Detecting and Modeling Changes in a Time Series of Continuous Proportions

An Application to Phytoplankton Taxa in a Freshwater Lake

Thomas J. Fisher

Joint work with Jing Zhang Stephen Colegate Michael J. Vanni

Miami University, Oxford, OH

18 July 2023

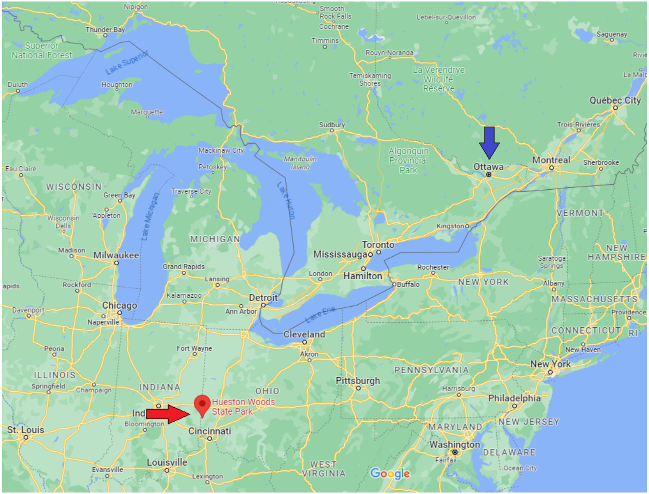
MIAMI UNIVERSITY
OXFORD, OH • EST. 1809

Some Background

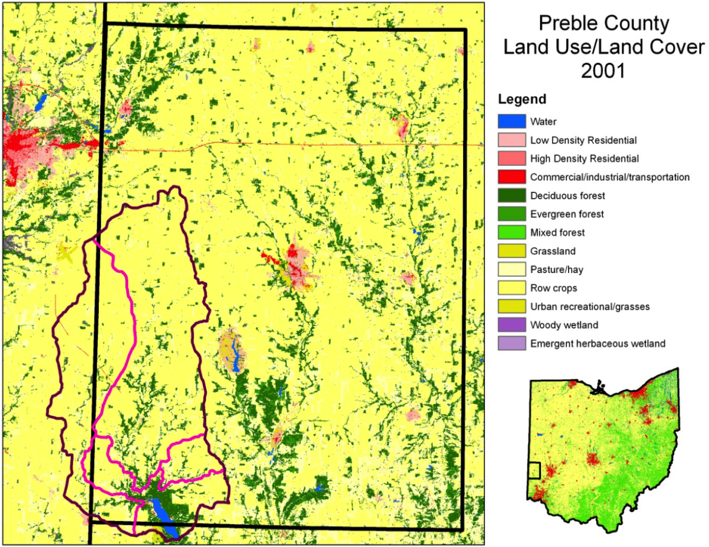
Acton Lake – Hueston Woods State Park

Acton Lake

Where is this?



Acton Lake Watershed



Acton Lake Sediment Bloom

Agricultural Practices

Changes in Agricultural Practices
over past 30 years

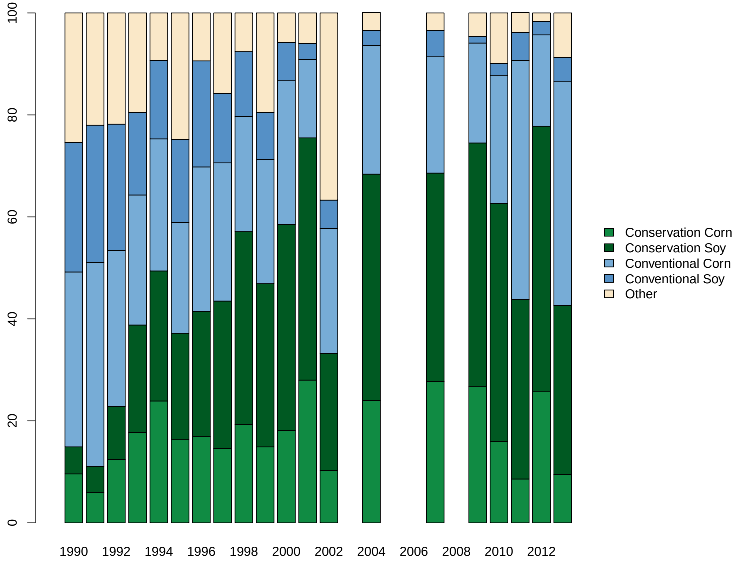
Less of this

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.

