Recovering / Remembering Ecology

Gavin L. Simpson · Aarhus University

Palaeo SIG · May 25–28th 2021

Acknowledgements

NSERC CRSNG

Slides: **bit.ly/isectopicmodels**

Recovering / Remembering Ecology

Representing data

How do we represent the world in data and models?

Vignettes:

1. Dimension reduction via **topic models**
2. Network representations & species interactions(?)
3. Models beyond the mean (Friday)

Tradition

- Typically proportional composition
- Dimension reductions via ordination
- Clustering
- Dissimilarity
- Approximations

Maslow's hammer

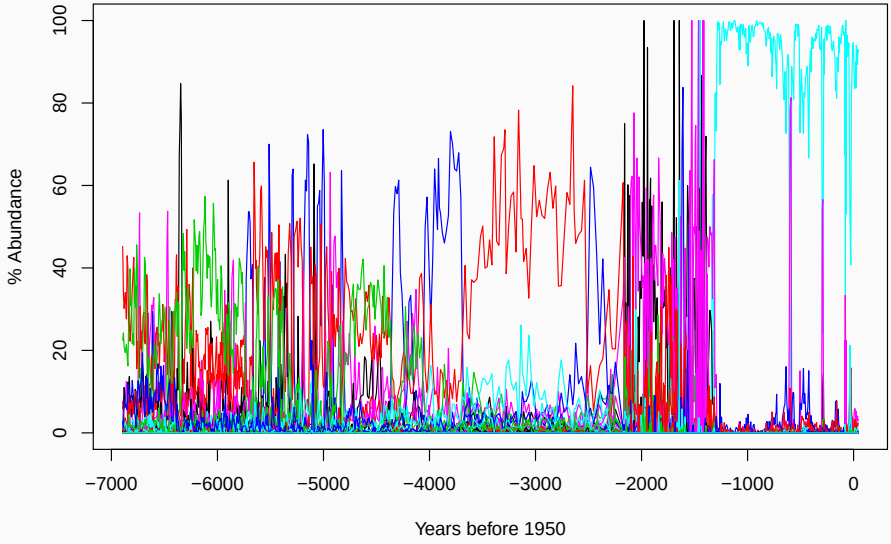
{Anne Nygard}

Topic Models

Community response to environmental change

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)}

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

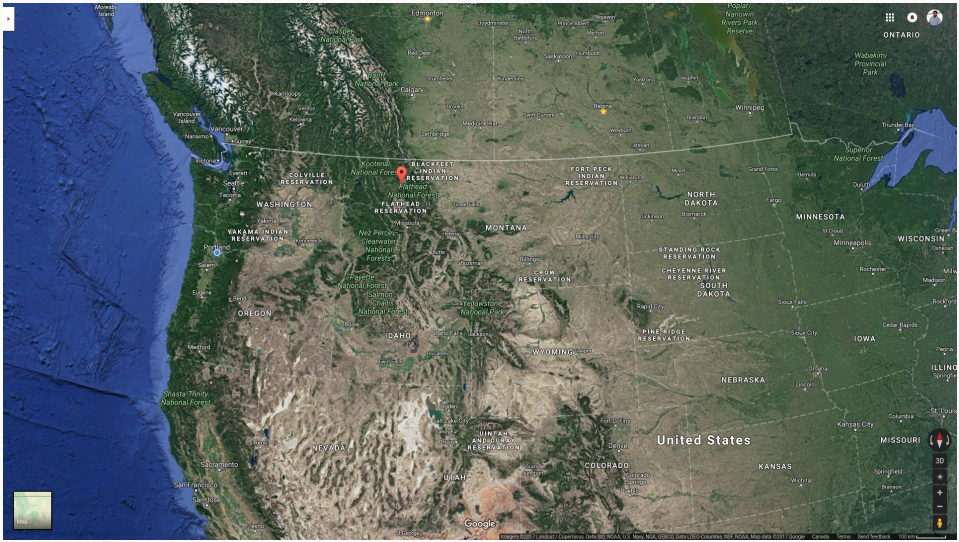
Existing dimension reduction methods aren't appropriate for questions we want to ask

- Interpretation of latent factors is complex (PCA, CA, Principal Curves)
- Linear combination of all taxa

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

Twinspan?

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 occurences 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* — $j = \{1, 2, \ldots, J\}$

