This research is funded by
U.S.EPA – Science To Achieve Results (STAR) Program
Cooperative Agreement # CR - 829095

Linking CWA Sections 305(b)/303(d): Small Area Estimation

F. Jay Breidt
Colorado State University

EMAP Symposium 2004, Newport, Rhode Island

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.

Outline

- Primer on small area estimation
  - direct and indirect estimation
  - synthetic and composite estimation
  - borrowing strength and shrinkage
  - simultaneous and ensemble estimation
- Small area estimation examples
  - semi-parametric small area estimation
  - constrained estimation for ensembles

Domains

- $Domain =$ subpopulation of interest in a survey
  - geographic domains = areas (EPA region, state, county, HUC)
- Major domains: addressed by CWA 305(b)
  - sufficient sample size allocated at the design stage
  - standard survey estimation procedures yields estimates of adequate precision

Major Domains: Use Direct Estimation

- Direct estimators:
  - use data only from the study units in the domain and time period of interest
  - include standard weighted survey estimators
  - good design properties: unbiased estimator and valid confidence intervals *without any statistical model!*
- Direct estimation is not reliable if sample size is extremely small

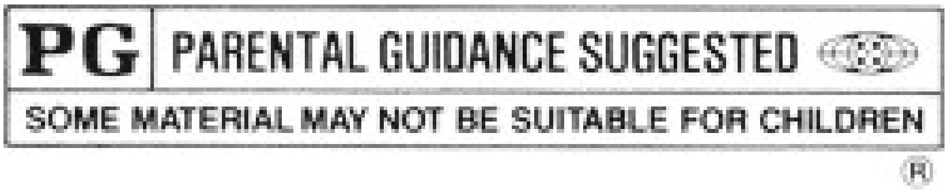
Small Domains: Direct Estimates Not Reliable

- Small domains/Small areas
  - sample size is small and may be zero in some domains
  - model-based inference is necessary to yield estimates of adequate precision
  - (definition depends on sampling resources and precision requirements)
- Might consider small area estimates for CWA 303(d)
  - rare to have adequate sample size everywhere

Indirect Estimation: Borrowing Strength

- Indirect estimators:
  - use data from outside the domain and/or time period of interest
  - (time indirect, domain indirect, domain and time indirect)
  - explicitly use statistical model to "borrow strength" across time or space
  - include various small area estimators

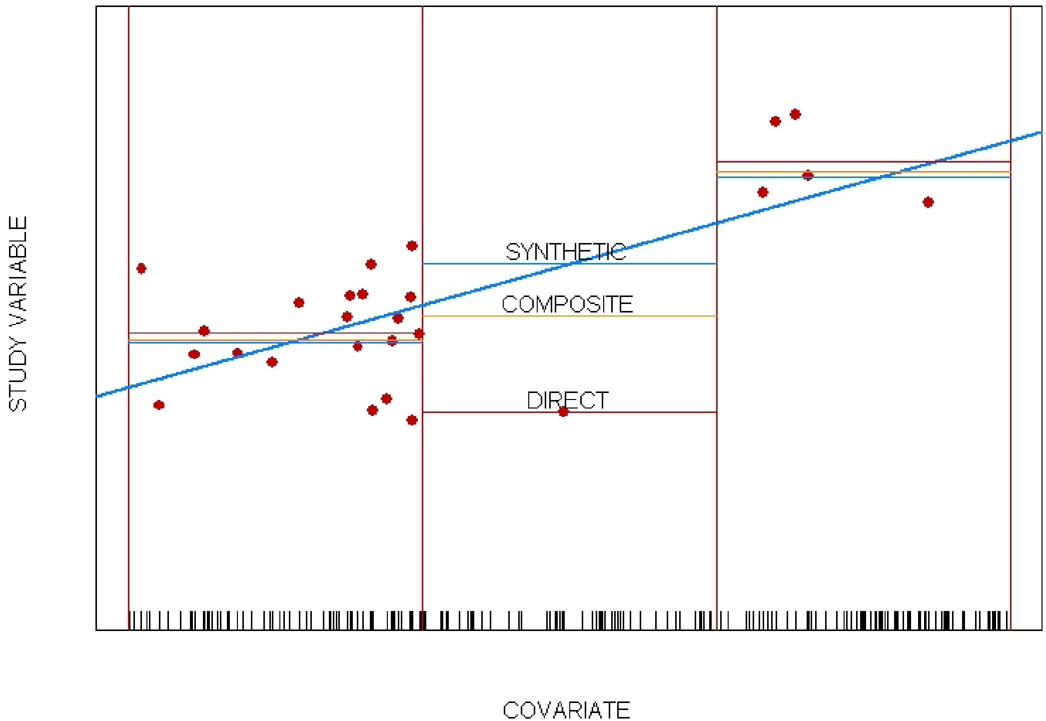
Indirect Estimation: Synthetic Estimator