Trends analyzed in Renwick et al. [2018].

Water Quality Conclusions

Overall findings

- *Ammonium* - Overall has decreased with roughly 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 flat 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

Overall findings

- *Ammonium* - Overall has decreased with roughly 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 flat 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

Overall findings

- *Ammonium* - Overall has decreased with roughly 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 flat 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.

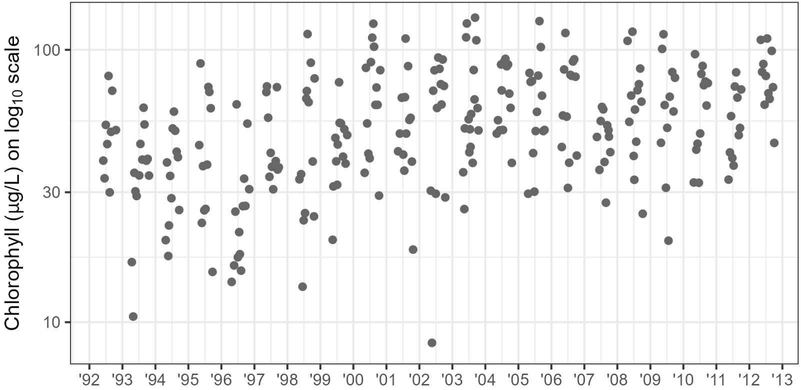
Questions from Ecology Friends – How does this effect the ecosystem?

- How has phytoplankton biomass changed?
- Is the composition of algal species types changing in time?

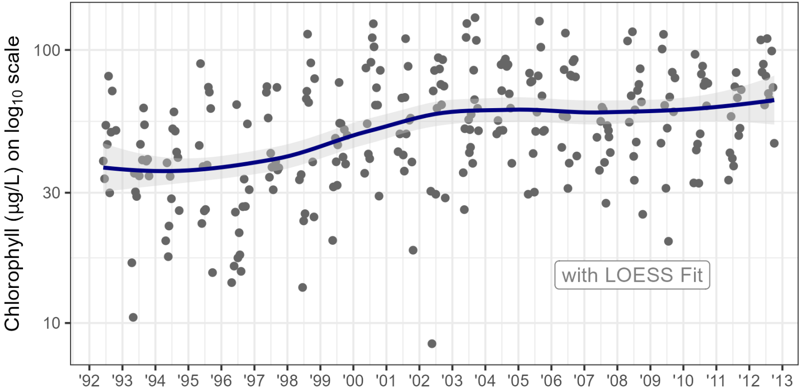
Phytoplankton

Analysis of Phytoplankton Biomass

Chlorophyll Measurements



Chlorophyll Trends?



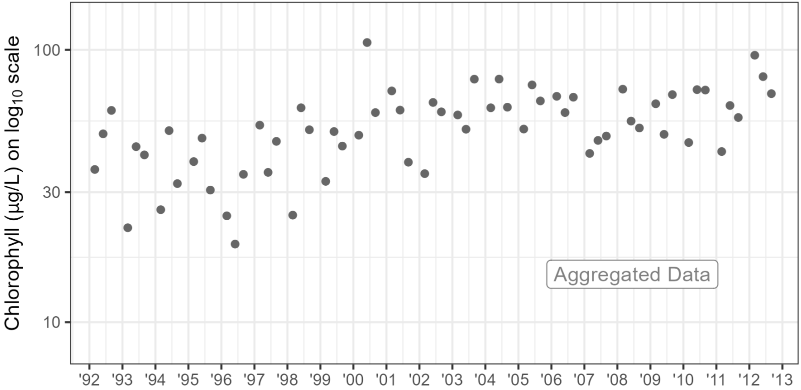
Data nuances

- Irregularly timed data.
- Roughly 12 or 13 measurements per year, on average.
- Recorded from May through September.
- Most measurements in June, July & August (bi-weekly).
- Lake can freeze in winter – Marina closed, lake access restricted.
- Difficult to collect samples during heavy mixing periods (early spring, late fall).

