subsequent regime shift coincide with regional climate change, indicating that the system is undergoing extrinsic forcing. Paleoecological records offer a unique opportunity to test tools for the detection of thresholds and stable-states, and thus to examine the long-term stability of ecosystems over periods of multiple millennia.

Citation: Spanbauer TL, Allen CR, Angeler DG, Eason T, Fritz SC, et al. (2014) Prolonged Instability Prior to a Regime Shift. PLoS ONE 9(10): e108936. doi:10.1371/journal.pone.0108936

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

Received February 11, 2014; **Accepted** September 3, 2014; **Published** October 3, 2014

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 University of Nebraska–Lincoln, the United States Fish and Wildlife Service, and the Wildlife Management Institute. This work was supported in part by the August T. Larsson Foundation of the Swedish University of Agricultural Sciences and the NSF's Integrative Graduate Education and Research Traineeship (IGERT) program (NSF #0903469) and the Sedimentary Geology & Paleobiology program (NSF #1203678). 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.

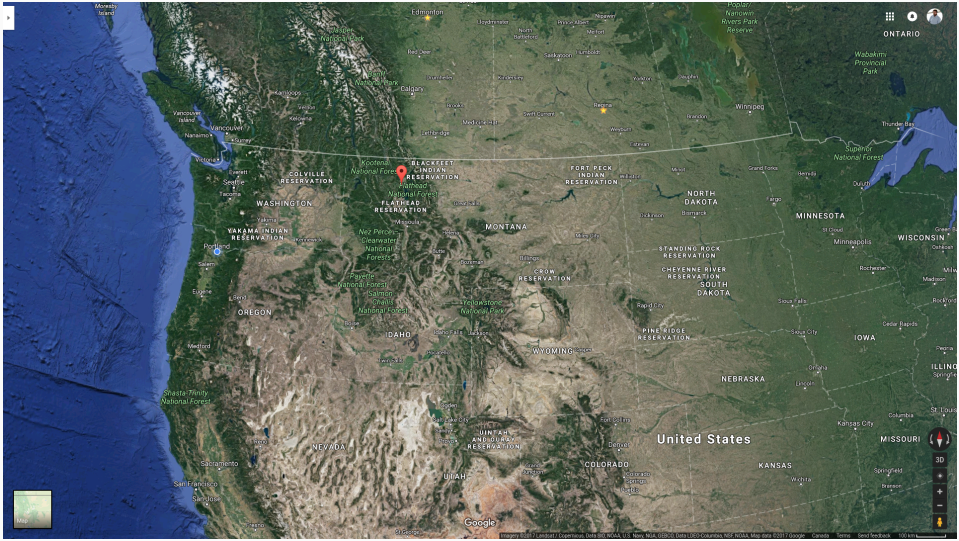
* Email: tspanbauer@unl.edu

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 how fast variables (e.g. nutrient loading/trade resilience, but it hasn't addressed how slow variables (e.g. long-term changes in climate) can alter ecosystem state. Paleoecological records can provide insight on the frequency and duration of transitions between alternative states in systems that are affected by both fast and slow variables, at timescales not accessible in the observed record.

To test for regime shifts in the paleoecological record, we used a long-term high-resolution sedimentological record from Foy Lake (Montana, USA) that showed abrupt changes in diatom community structure at ~ 1.3 ka (thousands of years before present, with present defined as AD 1950). Foy Lake (48.166ºN, 114.539ºW, 1003 m elevation) is a deep freshwater lake situated in the drought-sensitive Flathead River Basin in the Northern Rocky Mountains [4,5]. Diatom assemblages in this system are sensitive to changes in lake depth driven by changes in effective moisture [6] and represent one metric of ecological resilience. The percent abundances of 109 diatom species were collected from a lake sediment core that was sampled continuously at an interval of every ~5–20 years, yielding a ~7 kyr record of 400 time-steps.

