Spatial Ensemble Estimates of Temporal Trends in Acid Neutralizing Capacity

F. Jay Breidt
Colorado State University

Joint work with Mark Delorey, Colorado State University

The work reported here was developed under STAR Research Assistance Agreements CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This presentation has not been formally reviewed by EPA. The views expressed here are solely those of the authors. EPA does not endorse any products or commercial services mentioned in this report.

PRogram for **I**nterdisciplinary **M**athematics, **E**cology, & **S**tatistics

- **PRIMES:** NSF-funded IGERT program
  - degree-plus program in quantitative ecology
  - generous fellowships for students ($27,500 for 03–04)
  - workshops, short courses, etc. encouraging team research
  - internship support by CDC, US Forest Service, and NCAR
- Five research focus groups:
  - Dynamics of Introduced Disease
  - Evolution in Structured Populations
  - Ecology of Managed Ecosystems
  - Ecology of Global Change
  - Aquatic Resources Modeling (STARMAP)

Preliminary Work on Temporal Trends in ANC

- Acid Neutralizing Capacity (ANC)
  - surface waters are acidic if ANC < 0
  - supply of acids from atmospheric deposition and watershed processes exceeds buffering capacity
- Temporal trends in ANC within watersheds (8-digit HUC's)
  - characterize the spatial **ensemble** of trends
  - make a map, construct a histogram, plot an empirical distribution function

Data Set

- 88 HUC's in Mid-Atlantic Highlands
- ANC in at least two years from 1993–1998
- HUC-level covariates:
  - area
  - average elevation
  - average slope, max slope
  - percents agriculture, urban, and forest
  - spatial coordinates

Small Area Estimation

- Probability sample across region
  - regional-level inferences are model-free
  - sample sizes too small to support HUC-level inferences
  - need to construct statistical model to borrow strength across areas
- Two standard types of small area models (Rao, 2003)
  - **area-level**: watersheds
  - **unit-level**: site within watershed

Basic Area-Level Model

- Temporal trend estimates:

$$\begin{aligned}\hat{\beta}_h &= \text{within-HUC WLS slope} \\ &= \beta_h + e_h \\ \beta_h &= \mathbf{x}_h^T \boldsymbol{\theta} + \omega_h\end{aligned}$$

- Design properties:

$$\mathrm{E}_p[e_h \,|\, \beta_h] = 0 \text{ and } \mathrm{Var}_p(e_h \,|\, \beta_h) = \psi_h$$

– design variances assumed known

- Model properties:

$$\mathrm{E}_m[\omega_h] = 0 \text{ and } \mathrm{Var}_m(\omega_h) = \sigma_\omega^2 \geq 0$$

Two Inferential Goals

- Interested in estimating **individual** HUC-specific slopes
- Also interested in **ensemble**:
  spatially-indexed true values: $\{\beta_h\}_{h=1}^m$
  spatially-indexed estimates: $\{\beta_h^{\text{est}}\}_{h=1}^m$
  - **subgroup analysis**: what proportion of HUC's have ANC decreasing over time?
  - "empirical" distribution function (**edf**):

$$F_\beta(z) = \frac{1}{m} \sum_{h=1}^m I_{\{\beta_h \le z\}}$$

Deconvolution Approach

- Treat this as measurement error problem

$$\hat{\beta}_h = \beta_h + e_h$$
$$\{e_h\} \sim \mathrm{N}(0, \psi_h)$$

- Deconvolve:
  - *parametric:* assume F_β in parametric class
  - *semi-parametric:* assume F_β well-approximated within class (like splines, normal mixtures)
  - *non-parametric:* assume $\mathrm{E}_{F_\beta}[e^{i\lambda\beta}]$ is smooth
- Not so appropriate for heteroskedastic measurements, explanatory variables, two inferential goals

Hierarchical Area-Level Model

- Extend model specification by describing parameter uncertainty:

$$
\begin{aligned}
\hat{\beta}_h &= \beta_h + e_h,\ \{e_h\}\ \mathrm{NID}(0, \psi_h) \\
\beta_h &= \mathbf{x}_h^T \boldsymbol{\theta} + \omega_h,\ \{\omega_h\}\ \mathrm{NID}(0, \sigma_\omega^2)
\end{aligned}
$$