- Have: study variables for sample, covariates for entire landscape
- Fit "global" model relating study variable to covariates
- Predict study variable at unobserved locations using available covariates and fitted model
  - works even if no samples in the area
  - may be poor if model is incorrectly specified

Direct, Synthetic and Composite Estimators

- One covariate, three small areas



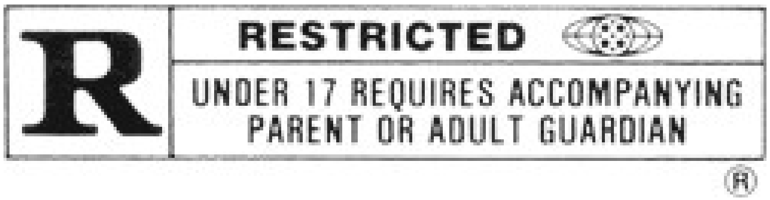
Shrinkage in the Composite Estimator

- Direct is moved toward synthetic to get composite estimator
  - equivalently, small-area specific effect "shrinks toward zero"
- Much of small area estimation involves choosing the shrinkage factor
- *Ad hoc* composite estimator

$$\text{composite} = w_h(\text{direct}) + (1 - w_h)(\text{synthetic})$$

  - still rated $\mathbf{PG}$

Formal Composite Estimation



- $w_h =$ function of parameters from a fitted mixed model
- Mature audiences only:
  - good auxiliary information
  - correctly-specified global regression structure
  - correctly-specified local correlation structure
  - (may require violence or coarse language)
  - sexy models and methods: EBLUP/EB, HB

Basic Small Area Models

- Model for direct estimates:

$$
\begin{aligned}
\hat{\theta}_h &= \text{direct estimate for small area } h \\
&= \theta_h + e_h \\
&= \text{truth+sampling error} \\
\theta_h &= \mathbf{x}_h^T \boldsymbol{\beta} + \omega_h \\
&= \text{regression + area-specific deviation}
\end{aligned}
$$

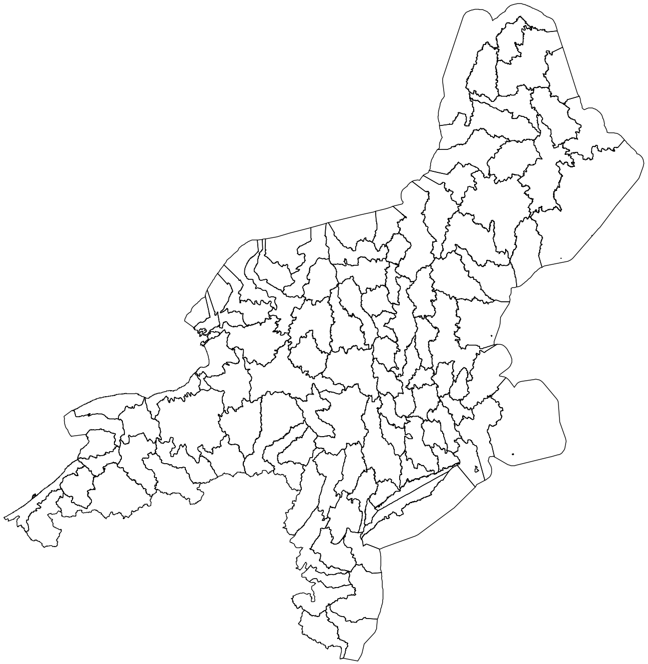
- Two ways to borrow strength:
  - globally, through regression fitted to all data
  - locally, through spatially (or temporally) correlated random effects