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

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

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

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

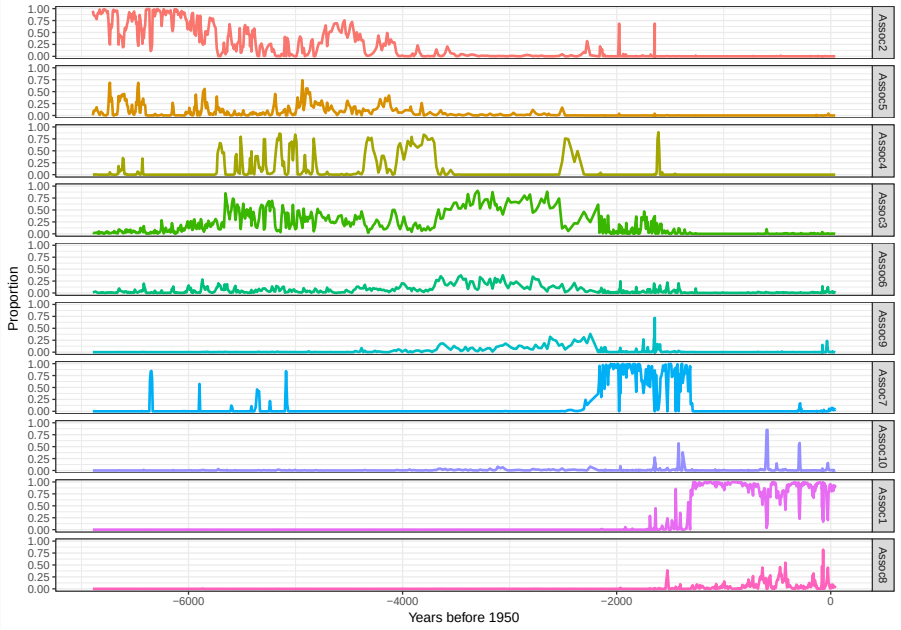
with $\eta \in \mathbb{R}^{J-1}$ and $\Sigma \in \mathbb{R}^{(J-1)\times(J-1)}$

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

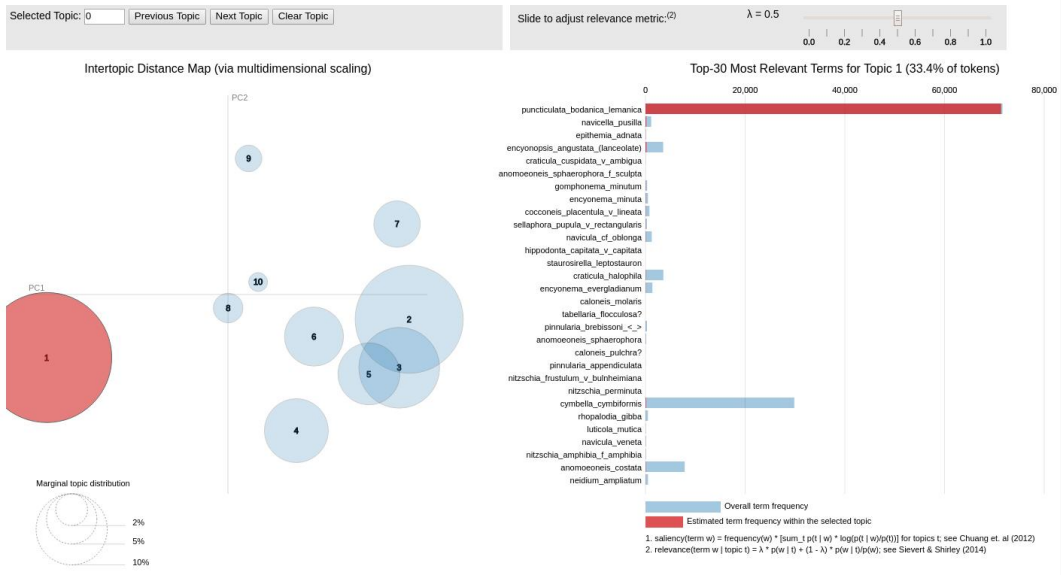


Displaying results

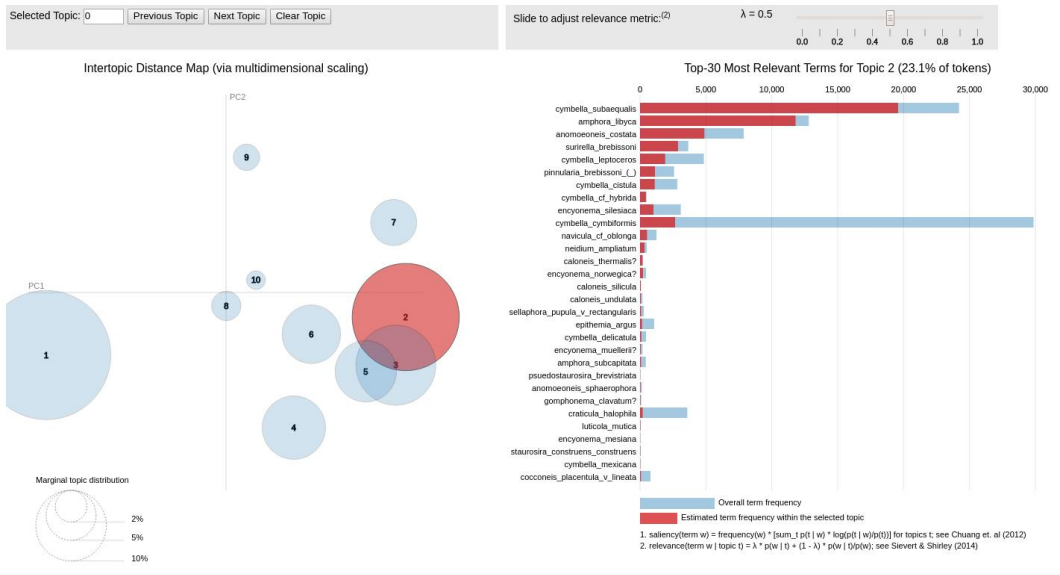
Traditionally large corpus — *interactive* exploration of results

