Detecting and modeling changes in a time series of proportions: An application to phytoplankton taxa in a freshwater lake

Thomas J. Fisher Jing Zhang Stephen Colegate Michael J. Vanni

Miami University, Oxford, OH

17 August 2017

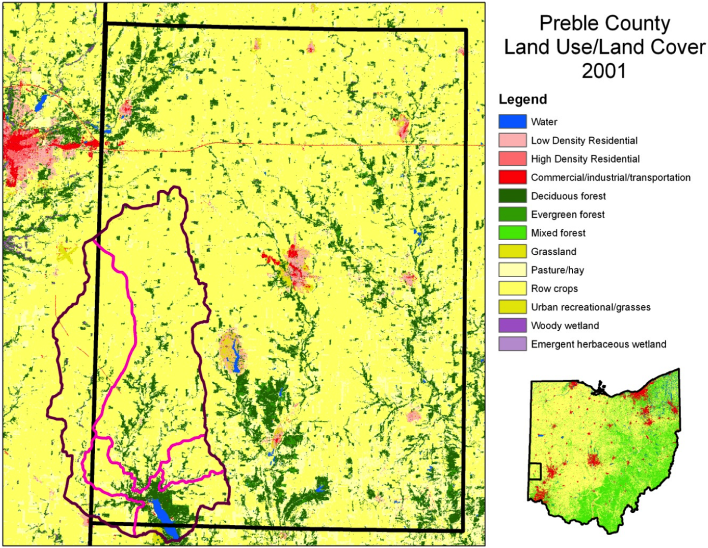
MIAMI UNIVERSITY
OXFORD, OH • EST. 1809

Acton Lake

Acton Lake – Hueston Woods State Park

Acton Lake

Acton Lake Watershed



Acton Lake Sediment Bloom

Agricultural Practices

Changes in Agricultural Practices

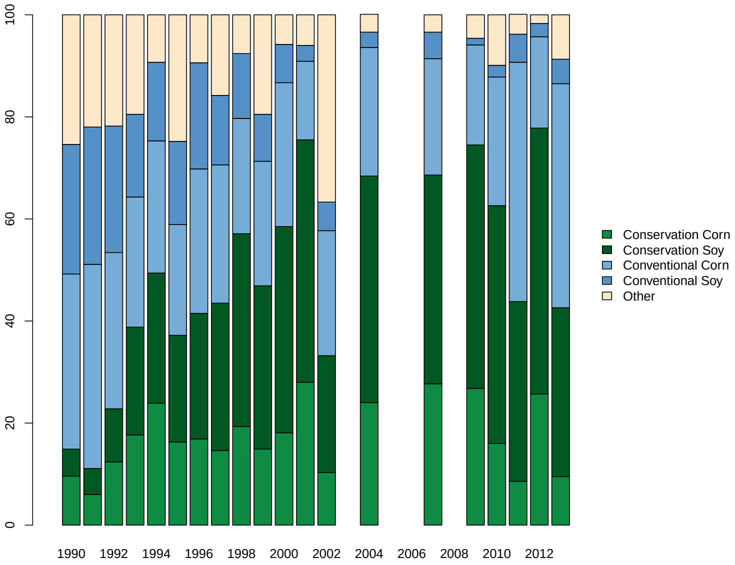
Less of this

More of this

Less of this

More of this

Farming Practices



Acton Lake Monitoring

Water Quality Monitoring and Analysis

Measurements

Since 1994 the following concentrations have been monitored:
Ammonium (NH_4), *Nitrate* (NO_3),
Phosphorus (SRP), and *Suspended Sediment* (SS).

with a known influence: Flow rate/discharge, in three streams:
Four Mile Creek,
Little Four Mile Creek, and
Marshall's Branch.

Addressed in Renwick et al. [2017].

Water Quality Conclusions

- *Ammonium* - Overall has decreased with two regimes: 1993 until 2004-ish levels decreased. Since 2004, much more variable.
- *Nitrate* - Overall decreased with two regimes: 1993 until 2006-ish levels decreased, reasonable level since.
- *Phosphorus* - No real overall change.
- *Suspended Sediment* - Overall decreased although the rate of decrease appears to be leveling off.