To determine if regimes shifts could be anticipated in this paleoecological data set we 1) plotted several indicators proposed to be early-warning signals of approaching critical thresholds (increasing variance, skewed responses, kurtosis, and the autocorrelation at lag-1) [7] against time, 2) collapsed the 109 species variables into the system's mean Fisher information (FI; [8]), and 3) used multivariate time series modelling based on canonical ordination [9]. Many of these statistical early-warning signals have been developed based on bifurcation theory, and they have successfully anticipated regime shifts in many [10–13], but not all [14] systems tested. Increasing variance, skewed responses, and kurtosis in time-series data may be indicative of flickering, the rapid alternating between two different states prior to a regime shift [15]. Along with autocorrelation at lag-1, increasing variance in time-series data can be caused by critical slowing down, where a system is slow to recover from minor disturbances as it approaches a critical transition [7]. These univariate metrics can be limited in their utility, because appreciable signals often occur at the onset of the regime shift, which is generally too late to implement effective management actions [16]. Hence, we sought methods (FI and multivariate time series modelling) that more effectively investigate the dynamics of complex multivariate systems. FI, an integrated index based on information theory, declines as it approaches a


ONTARIO
Edmonton
Saskatoon
Regina
Winnipeg
Thunder Bay
Superior National Forest
Vancouver Island
Vancouver
Calgary
Seattle
Tacoma
COLVILLE RESERVATION
BLACKFEET INDIAN RESERVATION
FORT PECK INDIAN RESERVATION
FLATHEAD RESERVATION
WASHINGTON
YAKAMA INDIAN RESERVATION
MONTANA
NORTH DAKOTA
MINNESOTA
WISCONSIN
Minneapolis
Portland
Salem
Eugene
OREGON
STANDING ROCK RESERVATION
CHEYENNE RIVER RESERVATION
SOUTH DAKOTA
CROW RESERVATION
IDAHO
Boise
WYOMING
PINE RIDGE RESERVATION
IOWA
NEBRASKA
Omaha
Lincoln
Kansas City
Columbia
St. Louis
MISSOURI
Shasta-Trinity National Forest
NEVADA
Sacramento
San Francisco
San Jose
UINTAH AND OURAY RESERVATION
United States
COLORADO
Denver
KANSAS
UTAH
Google

FOY LAKE — MONTANA

Machine learning approach for organizing text documents

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

generative model

Valle, D., Baiser, B., Woodall, C. W. & Chazdon, R. (2014) *Ecology Letters* **17**

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?

Aim is to infer the

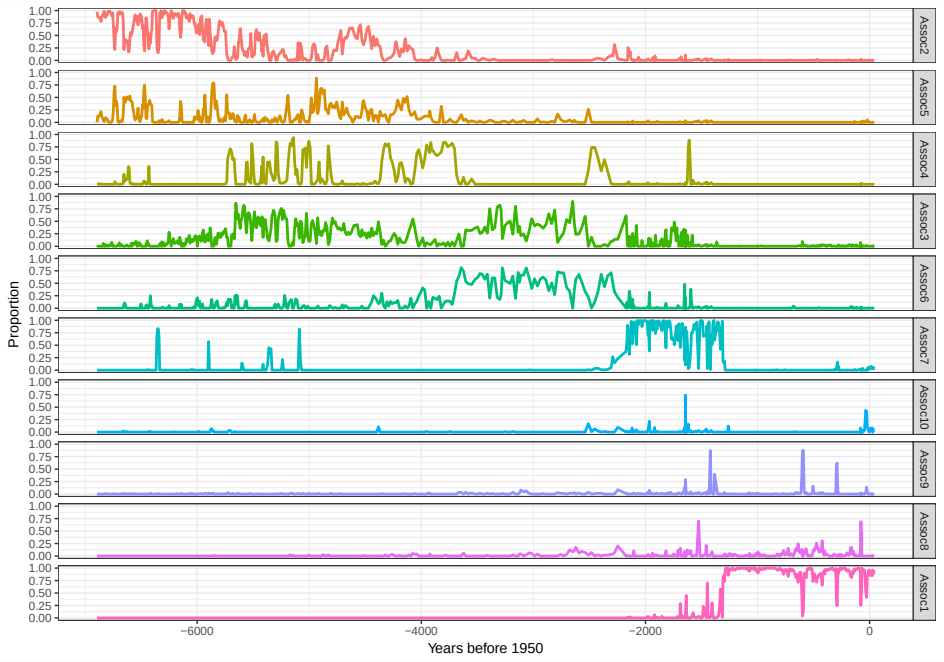
- distribution of skittle **flavours** within each pack β_j, and
- distribution of skittle **packs** within each community (*sample*)