- plot topic proportions over time
- ordinatate topic proportions
    - $\text{PCA}_{\text{Hellinger}}$, CA, etc

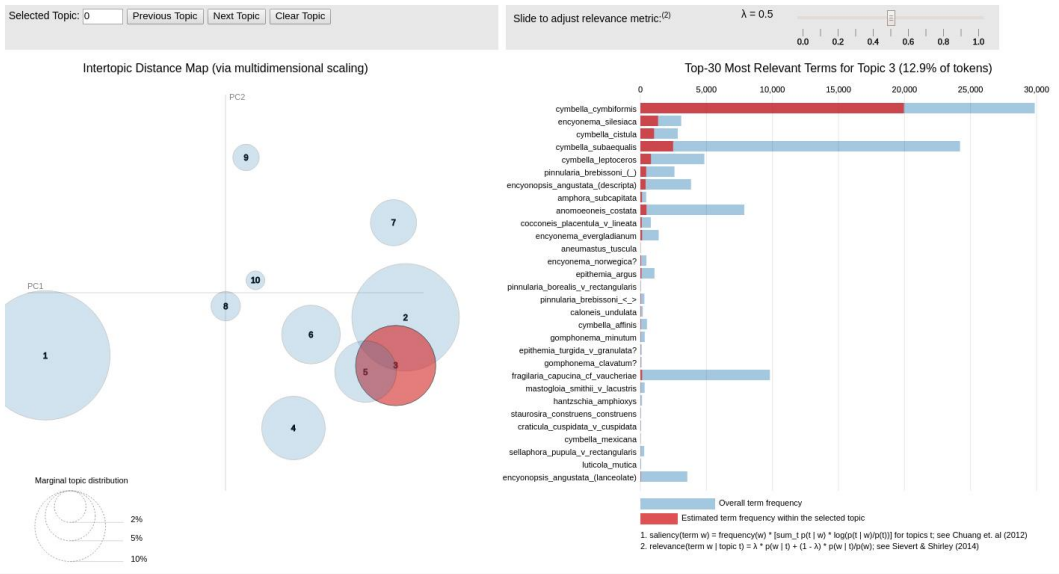
Latent Dirichlet Allocation — Association 1



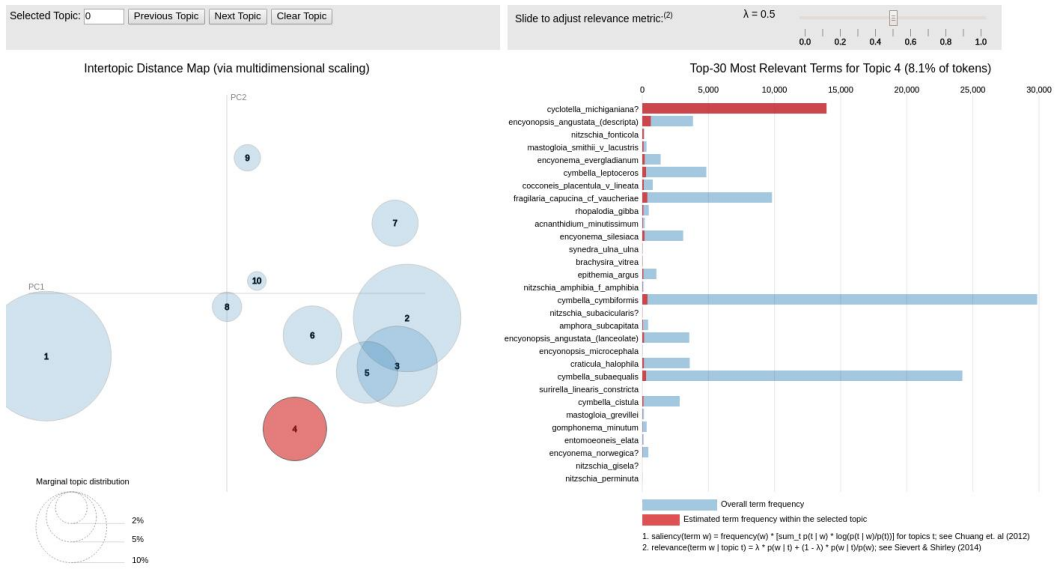
Latent Dirichlet Allocation — Association 2



Latent Dirichlet Allocation — Association 3



Latent Dirichlet Allocation — Association 4



Latent Dirichlet Allocation — Trend estimation

LDA knows nothing about the temporal ordering of the samples

A generative model for the data

Treat the topic proportions as data into some other model — *explanation*

Estimate trends in proportions of species associations using Dirichlet regression

