

Small Area Estimation Using Penalized Spline Regression

Jean Opsomer
Iowa State University

Joint work with
Gerda Claeskens, Katholieke Universiteit Leuven
Giovanna Ranalli, Università di Perugia
Göran Kauermann, Universität Bielefeld
Jay Breidt, Colorado State University

August 7, 2005



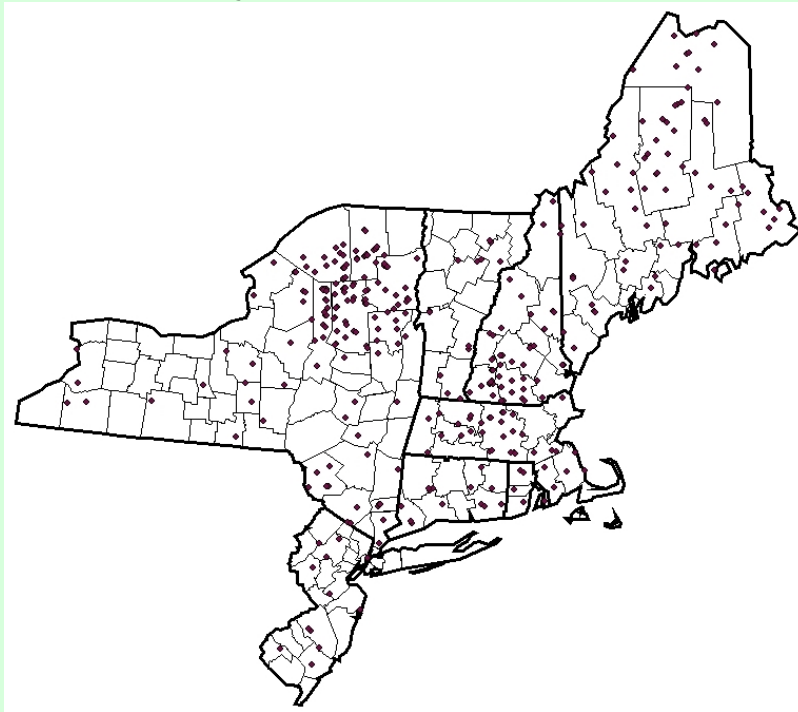
Outline

1. Introduction: Northeastern Lakes Survey
2. Methods
 - (a) Nonparametric regression using penalized splines
 - (b) Small area estimation using mixed models
3. Nonparametric small area estimation
 - (a) Estimator
 - (b) Bootstrap inference
4. Northeastern Lakes Survey
5. Conclusion

[Back](#)[Close](#)

1. Introduction: Lakes Survey

Ecological condition survey of Northeastern lakes conducted by EPA



Data collected for 334 lakes (551 observations) in 1991-1996

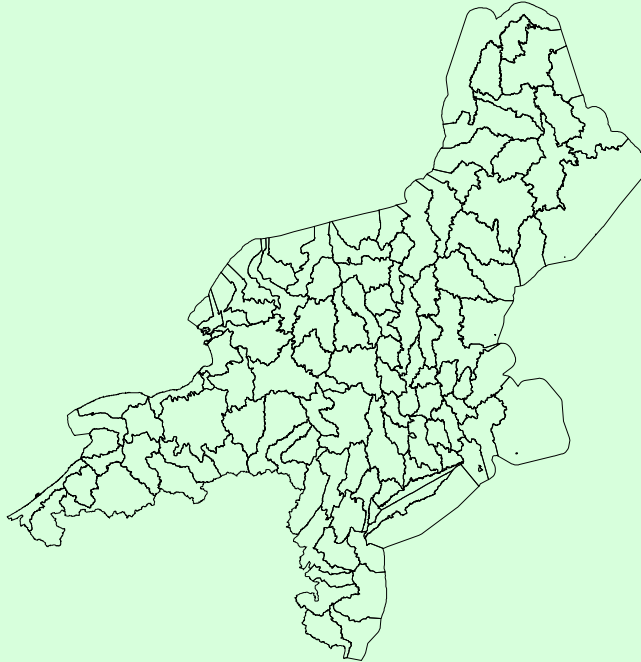


Back

Close

Northeastern Lakes Survey (2)

Region includes 113 8-digit “Hydrologic Unit Codes” (HUC)



