Small Area Estimation for Natural Resource Surveys

F. Jay Breidt
Colorado State University

Monitoring Science & Technology Symposium
Denver, Colorado, September 23, 2004

The work reported here was developed under STAR Research Assistance Agreements CR-829095 and CR-829096 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University and Oregon 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
- Small area estimation examples
  - semi-parametric small area estimation
  - constrained estimation for ensembles

Domains

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

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

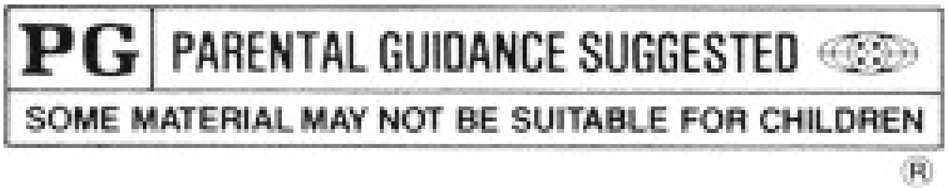
Small Domains: Direct Estimates Not Reliable

- 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)
- Typical map is covered with small areas

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

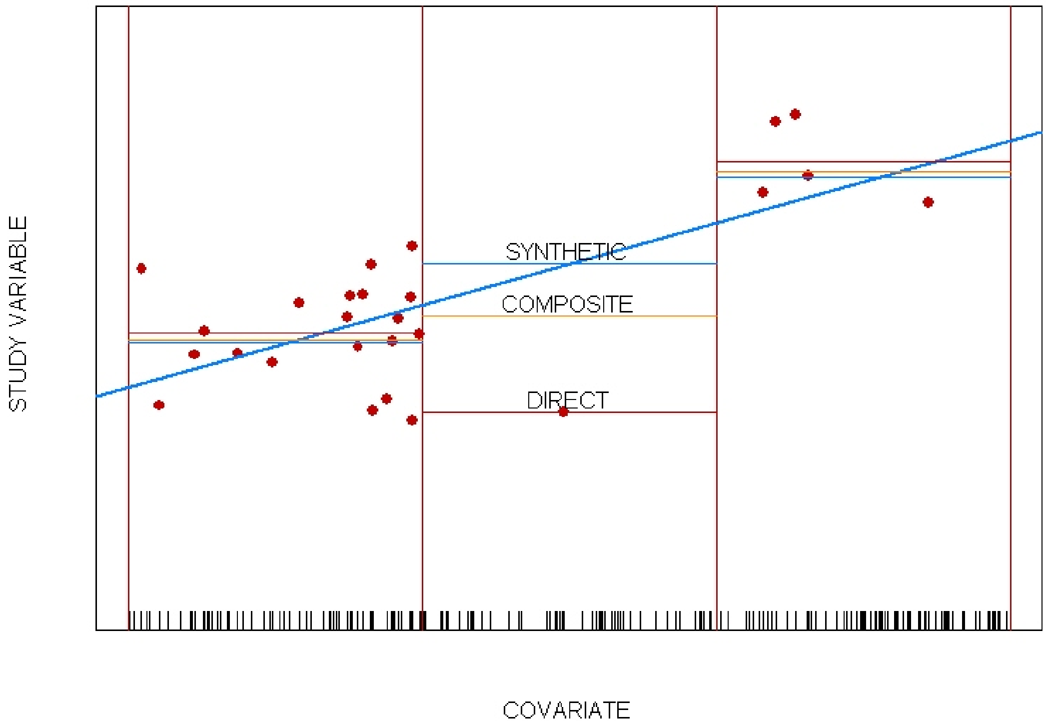
Indirect Estimation: Synthetic Estimator



- Have: response variables for sample, covariates for entire landscape
- Fit "global" model relating response variable to covariates
- Predict response 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



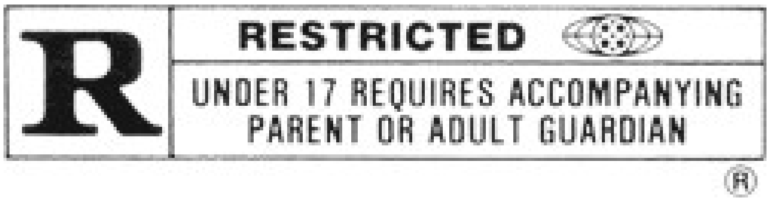
Shrinkage in the Composite Estimator

- 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

Small Area Models

- Model for direct estimates:

$$
\begin{aligned}
\hat{\theta}_h &= \text{direct estimate for small area } h \\
&= \theta_h + e_h \\
&= \text{truth+sampling error} \\
\theta_h &= \mathbf{x}_h^T \boldsymbol{\beta} + \omega_h, \quad [\omega_h] \sim (\mathbf{0}, \Gamma) \\
&= \text{regression + area-specific deviation}
\end{aligned}
$$