Dirichlet regression

Dirichlet distribution describes multivariate proportions on a *simplex*

% sand, silt, clay

Only need to know two of these for complete information — total is 100%

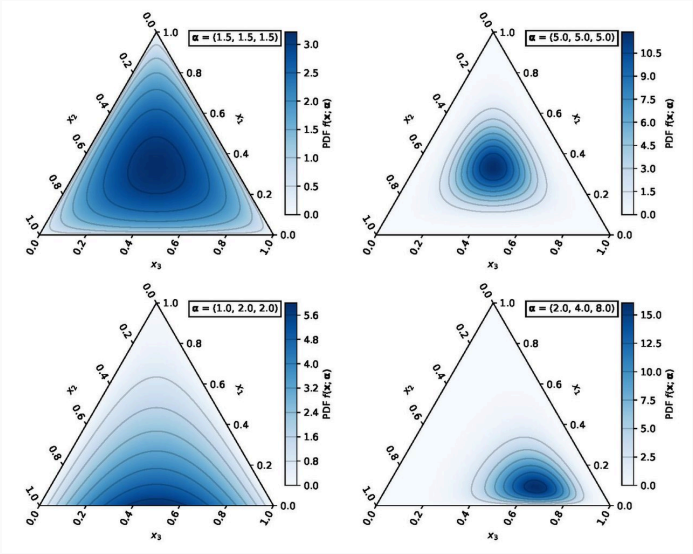
Closed compositional data

Multinomial is for *counts* from a total count

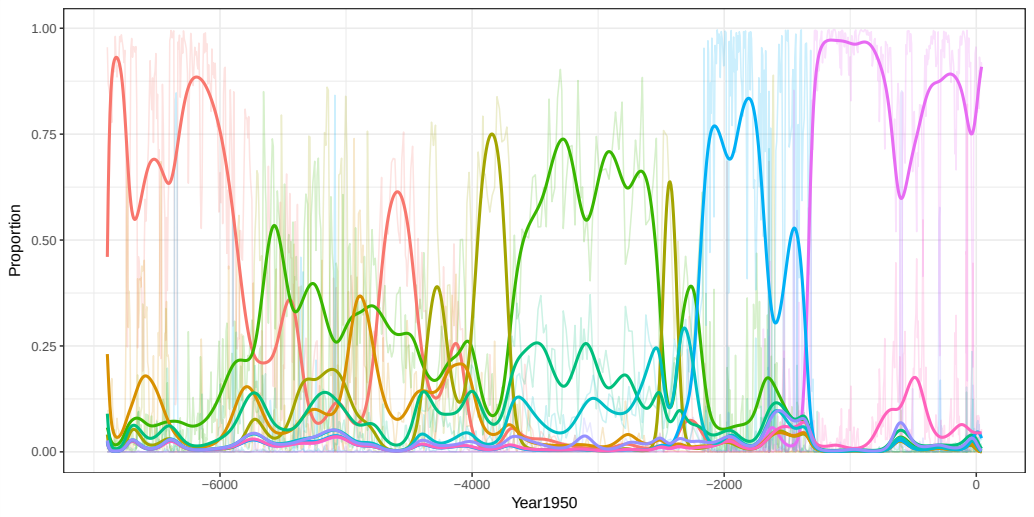
Dirichlet is for **true** proportions

Dirichlet regression

Model for the proportions α (or other parametrisation)



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*, &
- ...

Related topics

- *Structural topic model* — include covariates in the topic model
  - proportions of each species in the data set (corpus)
  - proportions of each association in samples
  - allow the proportions to vary over time
  - *dynamic topic model* is a special case
- Relationships with ecological theory
  - Harris *et al* (2015). Linking Statistical and Ecological Theory: Hubbell's Unified Neutral Theory of Biodiversity as a Hierarchical Dirichlet Process. *Proc. IEEE PP*, 1–14. `https://doi.org/10.1109/JPROC.2015.2428213`
  - In large samples Hubbell's UNTB $\longrightarrow$ HDP

Species interactions?

Species interactions

Community ecology is *the study of the interactions that determine the distribution and abundance of organisms* (Krebs, 2009)

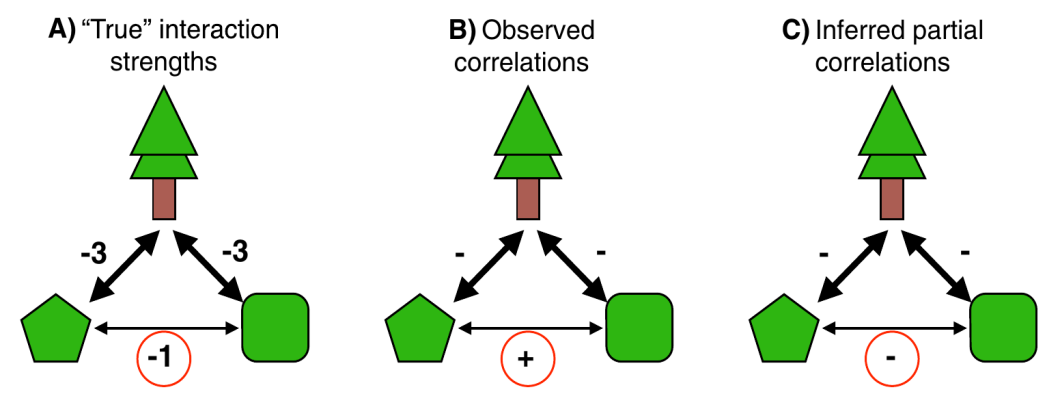
Yet change in community **composition** in space or time is often the primary focus of research — especially in palaeoecology