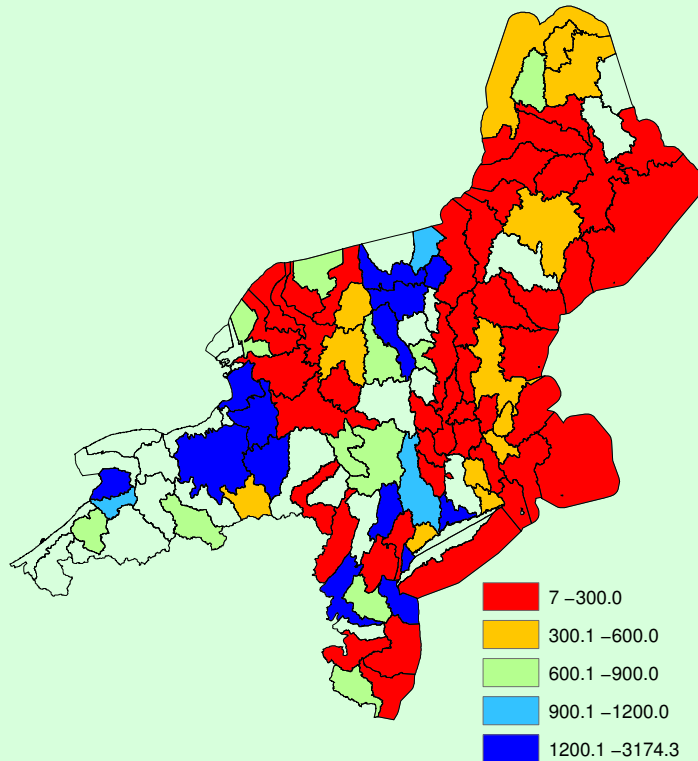
Goal: estimate mean lake *Acid Neutralizing Capacity* (ANC) for all HUCs



Back

Close

HUC Sample Means



Problems: unreliable estimates, missing HUCs



Back

Close

Nonparametric small area estimation?

- overall mean model captures relationship between available covariates and ANC, supplemented by HUC-specific effect
- no a priori obvious choice for parametric shape for mean model

[Back](#)[Close](#)

2.1) Methods: Nonparametric Regression Using Penalized Splines

Regression model $y_i = m(x_i) + \varepsilon_i$

Function $m(\cdot)$ is unknown but assumed well approximated by *polynomial spline*

$$m(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \sum_{k=1}^K \beta_{p+k} (x - \kappa_k)_+^p$$

p : degree of spline (fixed)

$\kappa_1 < \dots < \kappa_K$: set of K knots (fixed)

$\boldsymbol{\beta} = (\beta_0, \dots, \beta_{p+K})$: vector of parameters (unknown)

(Ruppert, Wand and Carroll, 2003)

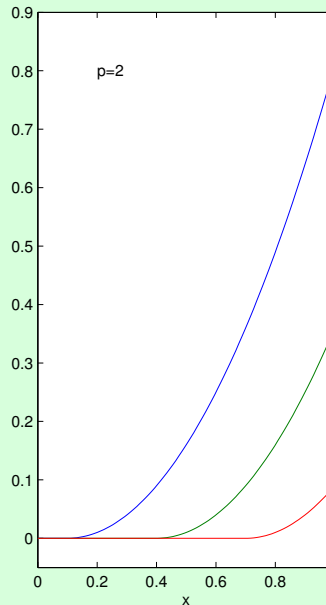
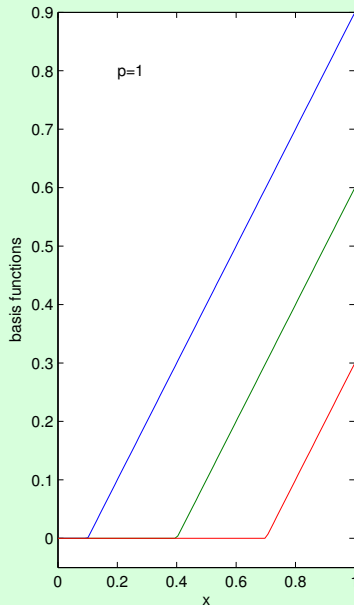


Back

Close

Polynomial Spline Basis Functions

$$(x - \kappa)_+^p \equiv \begin{cases} (x - \kappa)^p & \text{if } x - \kappa > 0 \\ 0 & \text{if } x - \kappa \leq 0 \end{cases}$$



Other spline basis functions are possible (B-splines, radial splines)



Expressing Spline Model as Parametric Model

$$\begin{aligned} m(x; \boldsymbol{\beta}) &= \beta_0 + \beta_1 x + \dots + \beta_p x^p + \sum_{k=1}^K \beta_{p+k} (x - \kappa_k)_+^p \\ &\equiv \mathbf{x}^* \boldsymbol{\beta} \end{aligned}$$

with

$$\begin{aligned} \mathbf{x}^* &= (1, x, \dots, x^p, (x - \kappa_1)_+^p, \dots, (x - \kappa_K)_+^p) \\ \boldsymbol{\beta} &= (\beta_0, \dots, \beta_{p+K})^T \end{aligned}$$