- Prior specification:

$$
f(\boldsymbol{\theta}, \sigma_\omega^2) = f(\boldsymbol{\theta}) f(\sigma_\omega^2) \propto f(\sigma_\omega^2)
$$

Bayesian Inference

- **Individual** estimates: use posterior means

$$\beta_h^B = \mathrm{E}[\beta_h \mid \hat{\boldsymbol{\beta}}] = \mathrm{E}\left[\gamma_h \hat{\beta}_h + (1-\gamma_h)\mathbf{x}_h^T \boldsymbol{\theta} \mid \hat{\boldsymbol{\beta}}\right]$$

where $\gamma_h = \sigma_\omega^2/(\psi_h + \sigma_\omega^2)$

- Do Bayes estimates yield a good **ensemble** estimate?
  – use edf of Bayes estimates to estimate F_β?
- **No!** Bayes estimates are "over-shrunk"
  – too little variability to give good representation of edf (Louis 1984, Ghosh 1992)

$$\sum_{h=1}^{m} (\beta_h^B - \bar{\beta}^B)^2 < \mathrm{E}\left[\sum_{h=1}^{m} (\beta_h - \bar{\beta})^2 \mid \hat{\boldsymbol{\beta}}\right]$$

Adjusted Shrinkage

- Posterior means not good for *both* individual and ensemble estimates
- Improve by reducing shrinkage
  - sample mean of Bayes estimates already matches posterior mean of $\{\beta_h\}$
  - adjust shrinkage so that sample variance of estimates matches posterior variance of true values
- Louis (1984), Ghosh (1992)

Constrained Bayes Estimates

- Compute the scalars

$$H_1(\hat{\boldsymbol{\beta}}) = \operatorname{tr}\left\{\operatorname{Var}\left(\boldsymbol{\beta} - \bar{\beta}\mathbf{1} \mid \hat{\boldsymbol{\beta}}\right)\right\}$$

$$H_2(\hat{\boldsymbol{\beta}}) = \sum_{h=1}^{m} \left(\beta_h^B - \bar{\beta}^B\right)^2$$

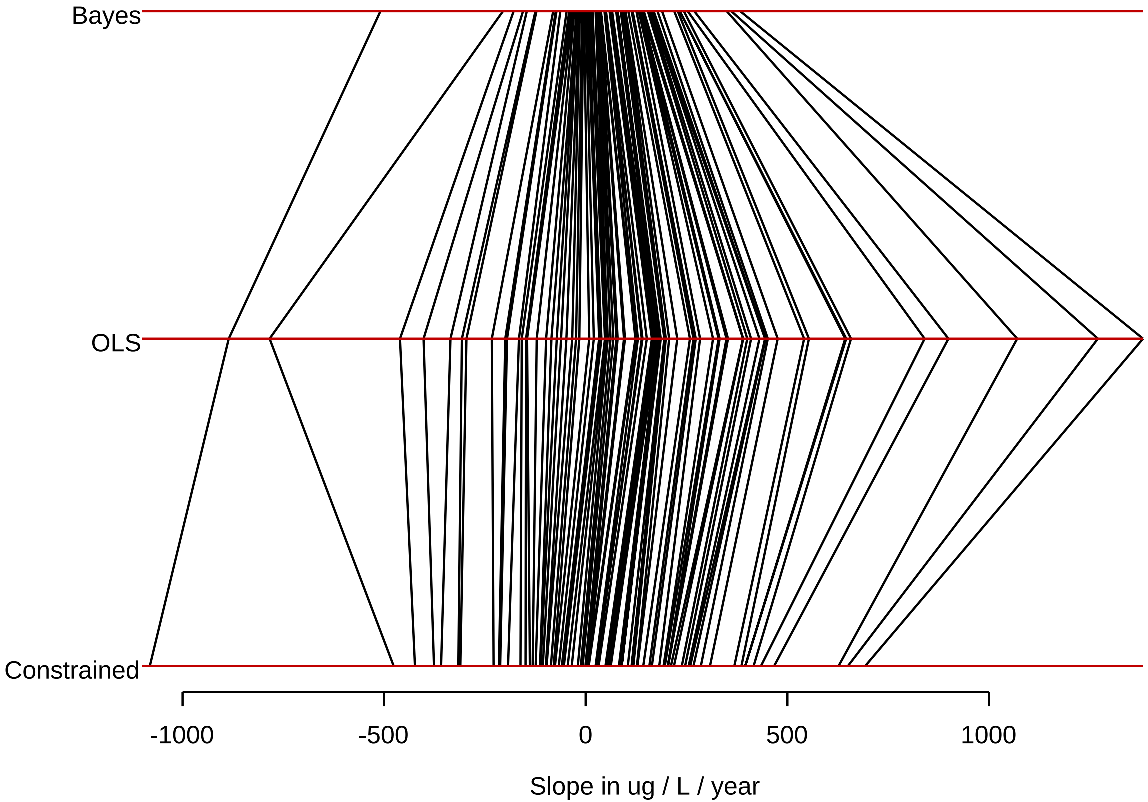
- Form the **constrained Bayes** (CB) estimates as