Network representation

Networks are a convenient way to represent community composition data

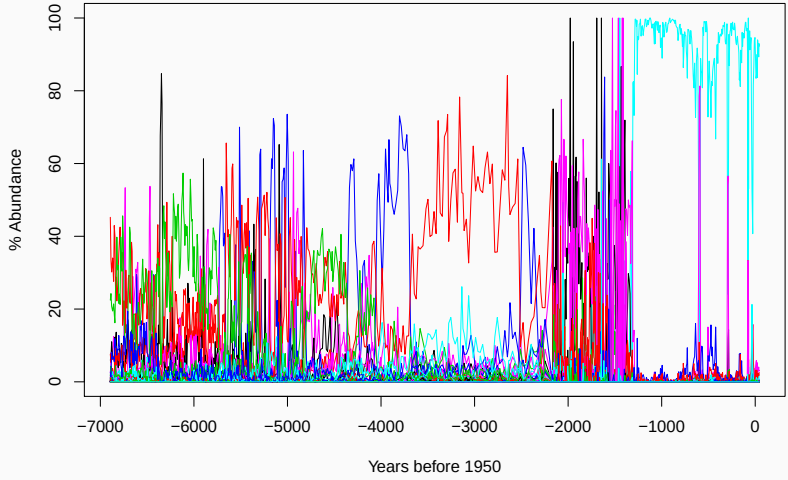
{Photo by Clint Adair on Unsplash}

Species interactions



Source: Harris (2016) *Ecology* **97**(12), 3308–3314

Foy Lake — Complex multivariate species data



Data provided by Jeffery Stone (Indiana State University); Spanbauer *et al* PLOS One 2014

Methods

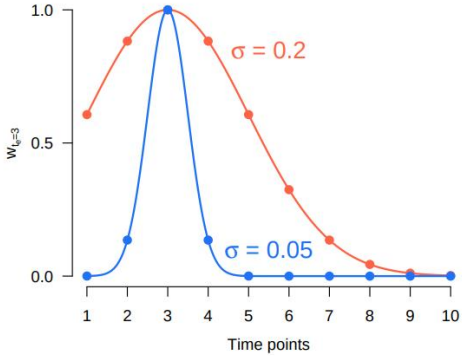
Flavours of undirected graphs — are species conditionally dependent given relationships with all other variables?

- Mixed Graphical Model
  - Gaussian responses, elastic net penalties, 1^{st}-order interactions — partial correlations
  - Poisson nodes (responses) allowed
  - Time-varying version works with irregularly spaced data
  - **mgm** R package Halsbeck & Waldorp, ArXiv
- Markov Random Fields
  - Ising-like model of binary responses
  - Conditional random fields approximated by pair-wise logistic regressions
  - Gaussian & Poisson nodes allowed
  - **MRFcov** R package Clark *et al* (2018) *Ecology*
  - Covariate (t) included as a node in the network

Time-varying MGM — Bandwidth

Assume the networks vary smoothly over time

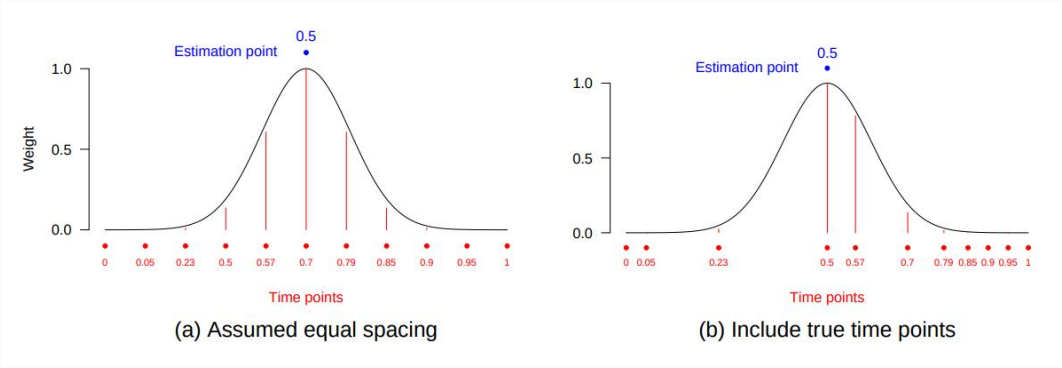
Bandwidth — σ — controls how much we smooth



Time	$X_{t,1}$	$X_{t-1,1}$	...	$X_{t,p}$	$w^{t_e=3}$	$w^{t_e^*=3}$
1	0.03	−0.97	...	−0.08	0.61	0.00
2	1.15	−1.07	...	−0.56	0.88	0.14
3	0.11	0.63	...	1.09	1.00	1.00
4	−1.08	0.13	...	1.88	0.88	0.14
5	−0.93	1.00	...	−0.29	0.61	0.00
6	−1.08	0.17	...	−1.36	0.32	0.00
7	0.27	−1.72	...	−1.13	0.14	0.00
8	0.03	−1.26	...	−0.97	0.04	0.00
9	−1.29	−1.05	...	−0.10	0.01	0.00
10	−0.07	−0.04	1.05	−0.12	0.00	0.00