Water Quality Conclusions

- *Ammonium* - Overall has decreased with two regimes: 1993 until 2004-ish levels decreased. Since 2004, much more variable.
- *Nitrate* - Overall decreased with two regimes: 1993 until 2006-ish levels decreased, reasonable level since.
- *Phosphorus* - No real overall change.
- *Suspended Sediment* - Overall decreased although the rate of decrease appears to be leveling off.

So...

- Water clarity is improving (less sediment)
- Less nitrogen is entering the lake
- Phosphorus levels appear to be stationary

Water Quality Conclusions

- *Ammonium* - Overall has decreased with two regimes: 1993 until 2004-ish levels decreased. Since 2004, much more variable.
- *Nitrate* - Overall decreased with two regimes: 1993 until 2006-ish levels decreased, reasonable level since.
- *Phosphorus* - No real overall change.
- *Suspended Sediment* - Overall decreased although the rate of decrease appears to be leveling off.

So...

- Water clarity is improving (less sediment)
- Less nitrogen is entering the lake
- Phosphorus levels appear to be stationary

Questions from Ecology Friends

How does this effect the ecosystem?

Water Quality Conclusions

- *Ammonium* - Overall has decreased with two regimes: 1993 until 2004-ish levels decreased. Since 2004, much more variable.
- *Nitrate* - Overall decreased with two regimes: 1993 until 2006-ish levels decreased, reasonable level since.
- *Phosphorus* - No real overall change.
- *Suspended Sediment* - Overall decreased although the rate of decrease appears to be leveling off.

So...

- Water clarity is improving (less sediment)
- Less nitrogen is entering the lake
- Phosphorus levels appear to be stationary

Questions from Ecology Friends

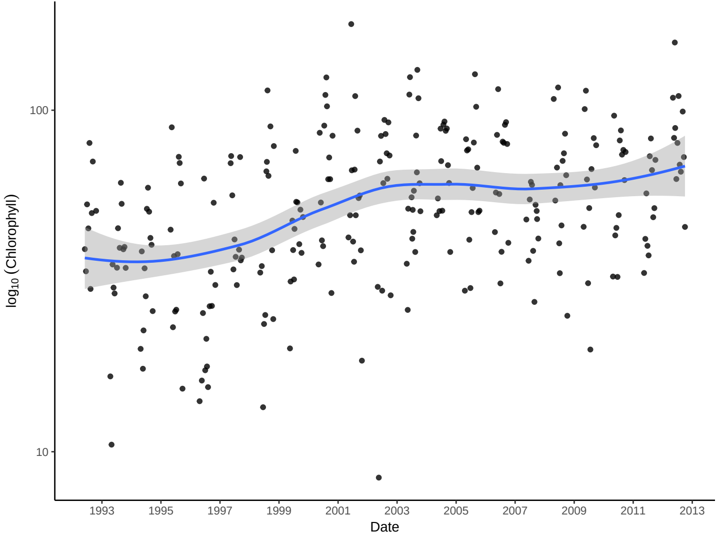
How does this effect the ecosystem?

- How has phytoplankton biomass changed?
- Are proportions of species types changing in time?