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

Borrowing Strength

- Only global, through regression fitted to all data

$$\mathbf{x}_h^T \hat{\boldsymbol{\beta}} = \mathbf{PG}\text{-rated synthetic estimator}$$

- Both global and local, allowing spatially (or temporally) correlated random effects

$$\mathbf{x}_h^T \hat{\boldsymbol{\beta}} + \hat{\omega}_h = \mathbf{R}\text{-rated composite estimator}$$

Fitting Small Area Models

- Model:

$$\begin{aligned}\hat{\theta}_h &= \theta_h + e_h \\ \theta_h &= \mathbf{x}_h^T \boldsymbol{\beta} + \omega_h, \quad [\omega_h] \sim (\mathbf{0}, \Gamma)\end{aligned}$$

- Empirical BLUP/ Empirical Bayes: Compute $\hat{\boldsymbol{\beta}}$ and $\hat{\Gamma}$ and plug in to get

$$\tilde{\theta}_h = \mathbf{x}_h^T \hat{\boldsymbol{\beta}} + \hat{\omega}_h$$

- Hierarchical Bayes (HB): Compute $\boldsymbol{\beta} \,|\, \text{data}$ and $\Gamma \,|\, \text{data}$ and plug in to get

$$\theta_h \,|\, \text{data}$$

Comparison of Estimation Methods

- Empirical BLUP/ Empirical Bayes:
  - relatively straightforward computation
  - can use SAS `proc mixed` or S-Plus function `lme()`
  - does not fully account for uncertainty due to unknown variance components
- Hierarchical Bayes:
  - entire posterior distribution, not just point estimates
  - full accounting for uncertainty (assuming correct model specification)
  - computation is typically much more involved, though sometimes can be done in `winbugs`

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

Cautions on Small Area Estimation

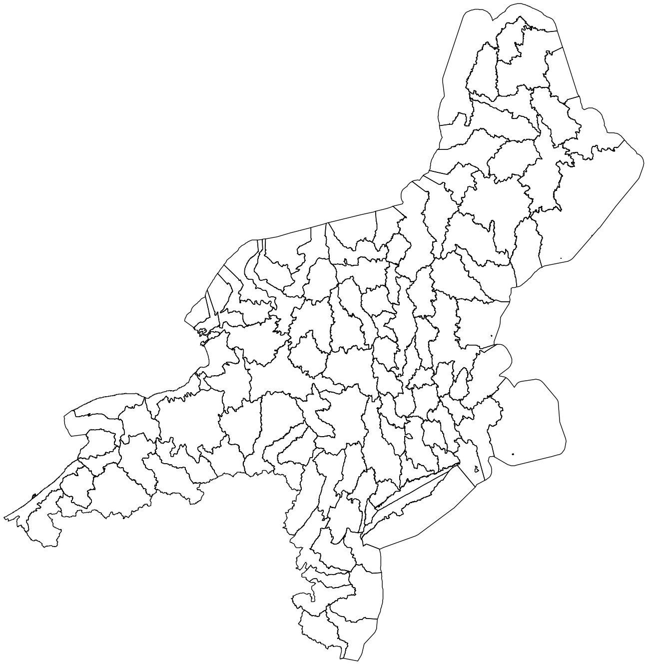
- Be wary of unreasonable expectations
- Global regression:
  - covariates from remote sensing may have poor explanatory power
  - particularly true for aquatic responses
- Local scale:
  - spatial correlations may be weak at scale of probability sample
  - particularly true once useful covariates taken into account
  - sampling may not be sufficiently dense for good estimates of spatial structure

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
  - not much available for global regression
  - local structure: can use correlation or flexible trend