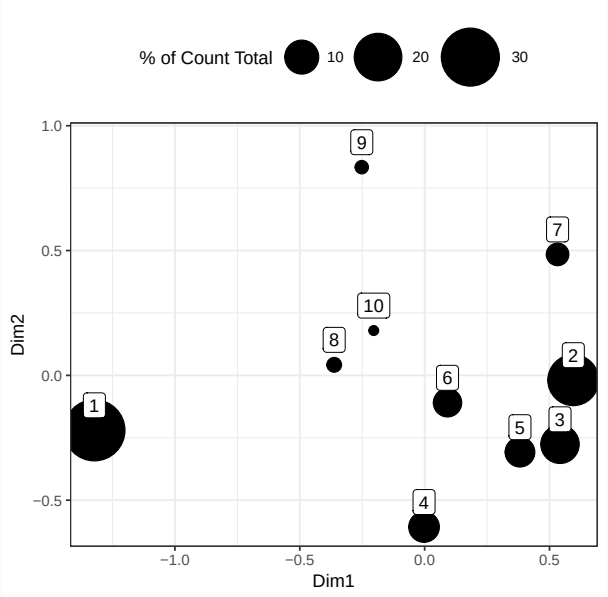
Achieve a soft clustering of samples — mixed membership model

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

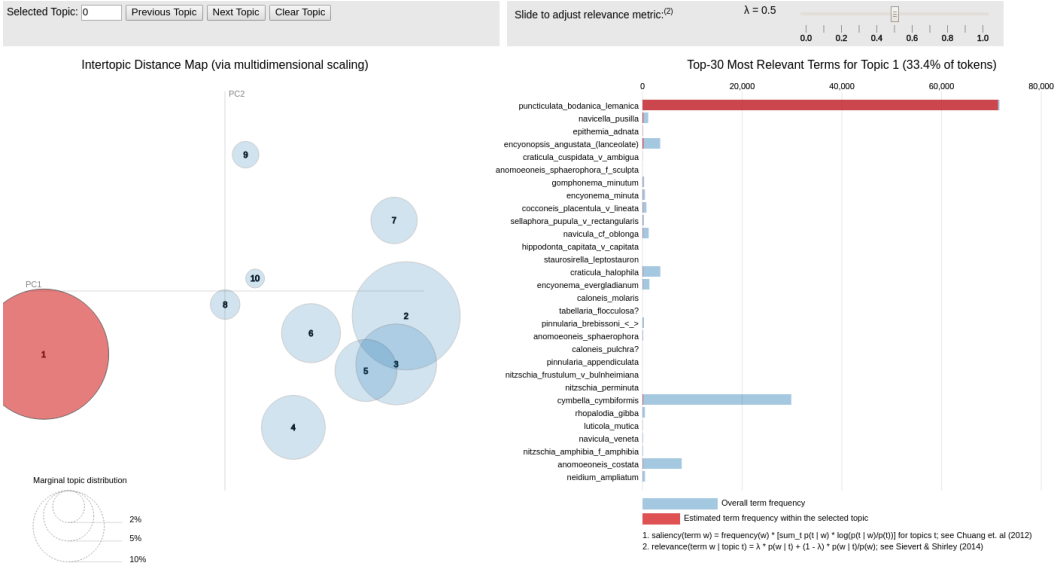
User supplies J — the number of associations *a priori* — $j = \{1, 2, \ldots, J\}$