Phytoplankton

Analysis of Phytoplankton

Chlorophyll Measurements



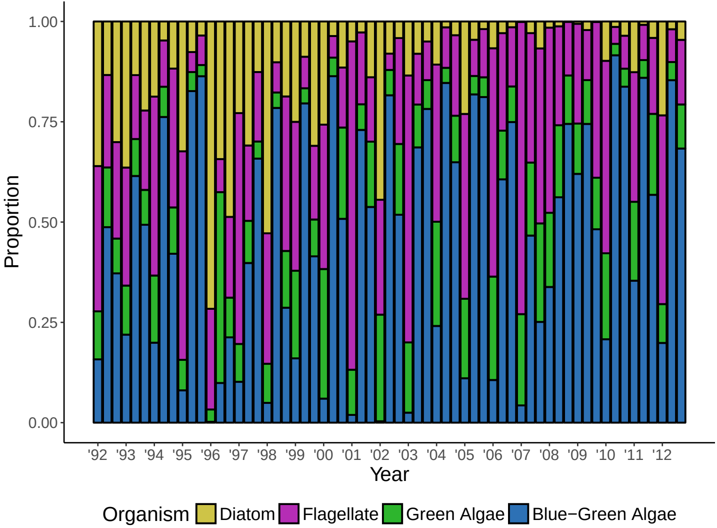
Data nuances

- Irregularly timed data
- Between 12 & 13 measurements per year, on average
- Recorded from May through September
- Most measurements in June, July & August (bi-weekly)
- Lake freezes over in winter – cannot collected
- Difficult to collect samples during heavy mixing periods (early spring, late fall)

Data nuances

- Irregularly timed data
- Between 12 & 13 measurements per year, on average
- Recorded from May through September
- Most measurements in June, July & August (bi-weekly)
- Lake freezes over in winter – cannot collected
- Difficult to collect samples during heavy mixing periods (early spring, late fall)
- We aggregate into three windows (other aggregation considered by not discussed today)
    - representing *late spring*, *summer* and *early fall*
    - Calculate the proportion of four taxa of phytoplankton: *Diatoms*, *Flagellate*, *Green algae* and *Blue-Green algae* (cyanobacteria)

Proportions in time



Time Series of Proportion

The time series of interest:

- Multivariate response on the Simplex of dimension $D = 4$ (*i.e.*, *compositional data*).
- Likely has seasonal influences
- Possible covariate influence (not explored today)

How to handle a time series of proportions:

- Traditional approach: log-ratio transformations and treated as *Normal* vector response; see Aitchison [1986].
- State space approach of Grunwald et al. [1993].
- New paper I have not read yet: Zheng and Chen [2017].

Time Series of Proportion

The time series of interest:

- Multivariate response on the Simplex of dimension $D = 4$ (*i.e., compositional data*).
- Likely has seasonal influences
- Possible covariate influence (not explored today)

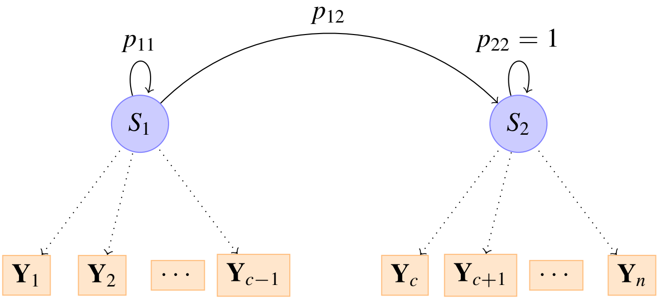
How to handle a time series of proportions:

- Traditional approach: log-ratio transformations and treated as *Normal* vector response; see Aitchison [1986].
- State space approach of Grunwald et al. [1993].
- New paper I have not read yet: Zheng and Chen [2017].

Our approach:

- Hidden Markov Model (HMM) with Dirichlet response where the HMM controls the parameters of a generalized linear model.

Hidden Markov Model



Each $\mathbf{Y}_i \sim Dirichlet_D(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha}' = (\alpha_1, \alpha_2, \ldots, \alpha_D)$.

To allow for covariates consider: $\alpha_j = \exp\left\{\mathbf{X}\boldsymbol{\beta}_j\right\}$ where $\mathbf{X}$ is a design matrix with coefficients $\boldsymbol{\beta}_j$.

Bayesian Estimation

We fit the HMM on Dirichlet response in the Bayesian framework. Specifically:

- The HMM is fit following Lystig and Hughes [2002]
- We consider at most one change in distribution, thus the transition matrix is limited to

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} \\ 0 & 1 \end{bmatrix}$$

- α_j are modeled by $\alpha_j = \exp\left\{\mathbf{X}\boldsymbol{\beta}_j\right\}$
- Consider two approaches for $\boldsymbol{\beta}_j$ parameters:

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \cdots \\ \boldsymbol{\beta}_D \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{D1} & \beta_{D2} & \cdots & \beta_{Dm} \end{bmatrix}$$

Two Model Approaches

Independent Components

- Prior on all β_{ij} terms are independent $N(0,2)$
- This corresponds to components within a response vector are treated as independent entities

Correlated Components

- Each column from **B** is treated as a mean zero multivariate Normal
- Assume compound symmetry covariate structure, use LKJ prior

Design matrix (for today)

$$\mathbf{X}_{1:3} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

and the prior on the transition probability p_{11} is $Beta(4,1)$.