K large, p small



Back

Close

Splines as Random Effect

$$\begin{aligned}y_i &= m(x_i; \boldsymbol{\beta}) + \varepsilon_i \\ &= \mathbf{x}_i^* \boldsymbol{\beta} + \varepsilon_i\end{aligned}$$

[Back](#)[Close](#)

Splines as Random Effect

$$\begin{aligned}
 y_i &= m(x_i; \boldsymbol{\beta}) + \varepsilon_i \\
 &= \mathbf{x}_i^* \boldsymbol{\beta} + \varepsilon_i \\
 &= \mathbf{x}_i^F \boldsymbol{\beta}^F + \mathbf{z}_i \boldsymbol{\gamma} + \varepsilon_i
 \end{aligned}$$

$$\mathbf{x}_i^F \boldsymbol{\beta}^F \equiv \beta_0 + x_i \beta_1 + \dots + x_i^p \beta_p \quad (\text{parametric, fixed component})$$

$$\begin{aligned}
 \mathbf{z}_i \boldsymbol{\gamma} &= z_{1i} \gamma_1 + \dots + z_{Ki} \gamma_K \\
 &\equiv (x_i - \kappa_1)_+^p \beta_{p+1} + \dots + (x_i - \kappa_K)_+^p \beta_{p+K} \\
 &\quad (\text{deviations from parametric,} \\
 &\quad \text{treated as random effect})
 \end{aligned}$$

$$\begin{aligned}
 \boldsymbol{\gamma} &= (\gamma_1, \dots, \gamma_K) \sim \text{iid } \mathcal{F}_\gamma(0, \sigma_\gamma^2) \\
 \varepsilon_i &\sim \mathcal{F}_\varepsilon(0, \sigma_\varepsilon^2)
 \end{aligned}$$



P-splines Estimator as BLUP

Assuming $\sigma_\gamma^2, \sigma_\varepsilon^2$ known, *Best Linear Unbiased Estimator/Predictor* (BLUE/BLUP) is solution to

$$\min_{\boldsymbol{\beta}^F, \boldsymbol{\gamma}} \sum_{i=1}^n (y_i - \mathbf{x}_i^F \boldsymbol{\beta}^F + \mathbf{z}_i \boldsymbol{\gamma})^2 + \frac{\sigma_\varepsilon^2}{\sigma_\gamma^2} \sum_{k=1}^K \gamma_k^2$$

(Henderson *et al.*, 1959)

$$\Rightarrow \begin{bmatrix} \hat{\boldsymbol{\beta}}^F \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \left(\mathbf{X}^{*T} \mathbf{X}^* + \frac{\sigma_\varepsilon^2}{\sigma_\gamma^2} \mathbf{D} \right)^{-1} \mathbf{X}^{*T} \mathbf{Y}$$

with

$$\begin{aligned} \mathbf{D} &= \text{diag}\{0, \dots, 0, 1, \dots, 1\} \\ \mathbf{X}^* &= [\mathbf{x}^F \ \mathbf{z}] \end{aligned}$$



Back

Close

P-splines Estimator as EBLUP

If $\sigma_\gamma^2, \sigma_\varepsilon^2$ are unknown, estimates can be obtained by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) (e.g. Searle, Casella and McCulloch, 1992), and

$$\begin{bmatrix} \hat{\beta}^F \\ \hat{\gamma} \end{bmatrix} = \left(\mathbf{X}^{*T} \mathbf{X}^* + \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\gamma^2} \mathbf{D} \right)^{-1} \mathbf{X}^{*T} \mathbf{Y}$$

Two interpretations for parameter estimators

- = Spline mean function estimator with penalty determined by data
- = *Empirical BLUP* (EBLUP) for spline parameters under random effects model



Back

Close

Spatial Smoothing using P-splines

- NE Lakes data require bivariate (spatial) smoothing
- Low-rank radial basis (\approx thin-plate spline)

$$\mathbf{z} = \left[C(\mathbf{x}_i - \boldsymbol{\kappa}_k) \right]_{\substack{1 \leq i \leq n \\ 1 \leq k \leq K}} \left[C(\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'}) \right]_{1 \leq k, k' \leq K}^{-1/2}$$

with $C(\mathbf{r}) = \|\mathbf{r}\|^2 \log \|\mathbf{r}\|$ (Ruppert *et al.* 2003)