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

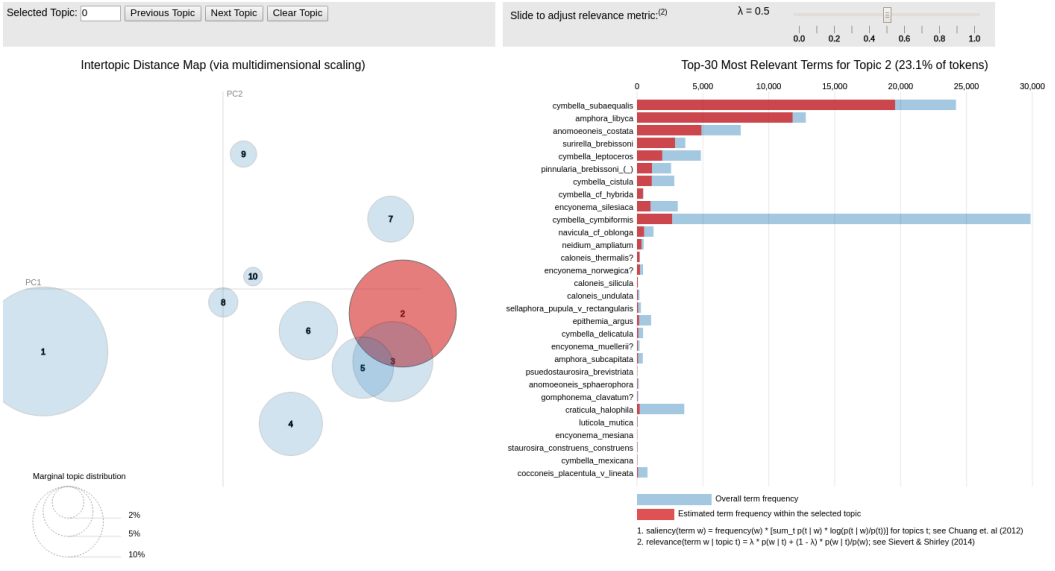

Proportion
1.00
0.75
0.50
0.25
0.00
Assoc2
Assoc5
Assoc4
Assoc3
Assoc6
Assoc7
Assoc10
Assoc9
Assoc8
Assoc1
−6000
−4000
−2000
0
Years before 1950


% of Count Total
10
20
30
1.0
0.5
0.0
−0.5
Dim2
−1.0
−0.5
0.0
0.5
Dim1
9
7
10
8
2
6
1
5
3
4

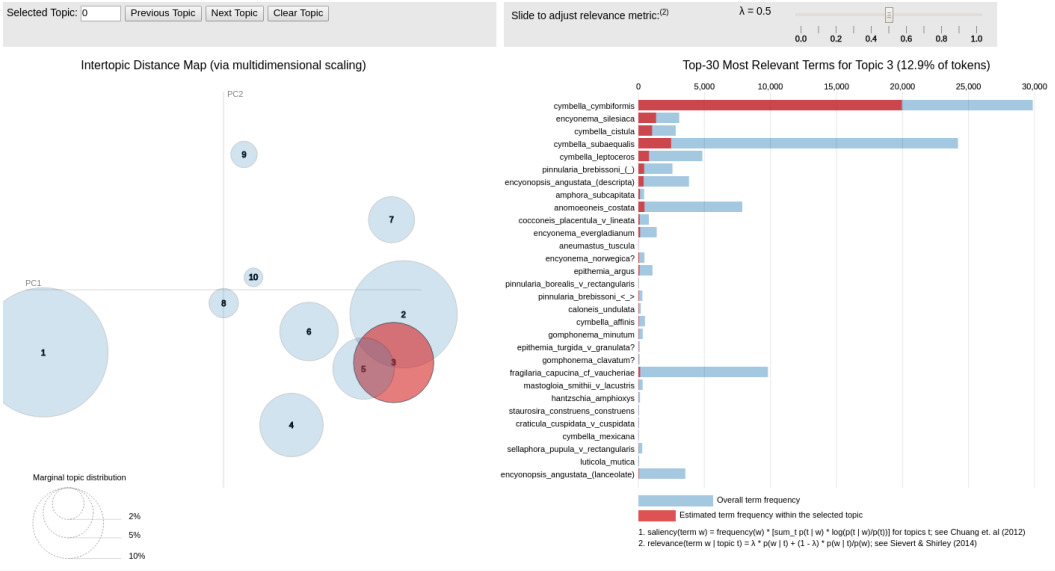
LATENT DIRICHLET ALLOCATION — ASSOCIATION 1