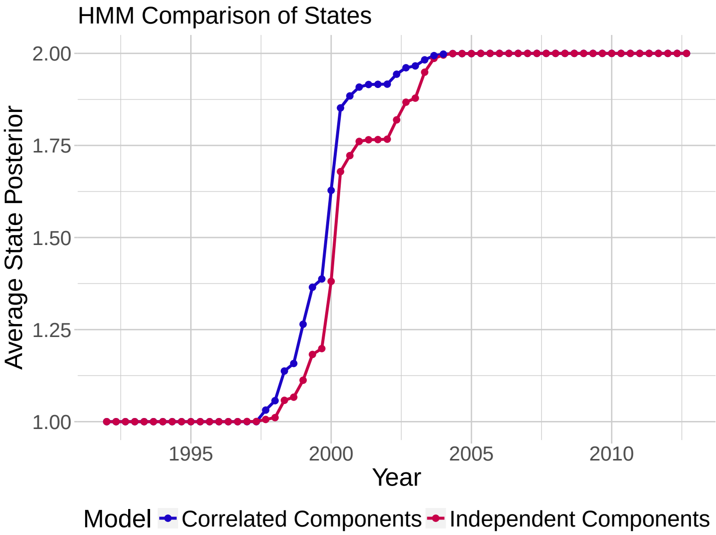
`rstan` details

Computational Details

- No-U-Turn sampler (NUTS)
- 2-chains
- 50,000 warm up samples
- 50,000 post-warm up samples
- thinning every 50 samples

Takes about 20 minutes to fit one model.

Change in States

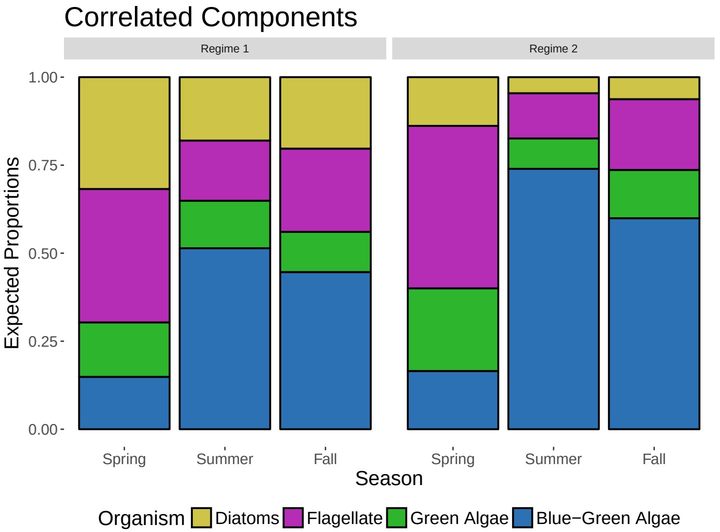


Correlated Component Details

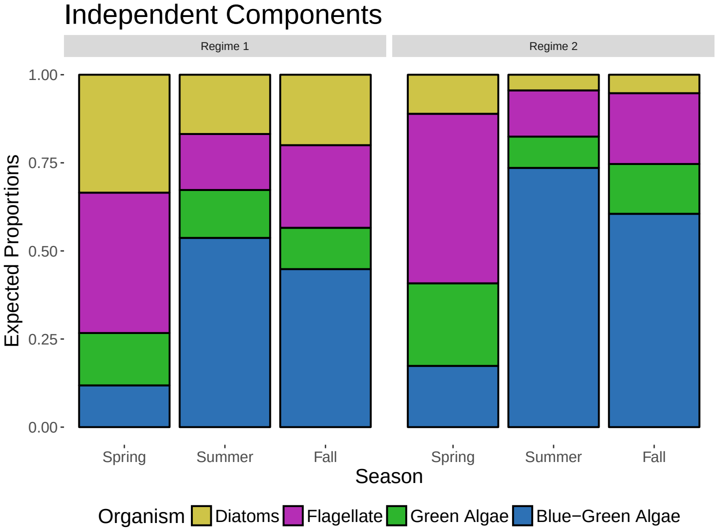
Table: Median from posterior distribution with 90% credible interval for α-parameters determining the shape of the Dirichlet distribution