Time-varying MGM

Can handle irregular spacing in time



Kassjön

Kassjön

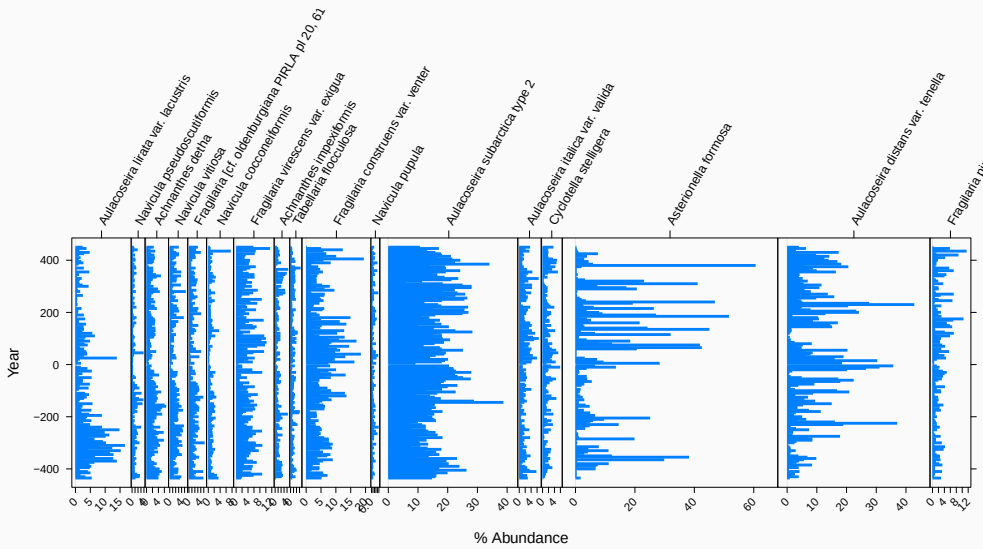
Source: © John Anderson

Kassjön

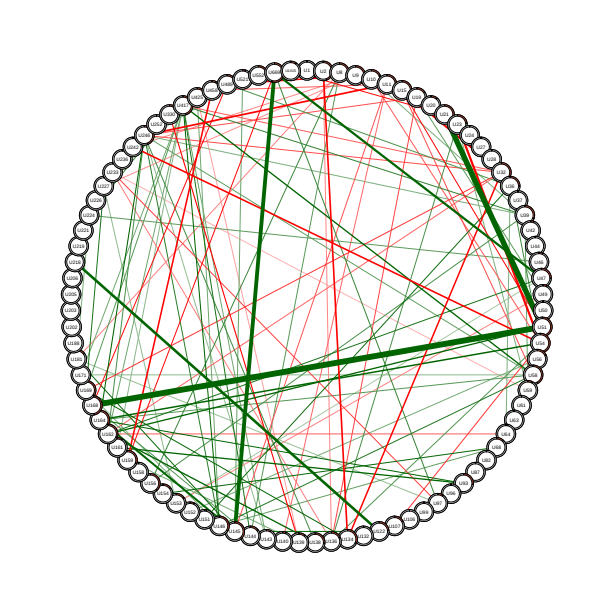
- Kassjön, Northern Sweden
- Small, dimictic, mesotrophic lake
- Seasonally anoxic; winter and summer
- Forms annual laminations due to strong spring minerogenic input (Spring melt)

Source: © John Anderson

Kassjön

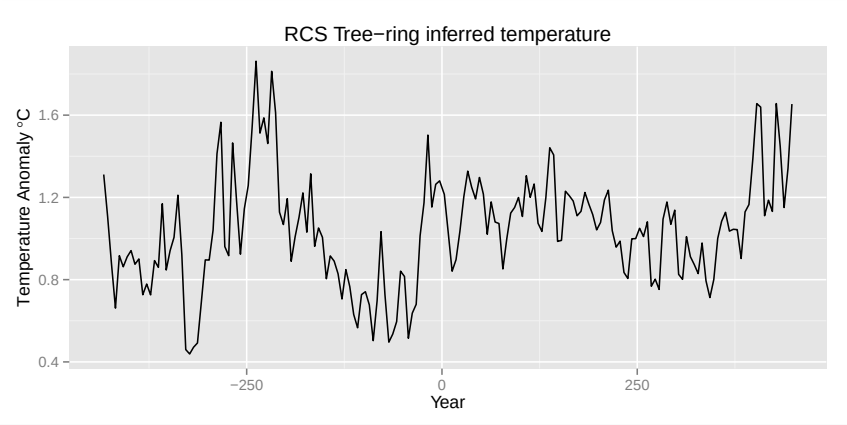

Aulacoseira lirata var. lacustris
Navicula pseudoscutiformis
Achnanthes deltha
Navicula viridula
Fragilaria cf. oldenburgiana PIRLA pl 20, 61
Navicula cocconeiformis
Fragilaria virescens var. exigua
Achnanthes impexiformis
Tabellaria flocculosa
Fragilaria construens var. venter
Navicula pupula
Aulacoseira subarctica type 2
Aulacoseira italica var. valida
Cyclotella stelligera
Asterionella formosa
Aulacoseira distans var. tenella
Fragilaria pinnata
Year
% Abundance

Kassjön interactions



Kassjön — do interactions change over time?