- Mixed model

$$y_i = \mathbf{x}_i^F \boldsymbol{\beta}^F + \mathbf{z}_i \boldsymbol{\gamma} + \varepsilon_i$$

$$\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K) \sim \text{iid } \mathcal{F}_\gamma(0, \sigma_\gamma^2)$$

$$\varepsilon_i \sim \mathcal{F}_\varepsilon(0, \sigma_\varepsilon^2)$$

- Knot selection: regular spatial grid, space-filling algorithm



Back

Close

2.2) Small Area Estimation as Mixed Effect Regression

“Classical” small area estimation (Battese, Harter and Fuller, 1988):

- T small areas, with u_t the random effect for small area $t = 1, \dots, T$
- Model

$$\begin{aligned}y_i &= \mathbf{x}_i\boldsymbol{\beta} + u_t + \varepsilon_i \\ &= \mathbf{x}_i\boldsymbol{\beta} + \mathbf{d}_i\mathbf{u} + \varepsilon_i\end{aligned}$$

$$\mathbf{d}_i = (d_{i1}, \dots, d_{iT}) \quad d_{it} = \begin{cases} 1 & \text{if } i \in \text{small area } t \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{u} = (u_1, \dots, u_T) \sim \text{iid } \mathcal{F}_u(0, \sigma_u^2)$$

$$\varepsilon_i \sim \mathcal{F}_\varepsilon(0, \sigma_\varepsilon^2)$$

- Target $\bar{Y}_t = \bar{\mathbf{X}}_t\boldsymbol{\beta} + u_t$ estimated by (E)BLUP $\bar{y}_t = \bar{\mathbf{X}}_t\hat{\boldsymbol{\beta}} + \hat{u}_t$



Back

Close

3. Nonparametric Small Area Model

Combine both random effects models

$$\begin{aligned}y_i &= m(x_i; \boldsymbol{\beta}) + \mathbf{d}_i \mathbf{u} + \varepsilon_i \\ &= \mathbf{x}_i^F \boldsymbol{\beta}^F + \mathbf{z}_i \boldsymbol{\gamma} + \mathbf{d}_i \mathbf{u} + \varepsilon_i\end{aligned}$$

Variance components

$$\boldsymbol{\gamma} \sim \text{iid } \mathcal{F}_{\boldsymbol{\gamma}}(0, \sigma_{\boldsymbol{\gamma}}^2)$$

$$\mathbf{u} \sim \text{iid } \mathcal{F}_{\mathbf{u}}(0, \sigma_{\mathbf{u}}^2)$$

$$\varepsilon_i \sim \mathcal{F}_{\varepsilon}(0, \sigma_{\varepsilon}^2)$$



Back

Close

3. Nonparametric Small Area Model

Combine both random effects models

$$\begin{aligned}y_i &= m(x_i; \boldsymbol{\beta}) + \mathbf{d}_i \mathbf{u} + \varepsilon_i \\ &= \mathbf{x}_i^F \boldsymbol{\beta}^F + \mathbf{z}_i \boldsymbol{\gamma} + \mathbf{d}_i \mathbf{u} + \varepsilon_i\end{aligned}$$

Variance components

$$\begin{aligned}\boldsymbol{\gamma} &\sim \text{iid } \mathcal{F}_\gamma(0, \sigma_\gamma^2) \\ \mathbf{u} &\sim \text{iid } \mathcal{F}_u(0, \sigma_u^2) \\ \varepsilon_i &\sim \mathcal{F}_\varepsilon(0, \sigma_\varepsilon^2)\end{aligned}$$

EBLUP can be computed by REML, and

$$\bar{y}_t = \bar{\mathbf{X}}_t^F \hat{\boldsymbol{\beta}}^F + \bar{\mathbf{Z}}_t \hat{\boldsymbol{\gamma}} + \hat{u}_t$$

estimates $\bar{Y}_t = \bar{\mathbf{X}}_t^F \boldsymbol{\beta} + \bar{\mathbf{Z}}_t \boldsymbol{\gamma} + u_t$



Back

Close

Inference for Random Effects

- Likelihood ratio tests for hypotheses
 - (1) $\sigma_u^2 = 0$
 - (2) $\sigma_\gamma^2 = 0$
 - (3) $\sigma_u^2 = \sigma_\gamma^2 = 0$
- Asymptotic distributions of LR statistics
 - $0.5\chi_0^2 + 0.5\chi_1^2$ for (1) and (2)
 - complicated mixture of $\chi_0^2, \chi_1^2, \chi_2^2$ for (3)