Two Small Area Estimation Problems

- Acid Neutralizing Capacity (ANC)
  - surface waters are acidic if ANC < 0
  - supply of acids from atmospheric deposition and watershed processes exceeds buffering capacity
- ANC level: Semiparametric small area estimation
  - HUCs in Northeast
- ANC trend: Constrained ensemble estimates
  - HUCs in mid-Atlantic highlands

Semiparametric Small Area Estimation of ANC Level

- Joint work with J. Opsomer, G. Ranalli, G. Claeskens, G. Kauermann
- 557 observations over 113 HUCs



HUCs as Small Areas

- Few sample observations available in most HUCs
  - Average sample size/HUC: 4.9
  - 64 HUCs contain less than 5 observations
  - 27 out of 113 HUCs contain no sample observations
- Site-specific covariates: lake location and elevation
  - need to account for spatial structure
  - worry about spatial model misspecification
- Simpler way to capture spatial effects?

Semiparametric Small Area Model

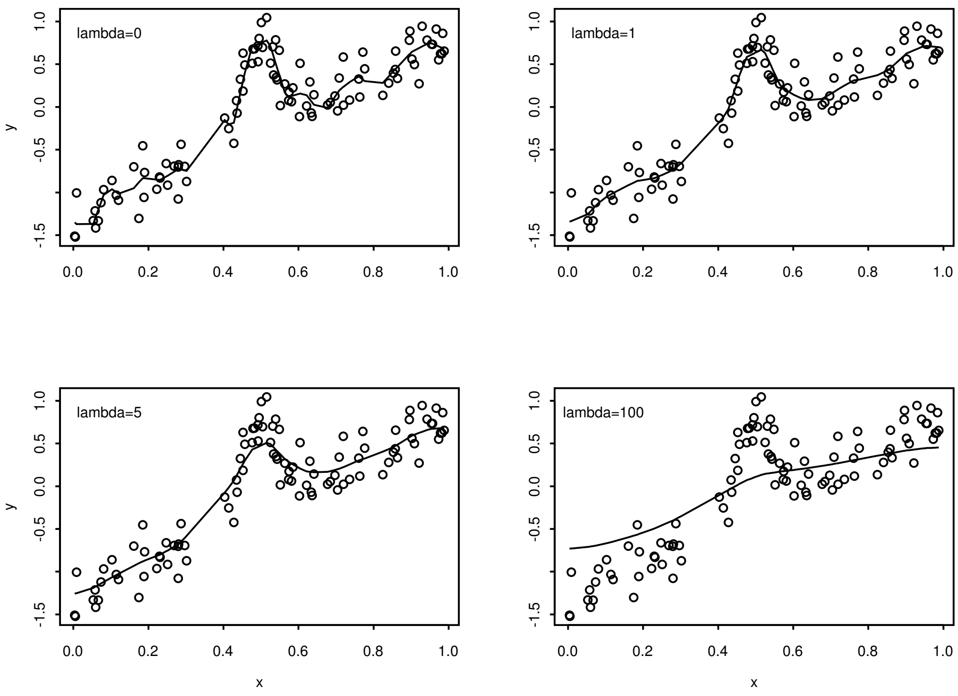
- Replace linear function of covariates by more general model:

$$\begin{aligned}
\text{direct} &= \text{truth}+\text{sampling error} \\
\text{truth} &= m(\mathbf{x}_h;\boldsymbol{\gamma}) + \omega_h \\
&= \text{semiparametric regression} + \text{area-specific deviation} \\
&= \mathbf{x}_h^T\boldsymbol{\beta} + \mathbf{z}_h^T\boldsymbol{\alpha} + \omega_h
\end{aligned}$$

- Semiparametric regression expressed as mixed linear model
  - penalized splines (P-splines)
  - thin plate splines
  - kriging
- EBLUP easily handled with standard software (SAS, SPlus)