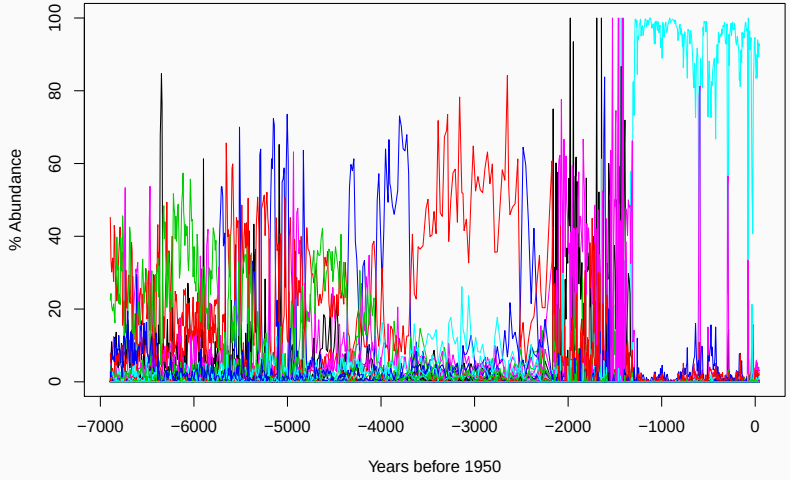
600 years of severe climate — particularly cold conditions ~330BC with glacial expansion at this time in Scandinavia



Kassjön — do interactions change over time?

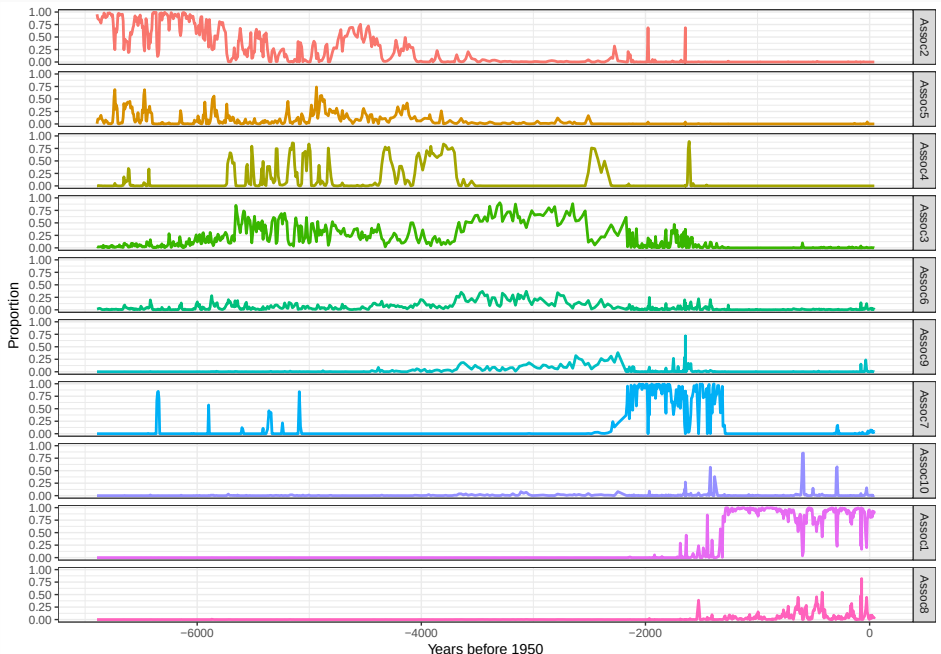
{Photo by Gemma Evans on Unsplash}

Foy Lake — Complex multivariate species data

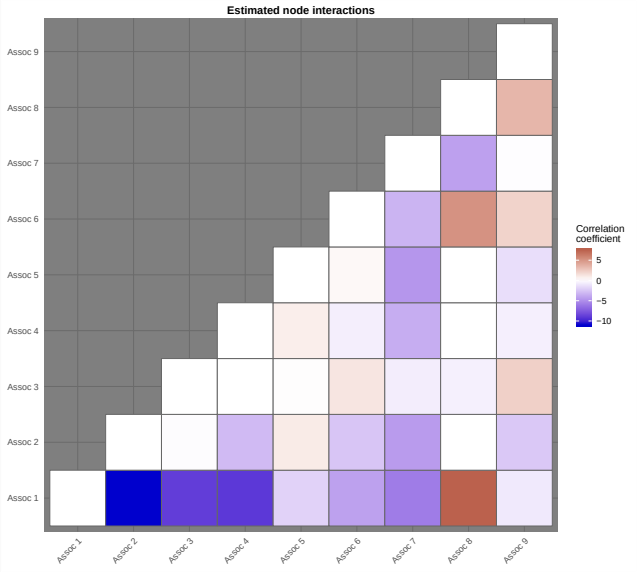


Data provided by Jeffery Stone (Indiana State University); Spanbauer *et al* PLOS One 2014