Back

Close

Inference for Random Effects

- Likelihood ratio tests for hypotheses
 - (1) $\sigma_u^2 = 0$
 - (2) $\sigma_\gamma^2 = 0$
 - (3) $\sigma_u^2 = \sigma_\gamma^2 = 0$
- Asymptotic distributions of LR statistics
 - $0.5\chi_0^2 + 0.5\chi_1^2$ for (1) and (2)
 - complicated mixture of $\chi_0^2, \chi_1^2, \chi_2^2$ for (3)
- Bootstrap approach
 - Parametric bootstrap: assume all distributions are Gaussian
 - Nonparametric bootstrap: resample from residuals/predictions?



Back

Close

Moment Adjustments for Nonparametric Bootstrap

- BLUP predictors $\hat{\gamma}$, $\hat{\mathbf{u}}$ and residuals $\hat{\boldsymbol{\varepsilon}}$ do not have same moments as $\boldsymbol{\gamma}$, \mathbf{u} , $\boldsymbol{\varepsilon}$

$$\text{Var}(\hat{\boldsymbol{\gamma}}) = \sigma_{\gamma}^4 \mathbf{Z}' \mathbf{V}^{-1} (\mathbf{I} - \mathbf{Q}) \mathbf{Z}$$

$$\text{Var}(\hat{\mathbf{u}}) = \sigma_u^4 \mathbf{D}' \mathbf{V}^{-1} (\mathbf{I} - \mathbf{Q}) \mathbf{D}$$

$$\text{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma_{\varepsilon}^4 \mathbf{V}^{-1} (\mathbf{I} - \mathbf{Q})$$

with $\mathbf{V} = \text{Var}(\mathbf{Y})$, $\mathbf{Q} =$ regression projection matrix

- In simulations, setting $\mathbf{Q} = \mathbf{0}$ had negligible effect, so that

$$\tilde{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}} (\mathbf{Z}' \mathbf{V}^{-1} \mathbf{Z})^{-1/2} / \sigma_{\gamma}$$

$$\tilde{\mathbf{u}} = \hat{\mathbf{u}} (\mathbf{D}' \mathbf{V}^{-1} \mathbf{D})^{-1/2} / \sigma_u$$

$$\tilde{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}} \text{diag}\{\mathbf{V}^{-1}\}^{-1/2} / \sigma_{\varepsilon}$$



Back

Close

Nonparametric Bootstrap

1. Obtain adjusted predictors/residuals $\tilde{\gamma}, \tilde{\mathbf{u}}, \tilde{\epsilon}$
2. Resample from $\tilde{\gamma}, \tilde{\mathbf{u}}, \tilde{\epsilon}$ with replacement and construct bootstrap data

$$\mathbf{Y}^* = \mathbf{X}^F \hat{\boldsymbol{\beta}}^F + \mathbf{Z} \hat{\boldsymbol{\gamma}}^* + \mathbf{D} \hat{\mathbf{u}}^* + \boldsymbol{\epsilon}^*$$

leaving out tested variance component(s)

3. Fit small area models under null and alternative to \mathbf{Y}^* , and compute LR statistic
4. p -value for model for \mathbf{Y} obtained from bootstrap distribution