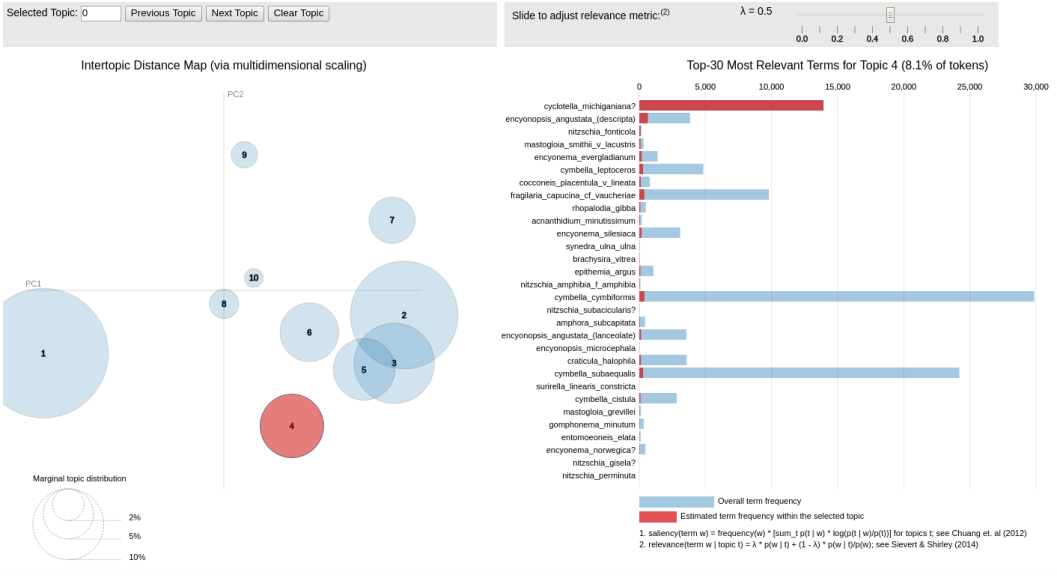


LATENT DIRICHLET ALLOCATION — ASSOCIATION 2



LATENT DIRICHLET ALLOCATION — ASSOCIATION 3




Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.5
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 4 (8.1% of tokens)
0 5,000 10,000 15,000 20,000 25,000 30,000
cyclotella_michiganiana?
encyonopsis_angustata_(descripta)
nitzschia_fonticola
mastogloia_smithii_v_lacustris
encyonema_evergladianum
cymbella_leptoceros
cocconeis_placentula_v_lineata
fragilaria_capucina_cf_vaucheriae
rhopalodia_gibba
acnanthidium_minutissimum
encyonema_silesiaca
synedra_ulna_ulna
brachysira_vitrea
epithemia_argus
nitzschia_amphibia_f_amphibia
cymbella_cymbiformis
nitzschia_subacicularis?
amphora_subcapitata
encyonopsis_angustata_(lanceolate)
encyonopsis_microcephala
craticula_halophila
cymbella_subaequalis
surirella_linearis_constricta
cymbella_cistula
mastogloia_grevillei
gomphonema_minutum
entomoeoneis_elata
encyonema_norwegica?
nitzschia_gisela?
nitzschia_perminuta
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

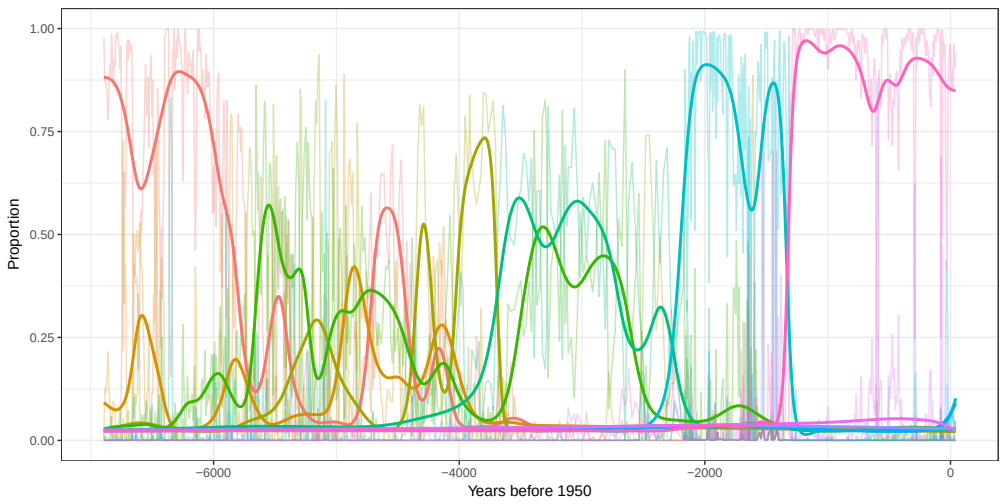
LDA knows nothing about the temporal ordering of the samples