Data nuances

- Irregularly timed data.
- Roughly 12 or 13 measurements per year, on average.
- Recorded from May through September.
- Most measurements in June, July & August (bi-weekly).
- Lake can freeze in winter – Marina closed, lake access restricted.
- 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* mixing, *summer* stratification and *early fall* mixing.

Aggregated Chlorophyll Measurements



Change Point Analysis

Many methods are available for univariate time series.

We apply the mean shift change point test from Robbins et al. [2011].

Change Point Analysis

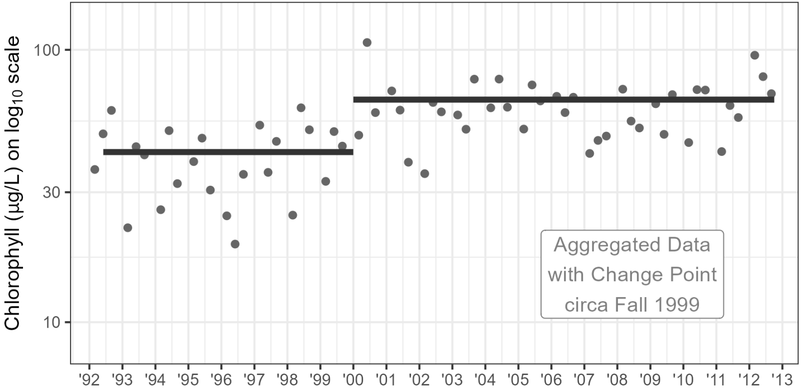
Many methods are available for univariate time series.

We apply the mean shift change point test from Robbins et al. [2011].

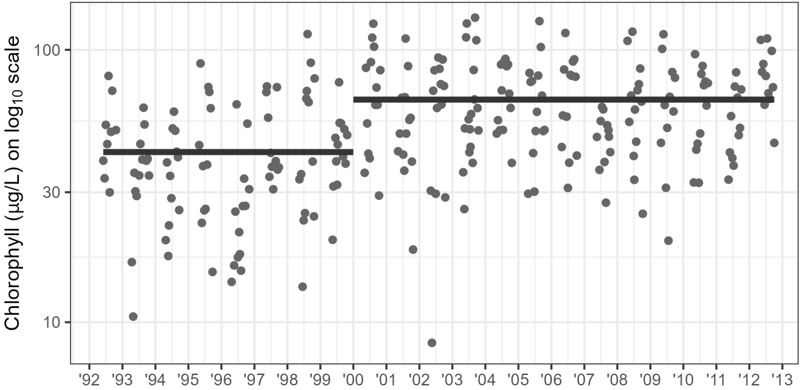
Stat	Location	p-value
1.6507	24	0.0086

Time point 24 corresponds to Fall 1999.

Aggregated Chlorophyll Measurements with Change Point



Chlorophyll Measurements with Change Point



Phytoplankton

That was easy...

What about the composition of algal species?