Back

Close

4. Small Area Estimation for NE Lakes

- 551 measurements on 334 lakes
- Dependent variable

ANC	Acid Neutralizing Capacity
-----	----------------------------

- Independent variables
 - Fixed effects

INT	intercept
ELEV	elevation

- Random effects

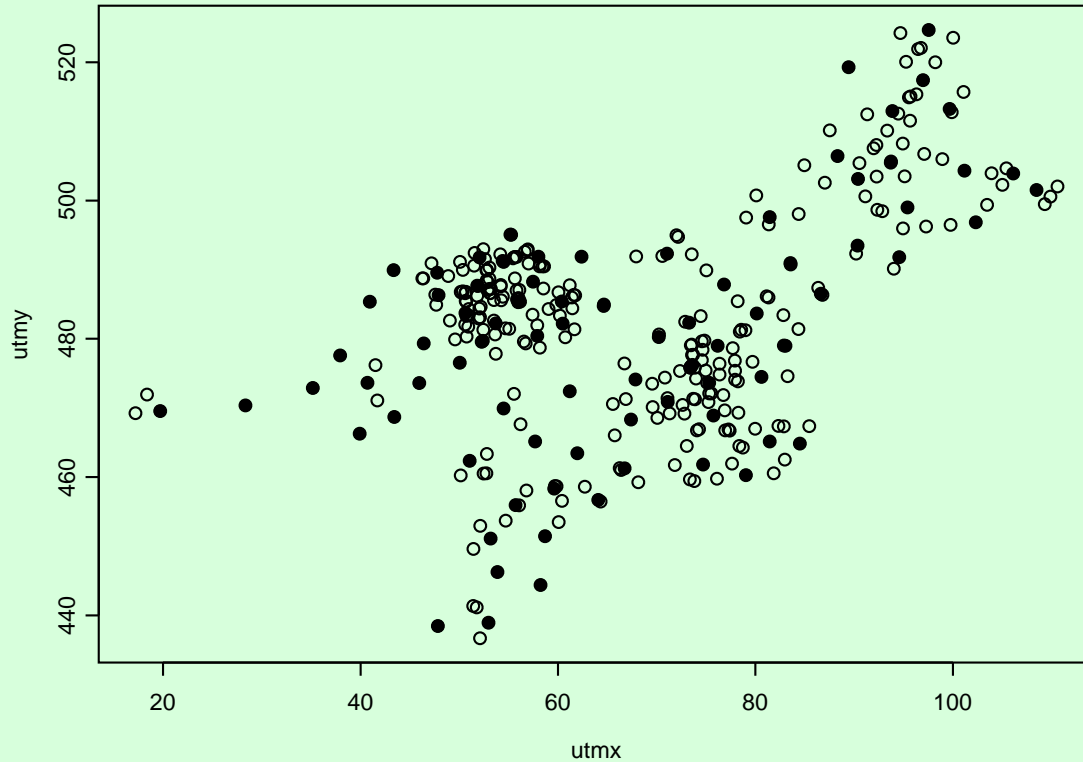
TPS	spatial thin-plate spline for $K = 80$
HUC	113 small areas



Back

Close

Knot Locations for Spatial Spline



Back

Close

Full Model: Model Fit

- Estimates
 - Fixed effects

	$\hat{\beta}^F$	p -value
INT	228.6	0.96
ELEV	-0.814	<0.001

- Random effects

	$\hat{\sigma}$	p -value
TPS	71.2	<0.001
HUC	365.7	<0.001
Error	179.5	<0.001

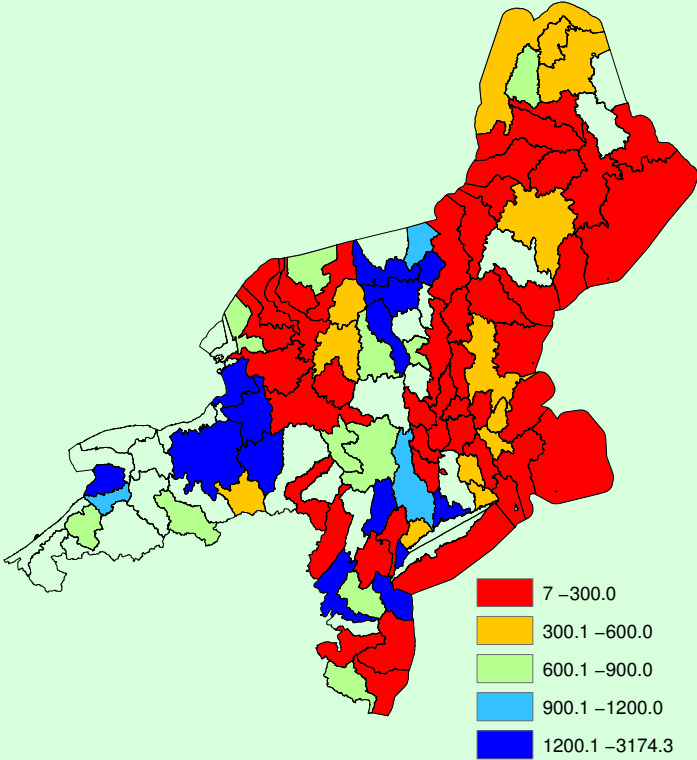
- p -values obtained by nonparametric bootstrap with 1,000 replicates