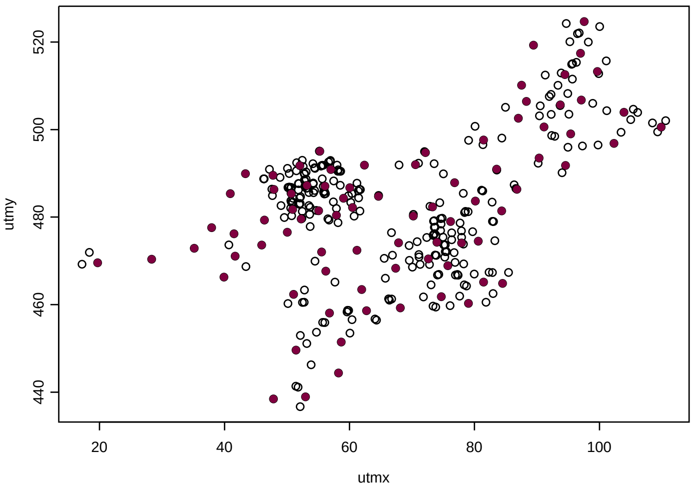
Fitting by Penalized Splines Regression

- Allow slope changes at each of many knots
  - penalize excessive slope changes via λ



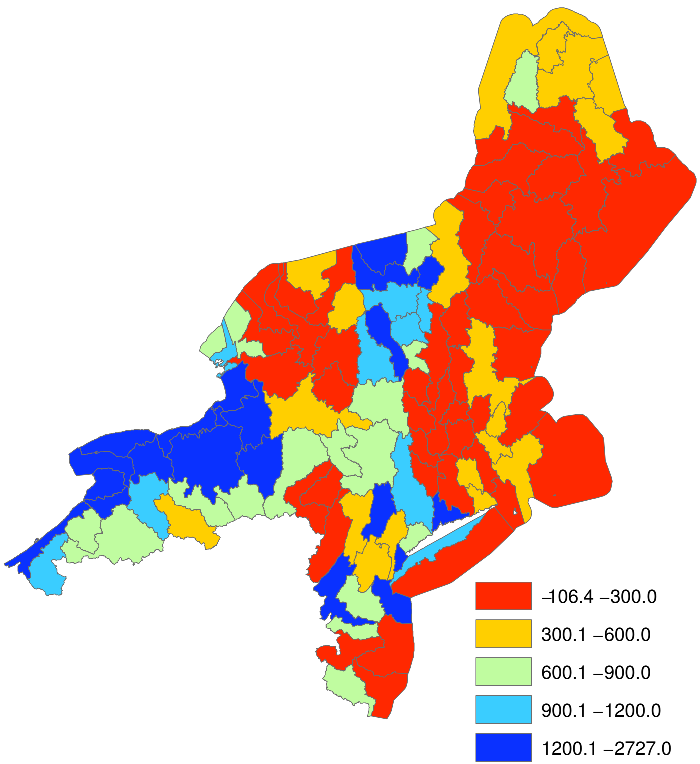
Spatial Smoothing Using P-Splines

- NE Lakes data require bivariate (spatial) smoothing $\approx$ thin-plate spline (Ruppert *et al.* 2003)
- Knot selection: space-filling algorithm



NE Lakes HUC Predictions

- Correlation between ANC and model prediction: 0.96



Constrained Bayes Estimation for ANC Trend

- Joint work with M. Delorey
- 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 Model for Trend Estimates

- Temporal trend estimates:

$$\begin{aligned}\hat{\tau}_h &= \text{within-HUC estimated slope} = \tau_h + e_h \\ &= \text{truth} + \text{sampling error} \\ \tau_h &= \mathbf{x}_h^T \boldsymbol{\beta} + \omega_h \\ &= \text{regression} + \text{area-specific effect}\end{aligned}$$

- Spatial correlation in $\{\omega_h\}$ modeled by conditional autoregression (CAR)

