STAR Progress Review Workshop
Old Town Alexandria, VA
June 16-18, 2004

Bayesian Methods for Regional Eutrophication Models

E. Conrad Lamon III
Dept. of Environmental Studies
Louisiana State University
and
Craig A. Stow
Department of Environmental Health Sciences
University of South Carolina

Overview

- Goals and Objectives
- Approach
- Preliminary Findings
- Significance
- Next Steps

Goals and Objectives

- Use modern classification and regression trees and hierarchical Bayesian techniques to link multiple environmental stressors to biological responses and quantify uncertainty in model predictions and parameters.

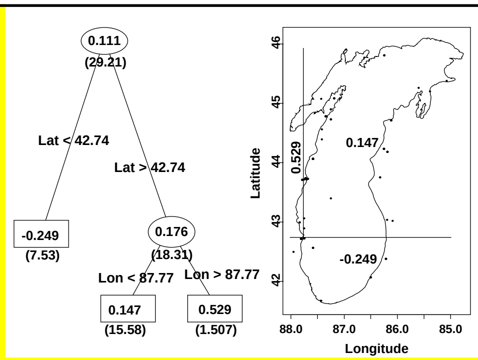
Guidance for TMDL model selection (NRC 2001)

- report prediction uncertainty
- be consistent with the amount of data available
- flexible enough to permit updates and improvements

Approach

Tree based methods

- are a flexible approach useful for variable subset selection
- when the analyst suspects global non-linearity
- and cannot (or does not want to) specify the functional form of possible interactions *a priori*.



Example tree

Methods

- Classification And Regression Trees (CART),
- it's Bayesian analogue, BCART
- a recently developed enhancement to the BCART procedure, which includes BCART as a model subclass, known as Bayesian Treed (BTREED) models, and
- Bayesian Hierarchical Models

BCART and BTREED Models

- Will be used with the EPA Nutrient Criteria Database to identify and estimate regional eutrophication stressor – response models for EPA STAR funded research.
- ✓ Lamon and Stow, 2004, Water Research, 38(11): 2764-2774.

Bayesian Treed models

- Bayesian Hierarchical model to:
  - Select subsets on $X \rightarrow X_s$
  - Fit linear models to these subsets X_s
- Tree structured models
  - "ANOVA in Reverse"
- "Leaves" contain linear models, not just a mean (like in CART models)

Bayesian Treed model specification

$y|x$, with $x = (x_1, x_2, \ldots, x_p)$,

where p = number of predictor variables.

two components of model

1. tree T with b bottom nodes,
2. parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_b)$,

where θ_i is associated with the ith bottom node. If x is in the ith node, then $y|x = f(y|\theta_i)$, where f is a parametric family indexed by θ_i.

Bayesian Treed model specification (cont.)

Tree is fully specified by (θ, T)

need a prior,

$$p(\theta, T).$$

Because θ indexes a parametric model for each T, we can use Bayes theorem such that

$$p(\theta, T) = p(\theta \mid T)p(T).$$

So, specify prior in two stages:

1 – on the tree space, $p(T)$, and

2 – on the distribution of Y at the bottom nodes, conditional on T, $p(\theta \mid T)$.

Bayesian Treed model search

- MCMC used to stochastically search for high posterior probability trees *T*.
- Metropolis –Hastings algorithm simulates a Markov chain with limiting distribution *p(T|Y,X)*
- Chipman, George and McColloch, 2000, JASA.

  http://gsbwww.uchicago.edu/fac/robert.mcculloch/research/papers/index.html

Data

- Response variables may be
  - either continuous (such as biological indices of abundance) or
  - discrete (such as designated use attainment classes).

EPA NES example: response variable is lake-wide, summer average $\log_{10}$ Chlorophyll *a* concentration.

Data

Predictor variables in tree based methods may also be continuous or discrete, and may include :

source agency, basin, sub-watersheds, states, EPA regions, latitude and longitude, and many continuous predictors related to water chemistry, water use, discharges or pollutant loading.

Data

For the EPA NES example, Latitude and Longitude were used in the tree portion,

and

$\log_{10}Q_{in,}$ $\log_{10}$ Z $\mathrm{Log}_{10}\ \tau_w$

In-lake $\log_{10}$ TP In-lake $\log_{10}$ TN

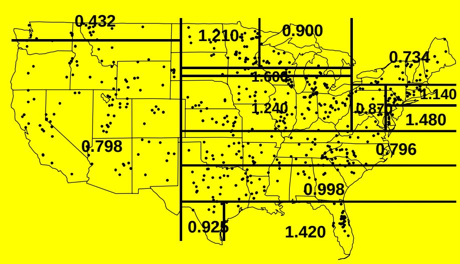
For the linear model within each bottom node (leaf)

This research is funded by
U.S. EPA - Science To Achieve Results (STAR) Program
Grant # RD-83088701-0

Preliminary Findings

Lamon, E.C., and C.A. Stow, 2004. Bayesian Methods for regional-scale eutrophication models, Water Research, 38(11): 2764-2774.