Estimate trends in proportions of species associations using a GAM

- assumes smoothly varying trends
- use adaptive spline to allow for rapid adaptation to changing data
- model each association as $\sim$ Beta(μ, ϕ)

Other methods would also be appropriate: eg. Bayesian Change point model

LATENT DIRICHLET ALLOCATION — TREND ESTIMATION (ADAPTIVE GAM)



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

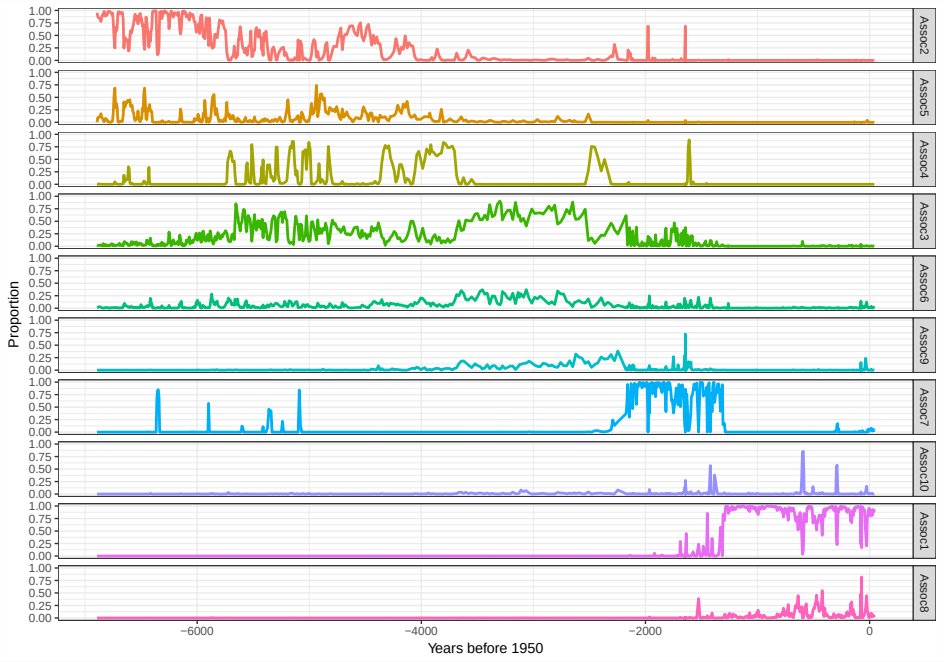
$$\eta \sim N(\mu, \Sigma)$$

with $\eta \in \mathbb{R}^{J-1}$ and $\Sigma \in \mathbb{R}^{(J-1)\times(J-1)}$
Then transform η_j to proportional scale

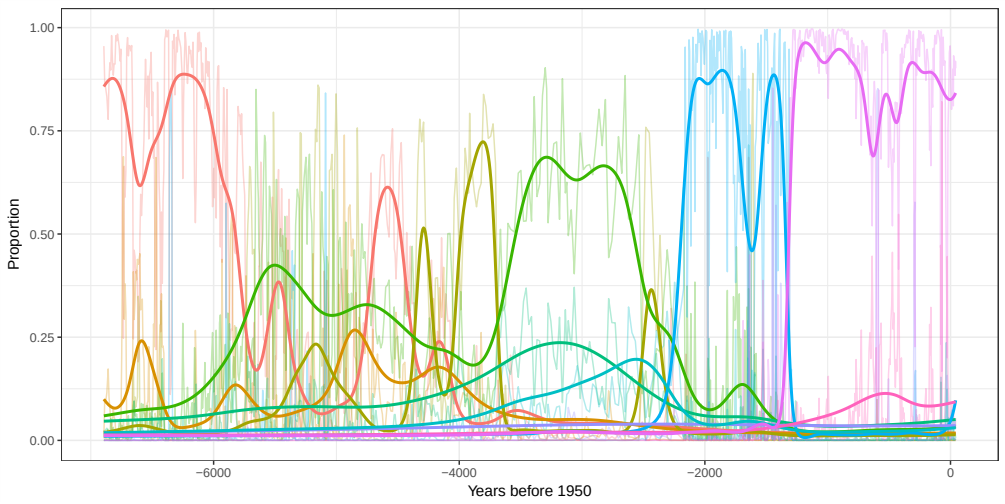
Σ controls the correlation between *associations*