Two Inferential Goals

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

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

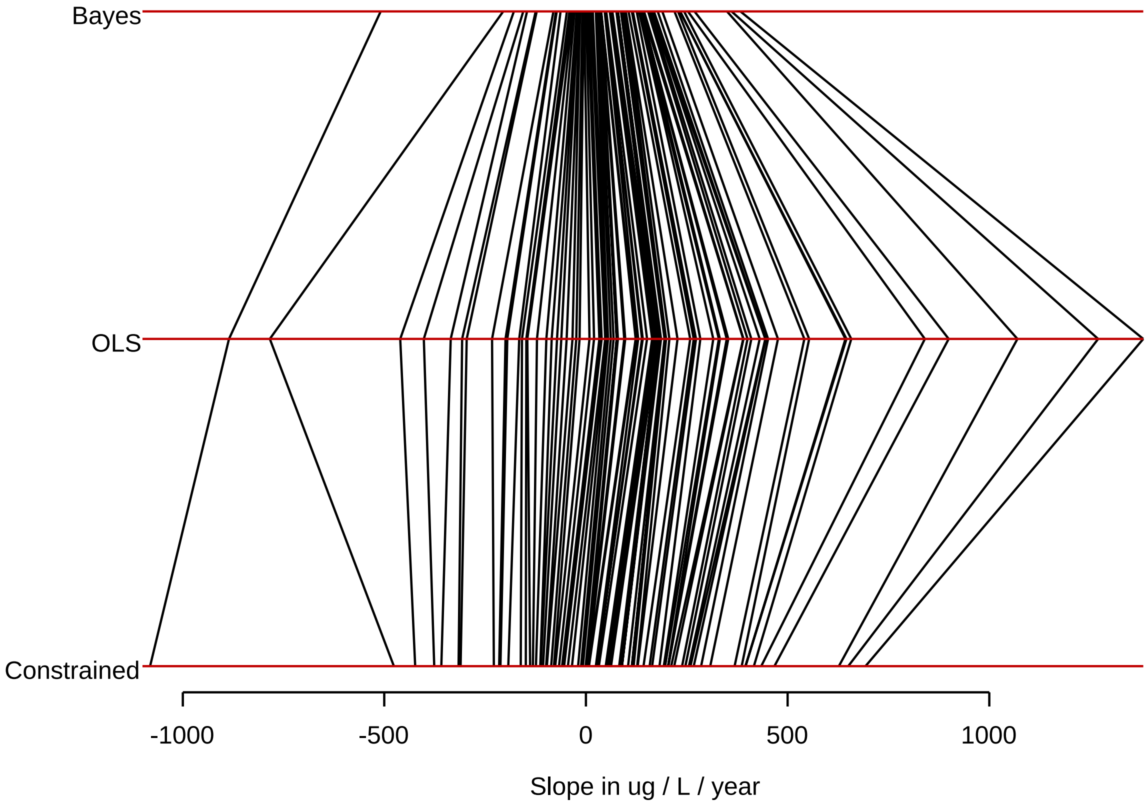
Bayesian Inference

- **Individual** estimates: use posterior means
  - pretty much sophisticated composite estimators
- Do Bayes estimates yield a good **ensemble** estimate?
  - use edf of Bayes estimates to estimate $F_{\mathcal{T}}$?
- **No!** Bayes estimates are "over-shrunk"
  - too little variability to give good representation of edf (Louis 1984, Ghosh 1992)

Constrained Bayes Adjusts the 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 $\{\tau_h\}$
  - adjust shrinkage so that sample variance of estimates matches posterior variance of true values
- Resulting estimates are called **Constrained Bayes**
  - Louis (1984), Ghosh (1992)
  - require posterior analysis

Shrinkage Comparisons for the Slope Ensemble



Numerical Implementation of Hierarchical Bayes

- Markov chain Monte Carlo (MCMC): often necessary to approximate posterior distribution of unknowns given data
- Idea: any distribution can be studied provided we can simulate from it
  - iid draws from distribution would be ideal
  - dependent, identically distributed draws would be fine if dependence is not too strong (ergodic theorem)
  - dependent, nearly identically distributed draws might be OK

Markov Chain Monte Carlo (MCMC)