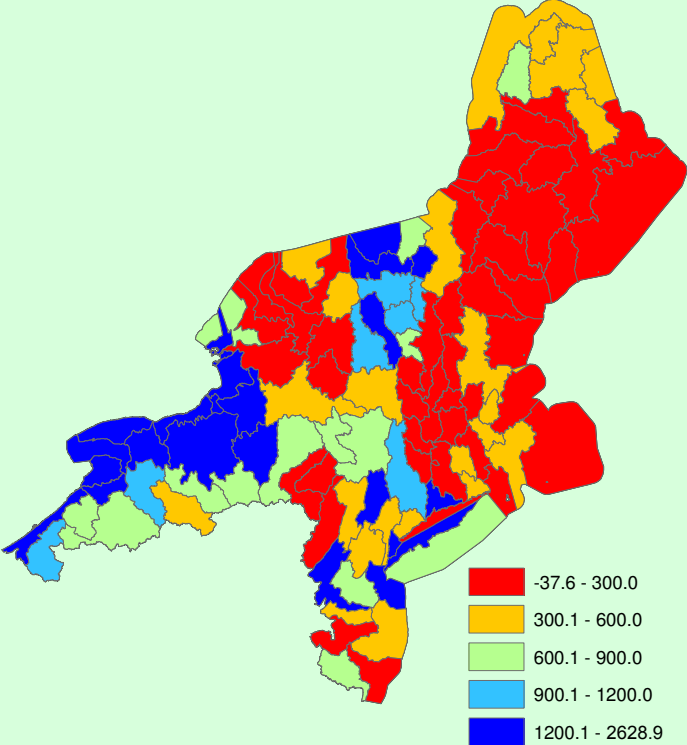
Back

Close

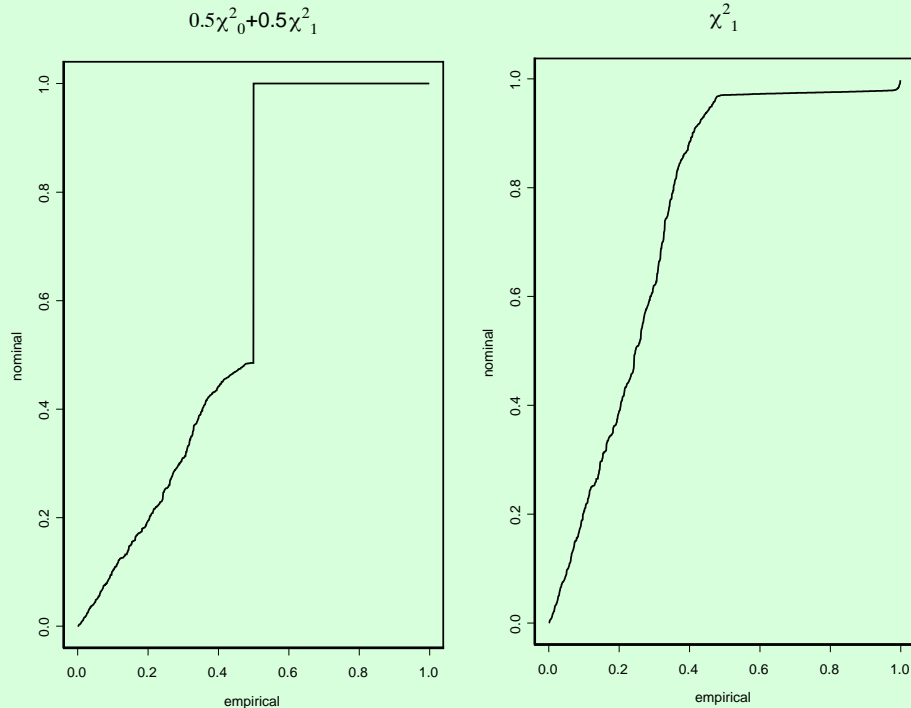
HUC Sample Means



Full Model: HUC Predictions



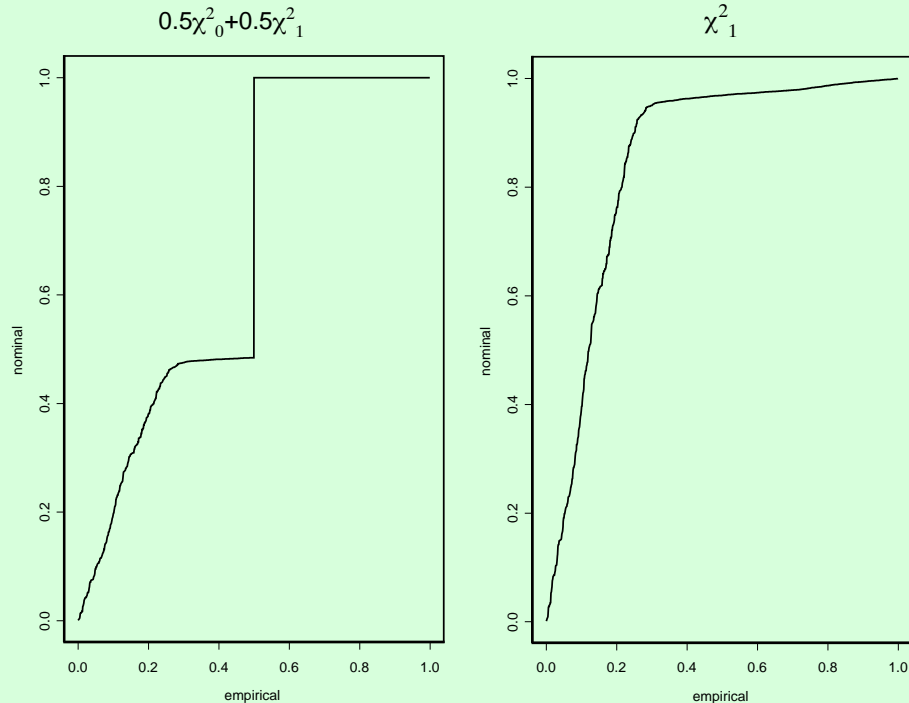
Bootstrap Distribution of Likelihood Ratio Test Statistic for $\sigma_u^2 = 0$