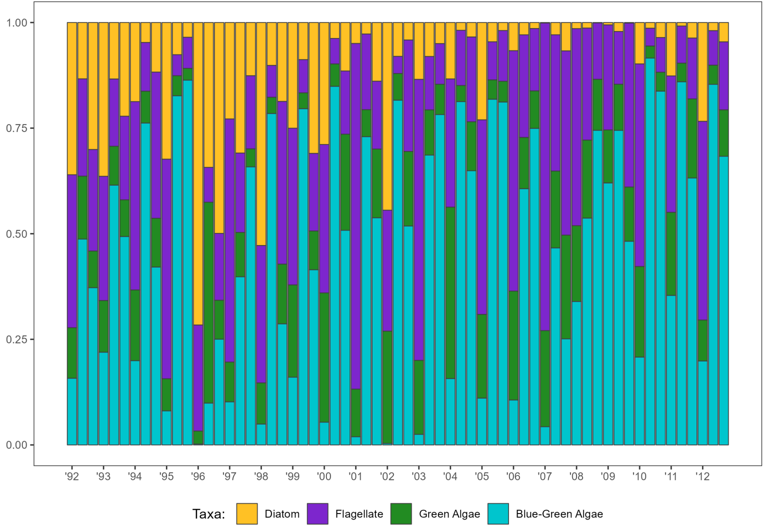
Switching to Proportions

The total biomass problem is fairly *easy* (well studied).

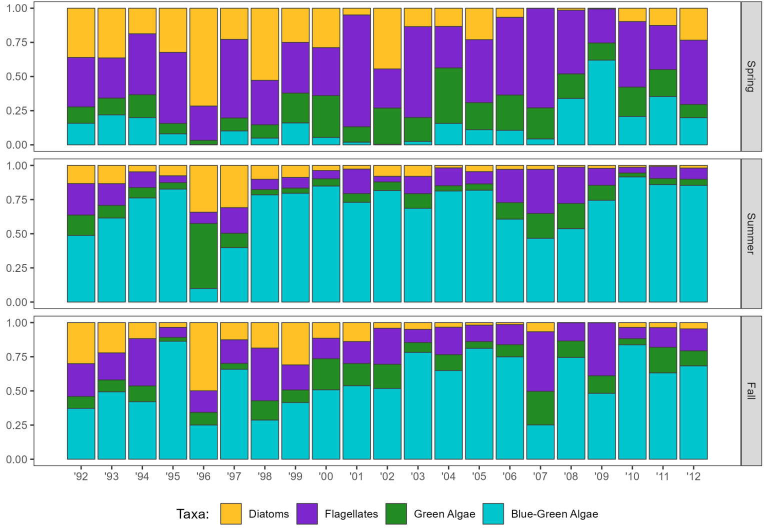
To tackle the question about the composition of algal species types:

- Calculate the proportion of four taxa of phytoplankton:
  - *Diatoms*.
  - *Flagellate*.
  - *Green algae*.
  - *Blue-Green algae* (cyanobacteria).
- Each measured when water is sampled (12-13 times per year).
- Aggregated into three measurements per year
  - *Spring* mixing, *Summer* stratification, *Fall* mixing.

Proportions in time



Proportions stratified by season



Time Series of Proportion

The time series of interest:

- Multivariate response in the Simplex of dimension $D = 4$ (*i.e., compositional data*).
- Clearly seasonal.
- Possible covariate influence (not explored today, see paper).

How to handle a time series of proportions:

- *Classic* approach: log-ratio transformations and treated as *Normal* vector response; see Aitchison [1986].
- State space approach of Grunwald et al. [1993].
- Dirichlet Regression (multivariate GLM) [Hijazi and Jernigan, 2009].
- Dirichlet ARMA Models [Zheng and Chen, 2017].
- Permutation based change point detection for single parameter Dirichlet [Prabuchandran et al., 2021].

Our Approach

Our approach [Fisher et al., 2022]:

- Hidden Markov Model (HMM) with Dirichlet response and predictor variables.
- the HMM controls the parameters of a system of generalized linear models.

Dirichlet Distribution

Consider $\mathbf{Y}_i \sim \text{Dirichlet}_D(\boldsymbol{\alpha})$

where $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_D)$ with $\alpha_i > 0$, known as the shape parameters.

A generalization of the Beta distribution.