$$\beta_h^{CB} = a\beta_h^B + (1-a)\bar{\beta}^B$$

where

$$a = \left(1 + \frac{H_1(\hat{\boldsymbol{\beta}})}{H_2(\hat{\boldsymbol{\beta}})}\right)^{1/2} > 1$$

Shrinkage Comparisons for the Slope Ensemble



Numerical Illustration

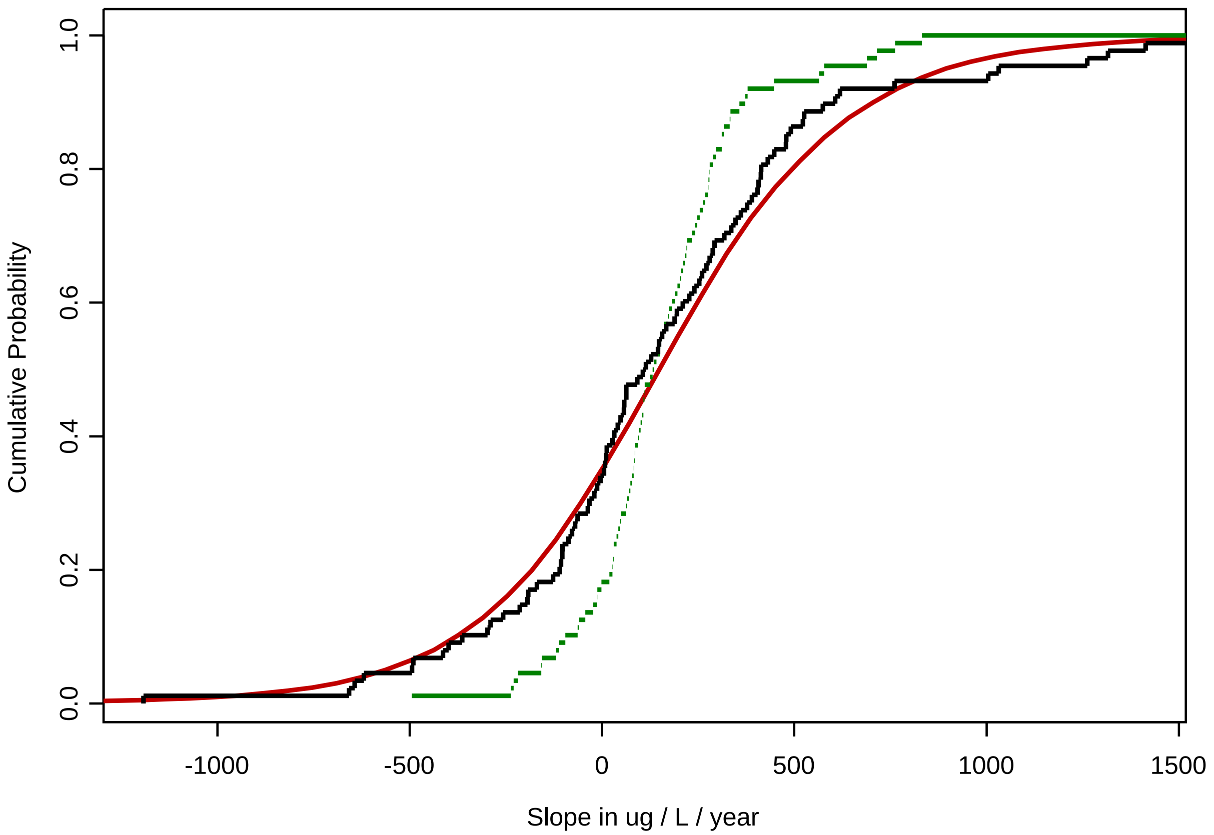
- Compare edf's of estimates to posterior mean of F_β:

$$F_\beta^B(z) = \frac{1}{m} \sum_{h=1}^{m} \mathrm{E}\left[I_{\{\beta_h \le z\}} \,|\, \hat{\boldsymbol{\beta}} \right]$$