Semiparametric Small Area Model

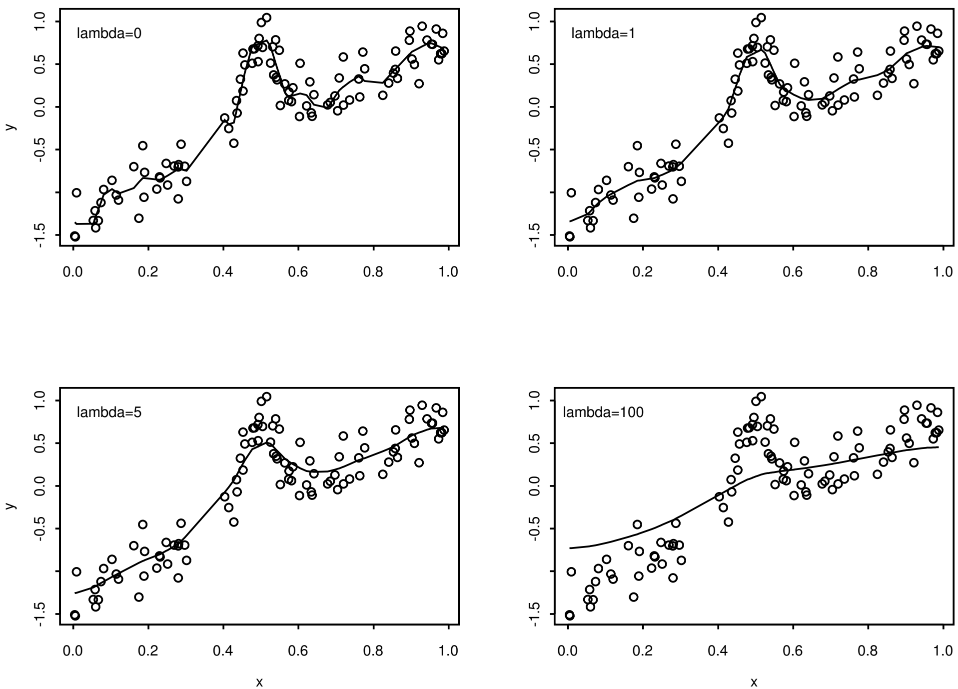
- Replace linear function of covariates by more general model:

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

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

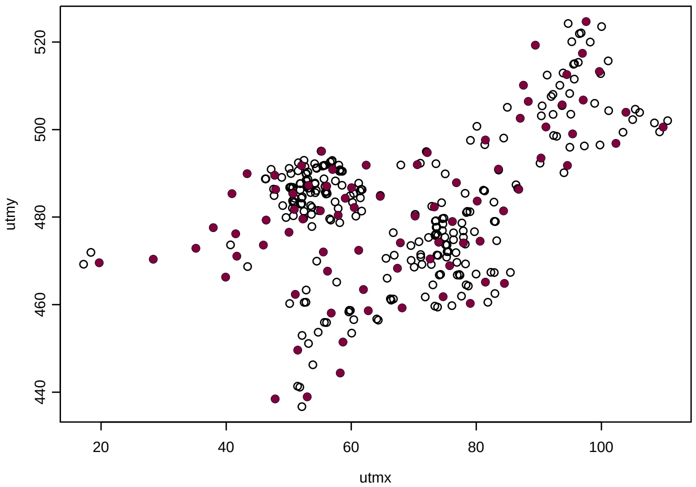
Fitting by Penalized Splines Regression

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



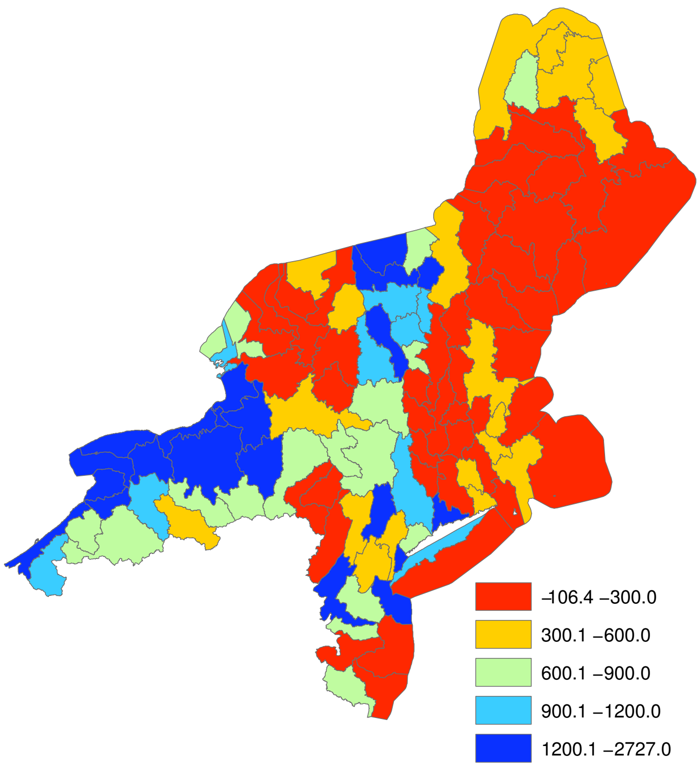
Spatial Smoothing Using P-Splines

- 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

- Mark Delorey (PhD student)
  - "Precision Monitoring Approaches for Research and Management Applications" session
- 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?
- Hierarchical Bayes "overshrinks" in this context

Small Area Estimation Summary

- $\mathbf{G}$-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