The expectation and variance of Y_j,
the j^{th} component of $\mathbf{Y}$, is

$$E[Y_j|\boldsymbol{\alpha}] = \alpha_j / \boldsymbol{\alpha}'\mathbf{1}_D \quad \text{and} \quad Var[Y_j|\boldsymbol{\alpha}] = \frac{\alpha_j(\boldsymbol{\alpha}'\mathbf{1}_D - \alpha_j)}{(\boldsymbol{\alpha}'\mathbf{1}_D)^2(\boldsymbol{\alpha}'\mathbf{1}_D + 1)}$$

where $\mathbf{1}_D$ is a D-dimensional vector of ones.

Reparameterized Dirichlet

Consider reparameterizing the shape parameter as such [Grunwald et al., 1993]

$$\boldsymbol{\theta} = \boldsymbol{\alpha}/\tau \quad \text{where} \quad \tau = \boldsymbol{\alpha}'\mathbf{1}_D$$

thus $\mathbf{Y} \sim \text{Dirichlet}_D(\boldsymbol{\alpha} = \tau\boldsymbol{\theta})$, with

$$E[\mathbf{Y}|\boldsymbol{\theta},\tau] = \boldsymbol{\theta} \quad \text{and} \quad Var[\mathbf{Y}|\boldsymbol{\theta},\tau] = \boldsymbol{\theta}\boldsymbol{\theta}'/(\tau+1).$$

- $\boldsymbol{\theta}$ is a *location* parameter in the simplex of dimension D, and
- τ is a *scale* parameter that inversely influences the variance.

Generalized Linear Models

The location parameter can be modeled by

$$\boldsymbol{\theta} = \boldsymbol{\eta}/(\boldsymbol{\eta}'\mathbf{1}_D), \quad \text{where } \log(\eta_i) = \beta_{i0} + \beta_{i1}X_1 + \beta_{i2}X_2 + \ldots + \beta_{ik}X_k, \tag{1}$$

and X_j, $j = 1, \ldots, k$, are predictor variables with β_{ij} as the coefficient on the j^{th} predictor for component i.

Model the scale parameter with

$$\log(\tau) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_k X_k, \tag{2}$$

where the γ_j terms are the coefficients on the j^{th} predictor.

Generalized Linear Models

The location parameter can be modeled by

$$\boldsymbol{\theta} = \boldsymbol{\eta}/(\boldsymbol{\eta}'\mathbf{1}_D), \quad \text{where } \log(\eta_i) = \beta_{i0} + \beta_{i1}X_1 + \beta_{i2}X_2 + \ldots + \beta_{ik}X_k, \tag{1}$$

and $X_j, j = 1, \ldots, k$, are predictor variables with β_{ij} as the coefficient on the j^{th} predictor for component i.

Model the scale parameter with

$$\log(\tau) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_k X_k, \tag{2}$$

where the γ_j terms are the coefficients on the j^{th} predictor.

This framework allows for a different set $\left(i.e., \{X_j\}_{j=1}^k\right)$ of predictor variables for the location and scale.

Hidden Markov Model

- Implement a HMM with the generalized Dirichlet formation from before.
  - the β_{ij} and γ_j terms are controlled by the HMM.
- This allows the HMM to detect changes in the underlying location and/or scale of the distribution.
- Constrain the transition matrix such that a Markov chain in state i can only jump to state $i+1$ or remain in state i at the next transition; i.e., $p_{ij} = 0$ for all $j \neq i, i+1$. [Chib, 1998].

Viterbi state assignments [Cappé et al., 2005] can be used to determine if a change point occurred—a change in Viterbi state indicates a shift in the observed distribution.