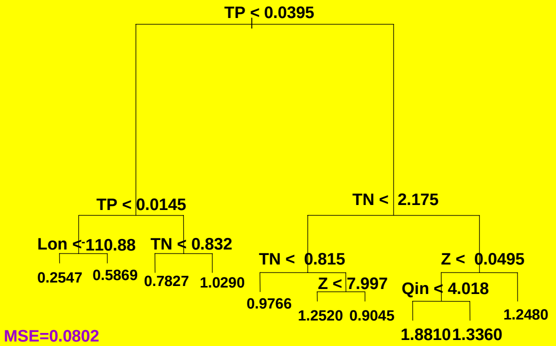
$\log_{10}(\mathrm{Chl}a) \sim$ latitude + longitude



MSE=0.1092
14 leaves

Preliminary Findings

$\log_{10}(\text{Chla}) \sim \text{TP}+\text{TN}+\text{Lat}+\text{Lon}+\text{Qin}+\text{Z}+\tau_w$



MSE=0.0802

10 leaves

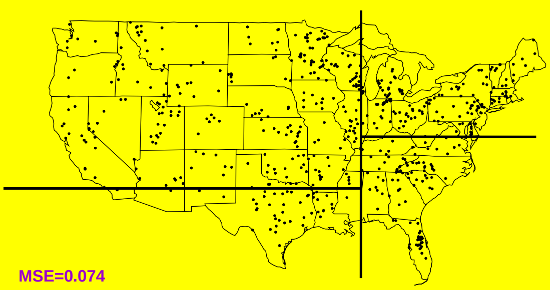
Preliminary Findings

Results

- Bayesian Treed Model

Preliminary Findings

BTREED Model – partitions (*T*)

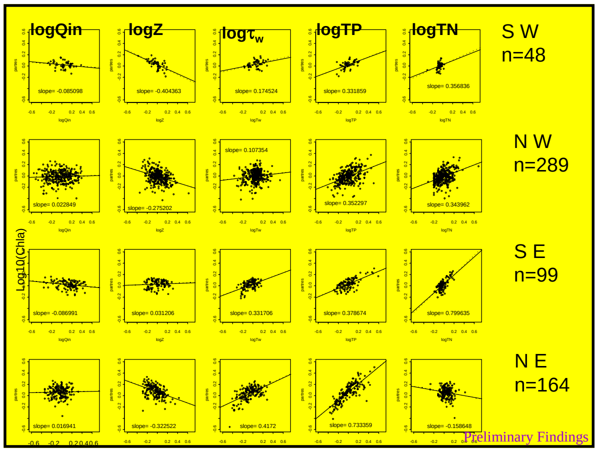


MSE=0.074

4 leaves

Region	Int.	$\log_{10}Q_{in}$	$\log_{10}Z$	$\log_{10}\tau_w$	In-lake $\log_{10}TP$	In-lake $\log_{10}TN$	MSE	n
SW	0.02166 0.0210 (0.0209)	-0.0851 -0.0691 (0.0927)	-0.4044 **-0.4390** (0.1426)	0.1745 0.2107 (0.1534)	0.3319 **0.3280** (0.1036)	0.3568 **0.4012** (0.1957)	0.027	48
NW	-0.0116 -0.0117 (0.0068)	0.0228 0.0241 (0.0478)	-0.2752 **-0.2763** (0.0647)	0.1074 0.1091 (0.0678)	0.3523 **0.3528** (0.0553)	0.3440 **0.3449** (0.0715)	0.095	289
SE	0.0290 **0.0299** (0.0090)	-0.0870 -0.0845 (0.0526)	0.0312 0.0385* (0.0734)	0.3317 **0.3325** (0.0862)	0.3787 **0.3683** (0.0772)	0.7996 **0.8456** (0.1514)	0.037	99
NE	0.0642 **0.0653** (0.0111)	0.0169 0.0222 (0.0734)	-0.3225 **-0.3306** (0.0997)	0.4172 **0.4275** (0.0854)	0.7334 **0.7398** (0.0653)	-0.1586 -0.1665 (0.1048)	0.073	164
total							**0.074**	**600**

Preliminary Findings



Preliminary Findings

Next Steps

Next Steps

- More predictor variables
- Apply these methods to the Nutrient Criteria Database
- Use resultant tree structures to identify important hierarchical structure
- Explore these structures with other Hierarchical Bayesian methods
- Non-linear specification? Spline basis functions in leaf model or inclusion of all predictors in tree
- Tools

Thanks!

This research is funded by U.S. EPA - Science To Achieve Results (STAR) Program
Grant # RD-83088701-0

- EPA STAR program for funding.
- Hugh Chipman, Univ. of Waterloo and Robert McColloch, University of Chicago for BCART/BTREED computer code.