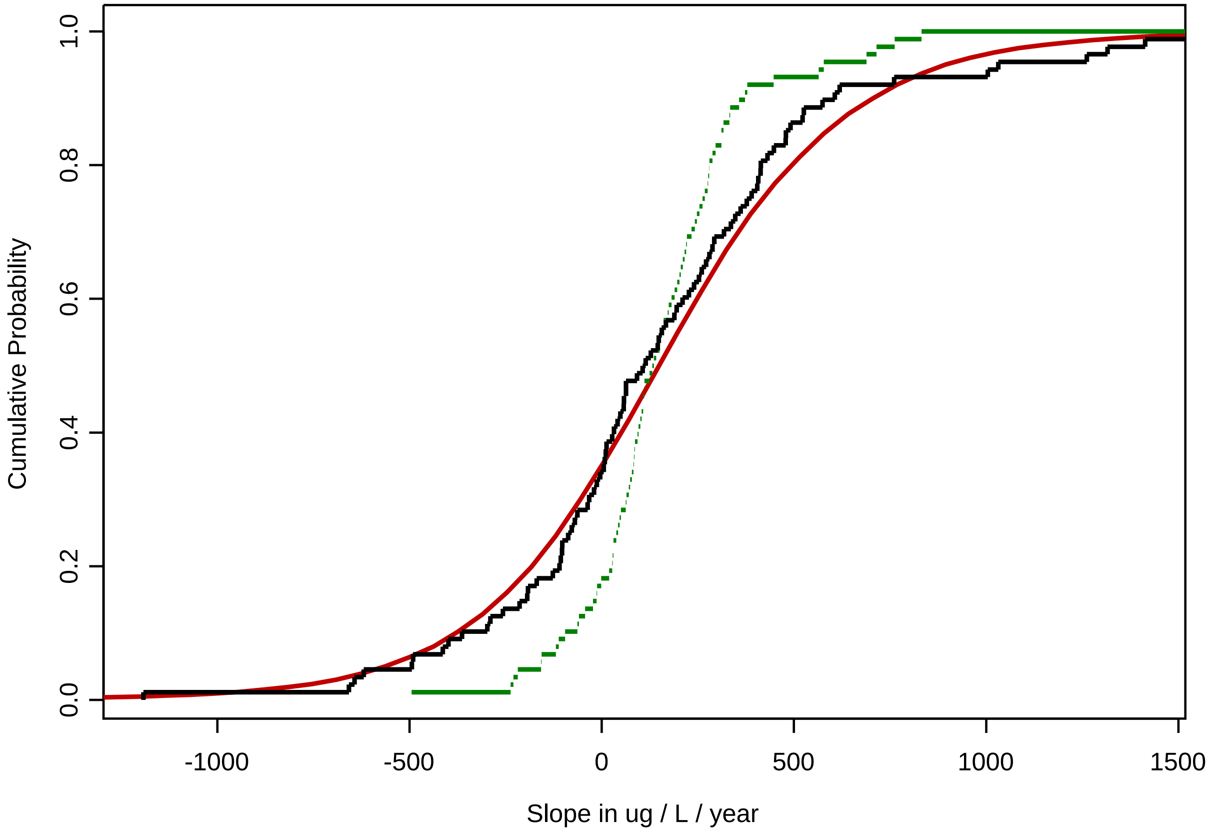
- MCMC generates Markov chain with invariant distribution equal to posterior distribution of interest
  - not independent due to Markov structure
  - not identically distributed except asymptotically, due to initialization problem
  - assessing convergence is critical
- MCMC recipes for constructing suitable Markov chains include
  - Gibbs sampler
  - Metropolis-Hastings algorithm

Gibbs Sampler: DISCO

- Derive full set of conditionals
- Initialize unknowns
- Sample sequentially from conditionals many times
- Check convergence, discarding a large number of "burn-in" draws
- Ordinary data analysis on remaining data set

$$\text{posterior mean of } \tau_h \simeq \text{sample mean of draws}$$
$$\text{posterior variance of } \tau_h \simeq \text{sample variance of draws}$$
$$\text{posterior median of } \tau_h \simeq \text{sample median of draws}$$

Estimated EDF's of the Slope Ensemble



Small Area Estimation Needed to Link 305(b) and 303(d)

- $\mathbf{G}$-rated direct estimates: no shrinkage
- Indirect estimates: $\mathbf{PG}$ or $\mathbf{R}$
  - need good covariates and/or useful correlations
  - rare in aquatic resources
- Shrinkage:
  - none = direct: $\mathbf{G}$-rated
  - total = synthetic: $\mathbf{PG}$-rated
  - ad hoc composite: $\mathbf{PG}$-rated
  - formal composite: $\mathbf{R}$-rated
- Two examples: semiparametric and constrained