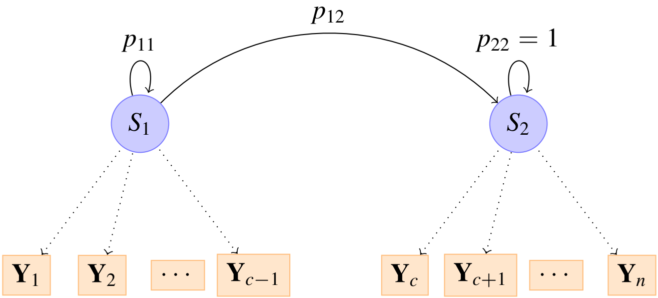
Hidden Markov Model

- Implement a HMM with the generalized Dirichlet formation from before.
  - the β_{ij} and γ_j terms are controlled by the HMM.
- This allows the HMM to detect changes in the underlying location and/or scale of the distribution.
- Constrain the transition matrix such that a Markov chain in state i can only jump to state $i+1$ or remain in state i at the next transition; i.e., $p_{ij} = 0$ for all $j \neq i, i+1$. [Chib, 1998].

Viterbi state assignments [Cappé et al., 2005] can be used to determine if a change point occurred—a change in Viterbi state indicates a shift in the observed distribution.

Allows us to address the ecological questions: did a considerable shift in phytoplankton phenology occur and what is the nature of that shift?

A Visual of a 2-State Hidden Markov Model



Each $\mathbf{Y}_i \sim \text{Dirichlet}_D\left(\boldsymbol{\alpha} = \tau\boldsymbol{\theta}\right)$ with $\boldsymbol{\theta}$ and τ modeled by equations (1) and (2), respectively.

Additional details, simulation studies, and variations of the model are available in Fisher et al. [2022].

Bayesian Estimation

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

Specifically:

- The HMM is fit following Lystig and Hughes [2002].
- No-U-Turn sampler (NUTS) in `rstan`, 2-chains, 50,00 warm up and 50,00 post-warm up samples with thinning every 50 samples.
- Priors:
  - $p_{ii} \sim \text{Beta}(9.5, 0.5)$ – Hesitant to jump states.
  - $\beta_{ij}, \gamma_j \sim N(0, 2)$ – centered at zero.

Design matrix (for today)

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

Some Findings

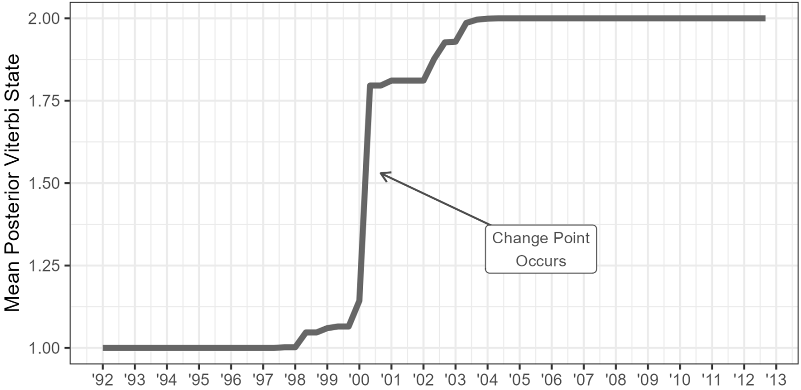
Results using this approach...

Model Selection

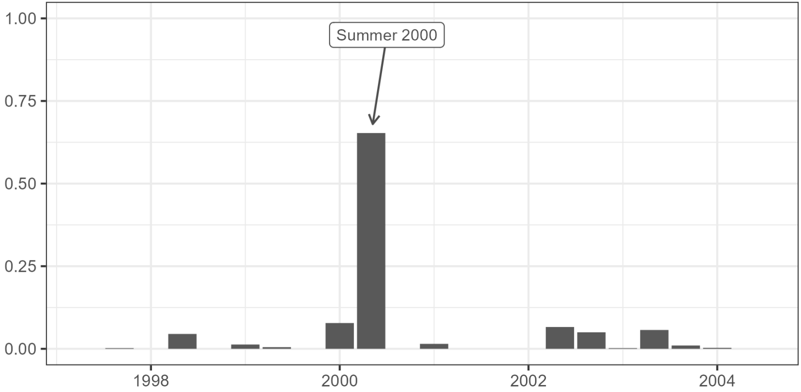
Using the Bayesian Leave-one-out cross validation based model selection [Vehtari et al., 2017]. (similar to a penalized model selection)