Correlated Topic Model

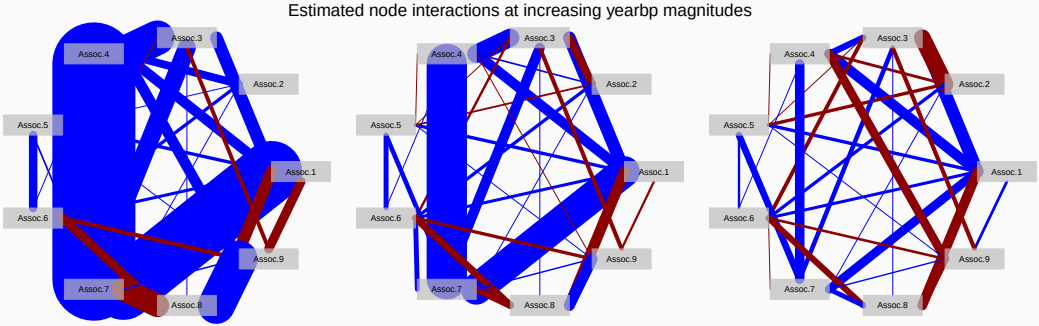


Foy Lake interactions



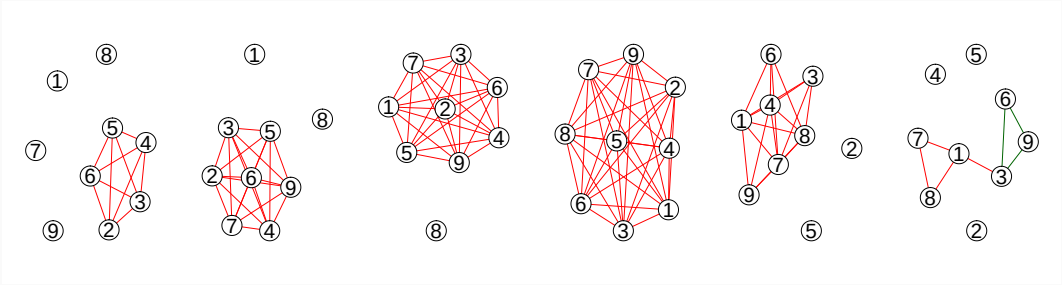
Foy Lake interactions — effect of time

How interaction coefficients change due to time



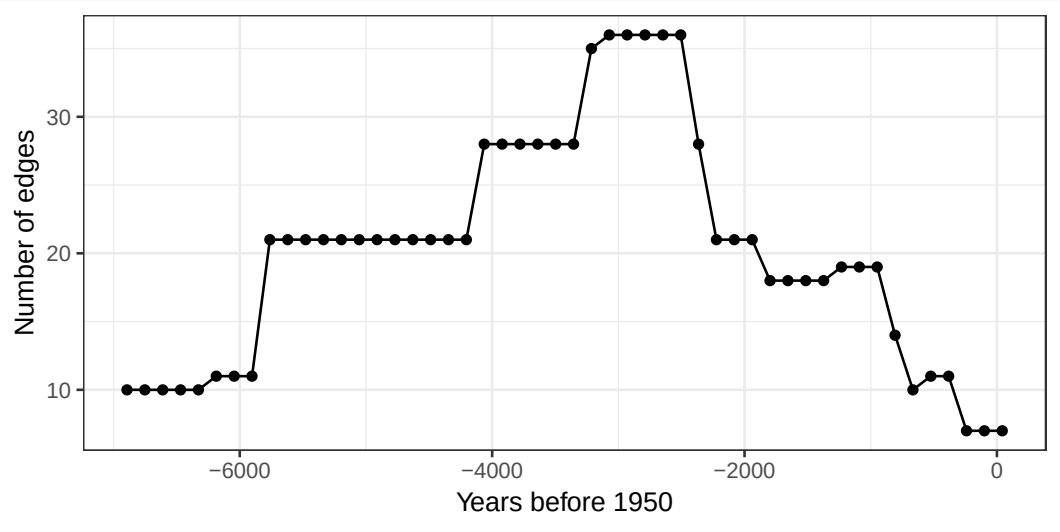
Foy Lake interactions — effect of time

How predicted interactions changes over time



Foy Lake interactions — effect of time

How network structure changes over time



What are these co-occurrence pattern-derived networks?

Attempting to divine process from pattern?

Range of methods failed to recover known interactions (Barner *et al*[1] & Freilich *et al*[2])

> *species interaction networks are a prediction of what [interactions] could be (Delmas et al*[3]*)*

Considerable power issue; estimating a huge number of parameters & assuming the result is sparse

[1] Ecology, 2018

[2] Ecology, 2018

[3] Biological Reviews, 2018

Conclusions

MGMs & (conditional) MRFs are a potentially useful way of viewing (palaeo)ecological data

With species-rich data sets, models are data hungry

- power problems; assumptions about sparsity

Some dimension reduction may help — complicates the interpretation

Work in progress — sensitivity to model settings & uncertainties?

{Photo by Steve Harvey on Unsplash}