		State 1	State 2
$\alpha_{Diatoms}$	Spring	2.109 **(1.125, 4.068)**	0.933 **(0.649, 1.111)**
	Summer	1.375 (0.842, 2.148)	0.962 (0.673, 1.261)
	Fall	1.837 (1.040, 3.083)	0.894 (0.588, 1.106)
$\alpha_{Flagellate}$	Spring	2.516 (1.275, 4.486)	3.133 (2.080, 4.684)
	Summer	1.309 (0.780, 2.024)	2.714 (1.791, 3.911)
	Fall	2.141 (1.213, 3.444)	2.875 (1.914, 4.168)
α_{Green}	Spring	1.027 (0.752, 1.452)	1.584 (1.137, 2.255)
	Summer	1.031 (0.701, 1.524)	1.818 (1.216, 2.581)
	Fall	1.033 (0.744, 1.605)	1.960 (1.292, 2.951)
$\alpha_{Blue-green}$	Spring	0.988 (0.557, 1.583)	1.115 (0.716, 1.794)
	Summer	3.925 **(2.080, 6.550)**	15.615 **(10.466, 22.008)**
	Fall	4.040 (2.124, 6.609)	8.568 (5.502, 12.614)

Correlated Component Summary



Independent Component Summary



Contextual findings

Overall phytoplankton

- Change point in chlorophyll measurements circa 2000
- Overall levels of chlorophyll (hence algae biomass) has increased

Taxa of phytoplankton

- Change point occurs at roughly the same time, definite by 2003
- Proportion of Flagellate and Green algae has undergone some minor changes
- Large increase in the proportion of cyanobacteria
- Substantial decrease in proportion of Diatoms

Future work

- Include covariate influence, try and determine some sort of *causal* (or at least suggestive) type effect
- From a biological perspective, why the increase in algae (think we have an answer) but why the changing dynamics in types of algae (do not have an answer)

Thanks!

Collaborators & contributers

- Dr. Jing Zhang - Colleague & Bayes person
  Department of Statistics - Miami University
- Mr. Stephen Colegate - Former MS Student
  Department of Mathematics - Xavier University
- Dr. Mike Vanni - Ecologist (Algae guy)
  Department of Biology - Miami University
- Dr. Bill Renwick - Geographer (Soil Guy)
  Department of Geography - Miami University
- Ms. Emily Morris - Former undergraduate Student
  University of Michigan-Biostats PhD student

Questions? Comments? Suggestions?

References

J. Aitchison. *The statistical analysis of compositional data*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986. ISBN 0-412-28060-4. doi: 10.1007/978-94-009-4109-0. URL `http://dx.doi.org/10.1007/978-94-009-4109-0`.

Gary K. Grunwald, Adrian E. Raftery, and Peter Guttorp. Time series of continuous proportions. *Journal of the Royal Statistical Society, Series B*, 55:103–116, 1993. URL `http://www.jstor.org/stable/2346067`.

Theodore C. Lystig and James P. Hughes. Exact computation of the observed information matrix for hidden markov models. *Journal of Computational and Graphical Statistics*, 11(3):678–689, 2002. ISSN 10618600. URL `http://www.jstor.org/stable/1391119`.

William H. Renwick, Michael J. Vanni, Thomas J. Fisher, and Emily L. Morris. Water quality trends and agricultural practices in a midwest u.s. watershed over a 21-year period. *pre-print*, 2017.

Tingguo Zheng and Rong Chen. Dirichlet arma models for compositional time series. *J. Multivar. Anal.*, 158 (C):31–46, June 2017. ISSN 0047-259X. doi: 10.1016/j.jmva.2017.03.006. URL `https://doi.org/10.1016/j.jmva.2017.03.006`.