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

CORRELATED TOPIC MODEL



CORRELATED TOPIC MODEL — TREND ESTIMATION (ADAPTIVE GAM)



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
    - smooth, slowly varying trends, and
    - rapid (regime shift?) state change ~ 1.3 ky BP

Choosing J is inconvenient

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

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

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

- dynamic topic models allow distributions to vary smoothly with time

Many developments in this field:

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

1. Flavour distribution for jth type of Skittle

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

2. Proportions of each type in the Skittle community

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

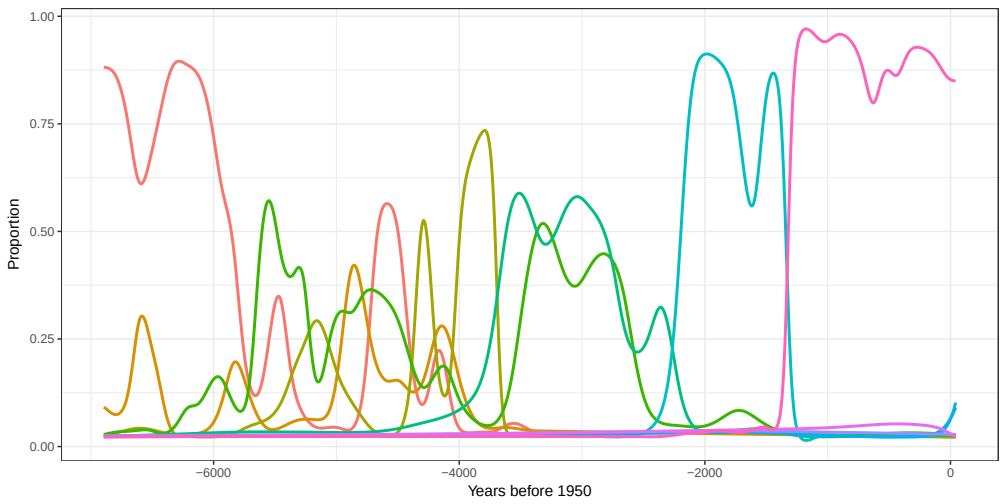
3. For each skittle s_i
   - Choose a pack in proportion

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

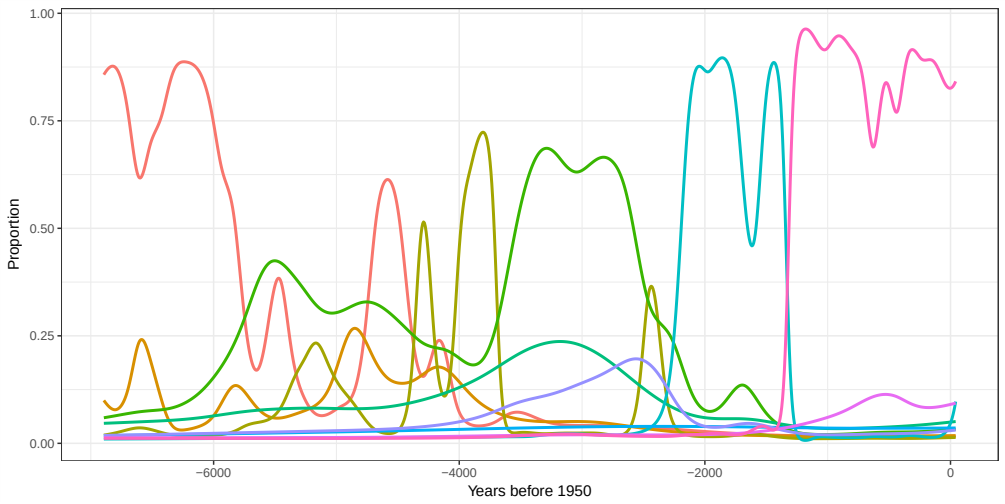
   - Choose a flavour from chosen pack with probability

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

LATENT DIRICHLET ALLOCATION — TREND ESTIMATION (ADAPTIVEGAM)



CORRELATED TOPIC MODEL — TREND ESTIMATION (ADAPTIVE GAM)



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