Back

Close

Bootstrap Distribution of Likelihood Ratio Test Statistic for $\sigma_\gamma^2 = 0$



Back

Close

Are both random effects really needed?

- AIC model selection criterion

		HUC	
		yes	no
TPS	yes	7755	7894
	no	7968	8497

- Correlation between small area mean ANC and model predictions

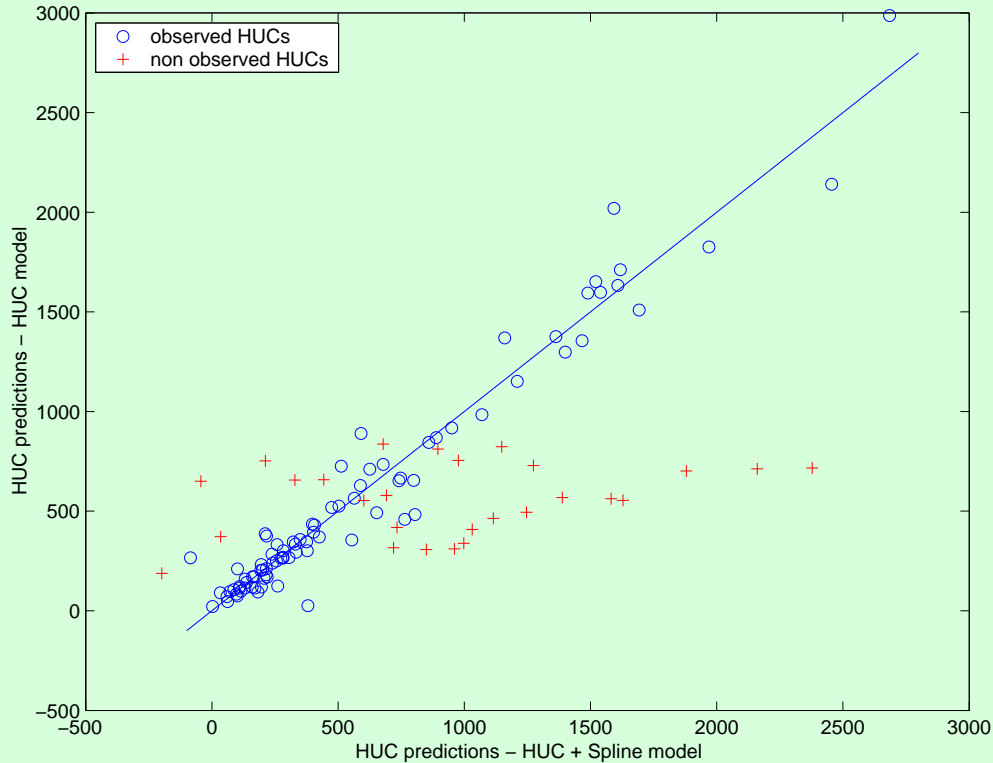
		HUC	
		yes	no
TPS	yes	0.98	0.88
	no	0.99	0.02



Back

Close

Spline in Small Area Model?



Spline model provides better predictions for “empty” HUCs



Back

Close

5. Conclusions

- P-splines incorporate deviations nonparametric mean model into small area estimation through mixed-model formulation
- Fitting with existing mixed model software
- Nonparametric bootstrap corrects for lack of fit of asymptotic distribution
- Northeastern Lakes survey
 - spline mean function improves prediction for HUCs without observations

Contact information:

- jopsomer@iastate.edu
- <http://www.public.iastate.edu/~jopsomer/home.html>



Back

Close