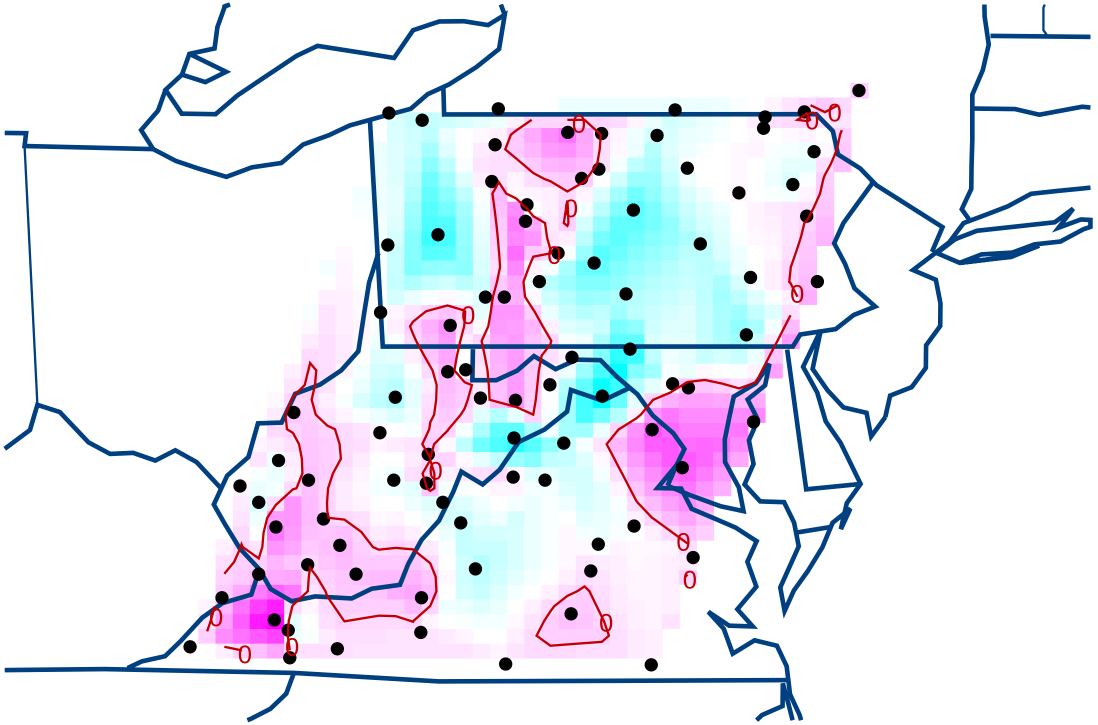
- Comparison of ensemble estimates at selected quantiles:

Estimate	$F_\beta(0)$	$F_\beta(400)$
edf of $\{\beta_h^B\}$	0.205	0.932
posterior mean	0.356	0.743
edf of $\{\beta_h^{CB}\}$	0.352	0.739

Estimated EDF's of the Slope Ensemble



Spatial Structure for the Slope Ensemble



Further Work: Spatial Model

- Let A_h denote set of neighboring HUC's for HUC h
- Conditional autoregression (CAR) model:

$$\begin{aligned}\hat{\beta}_h &= \beta_h + e_h \\ \beta_h &= \mathbf{x}_h^T \boldsymbol{\theta} + \omega_h \\ \omega_h \,|\, \{\omega_k, k \neq h\} &\sim \mathrm{N}\left(\rho \sum_{k \in A_h} q_{hk}\omega_k, \sigma_\omega^2\right)\end{aligned}$$

- Adjacency matrix $[q_{hk}]$ can reflect watershed structure

Further Work

- Restrict to acid-sensitive waters
- Combine probability and convenience samples
  - weights from selection functions to get $E_p[e_h \mid \beta_h] = 0$
- Other covariates?
  - deposition maps/trends from CASTNet?
- Other trend summaries?
- Site-level model?
  - useful sub-watershed covariates?
  - spatial scales: HUC to HUC, site to site
  - more concern with design, normality assumptions