No-Change	One Change Point	Two Change Points
-371.00	$-\mathbf{379.25}$	-336.72

Change in States



Posterior Distribution of Change Point Locations

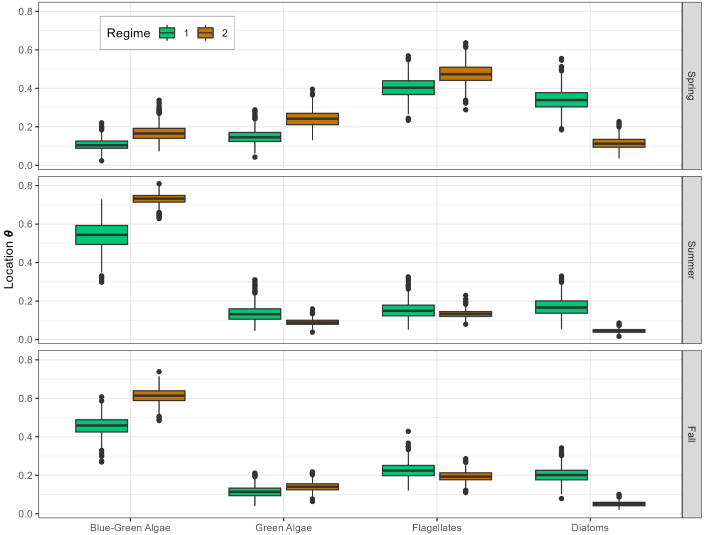


Change Point Occurred

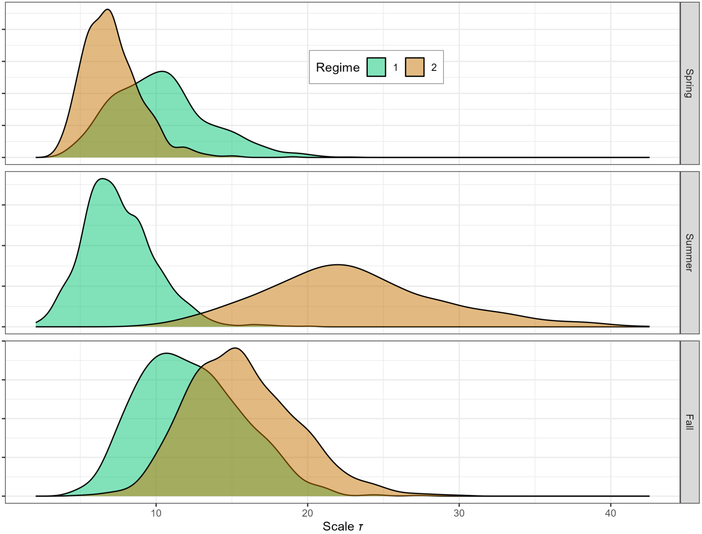
Pretty confident a change occurred.

What is the nature of that change?

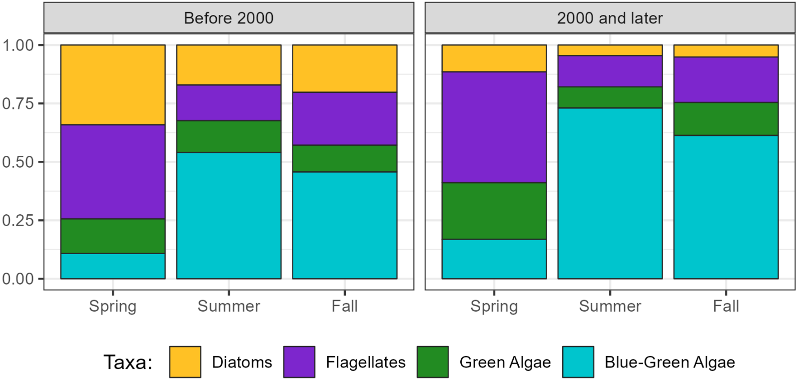
Posterior Distribution of θ



Posterior Distribution of τ



Expected Proportion of Phytoplankton



Contextual Findings

Overall phytoplankton biomass

- Change point in chlorophyll measurements circa 1999/2000.
- 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 minor changes.
- Large increase in the proportion of cyanobacteria.
- Substantial decrease in proportion of Diatoms.

Other work (not included today)

- Covariate influence (*e.g.*, water temperature, water clarity).
- Other aggregation (5 measurements per year) – same general result, change point is a little earlier.

Statistical Findings

The HMM can be a useful in change point analysis!
(feels like it is a forgotten tool in the toolbox)

Can simultaneously detect a change point and model the changes.
Fairly straightforward to add additional structure (*e.g.*, generalized linear models).

Has the added benefits

- State probabilities (similar to Viterbi states, not shown today).
  - Provides a measure of uncertainty on the state of each time point.
- With a Bayesian implementation:
  - Posterior distribution provides a measure of variability on the change point location.
  - Allows for the construction of credible intervals on the change point location.

Some computational costs is a drawback.

Thanks!

Collaborators & contributors

- Dr. Jing Zhang - Colleague & Bayes person
  Department of Statistics - Miami University
- Dr. Stephen Colegate - Former MS Student
  Cincinnati Children's Hospital Medical Center
- Dr. Mike Vanni - Ecologist (Algae guy)
  Department of Biology - Miami University
- Dr. Bill Renwick - Geographer (Soil Guy)
  Department of Geography - Miami University
- Dr. Emily Morris - Former undergraduate Student
  Food & Drug Administration

Questions? Comments? Suggestions?

References I

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.

Olivier Cappé, Eric Moulines, and Tobias Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387402640.

Siddhartha Chib. Estimation and comparison of multiple change-point models. *J. Econometrics*, 86(2):221–241, 1998. ISSN 0304-4076. doi: 10.1016/S0304-4076(97)00115-2.

Thomas J. Fisher, Jing Zhang, Stephen P. Colegate, and Michael J. Vanni. Detecting and modeling changes in a time series of proportions. *The Annals of Applied Statistics*, 16(1):477 – 494, 2022. doi: 10.1214/21-AOAS1509.

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.

Ragiq H. Hijazi and Robert W. Jernigan. Modeling compositional data using dirichlet regression models. *Journal of Applied Probability & Statistics*, 4(1):77–91, 2009.

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.

K.J. Prabuchandran, Nitin Singh, Pankaj Dayama, and Vinayaka Pandit. Change point detection for compositional multivariate data. *Applied Intelligence*, 2021. doi: 10.1007/s10489-021-02321-6.

References II

William H. Renwick, Michael J. Vanni, Thomas J. Fisher, and Emily L. Morris. Stream nitrogen, phosphorus, and sediment concentrations show contrasting long-term trends associated with agricultural change. *Journal of Environmental Quality*, 47(6):1513–1521, 2018. doi: 10.2134/jeq2018.04.0162.

Michael Robbins, Colin Gallagher, Robert Lund, and Alexander Aue. Mean shift testing in correlated data. *J. Time Series Anal.*, 32(5):498–511, 2011. ISSN 0143-9782. doi: 10.1111/j.1467-9892.2010.00707.x.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432, 2017. ISSN 1573-1375. doi: 10.1007/s11222-016-9696-4.

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.

Self-References

These slides are available on my github site:
`https://tjfisher19.github.io/`

Github handle: **`tjfisher19`**

Email: **`fishert4@miamioh.edu`**

Paper (with many more details):
Thomas J. Fisher, Jing Zhang, Stephen P. Colegate, and Michael J. Vanni. "Detecting and modeling changes in a time series of proportions." *The Annals of Applied Statistics*, 16 (1): 477 – 494, 2022. `10.1214/21-AOAS1509`.

`https://doi.org/10.1214/21-AOAS1509`

Code available:
`https://github.com/tjfisher19/hmmDirichletModel`