

# Penalized Splines and Small Area Estimation

**Jean Opsomer**  
*Iowa State University*

Joint work with  
Jay Breidt, Colorado State University  
Gerda Claeskens, Université Catholique de Louvain  
Göran Kauermann, Universität Bielefeld  
Giovanna Ranalli, Università di Perugia

July 23, 2004



Back

Close

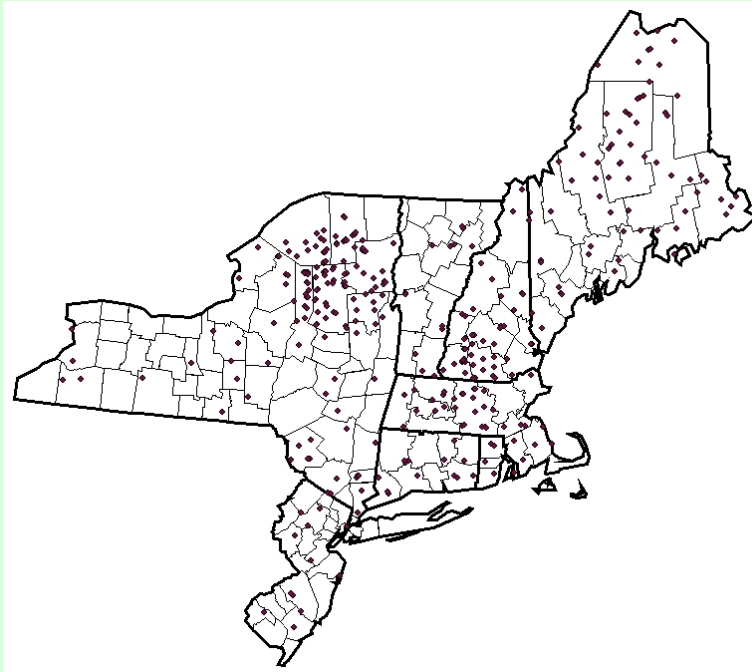
# Outline

1. Introduction
2. Nonparametric regression using penalized splines
3. Small area estimation
4. Nonparametric small area estimation
5. Northeastern lakes survey
6. Conclusion

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

# 1. Introduction: Lakes Survey

Ecological condition survey of Northeastern lakes conducted by U.S. Environmental Protection Agency



Data collected for 338 lakes

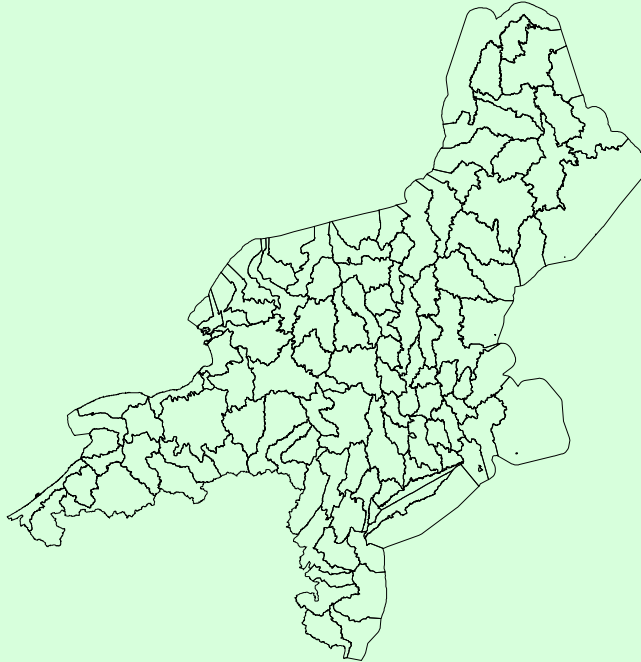


Back

Close

# 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

## 2. Nonparametric Regression Using Penalized Splines

Many nonparametric regression methods are available

- Kernel and local polynomial methods
- Splines
  - smoothing splines
  - regression splines
  - penalized splines (P-splines)
- Orthogonal decomposition (wavelet, Fourier series)

Penalized spline regression (Eilers and Marx, 1996) is simple, flexible and computationally attractive smoothing method



Back

Close

# Definition of Penalized Splines Model

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

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

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

$p$  : 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)

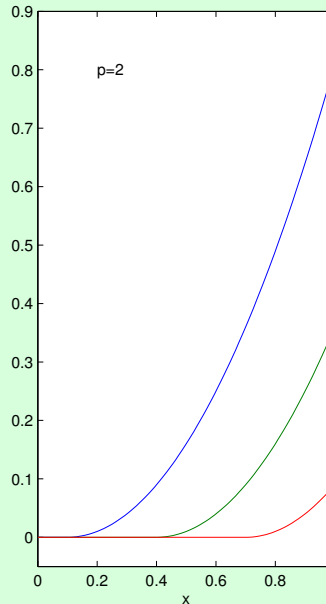
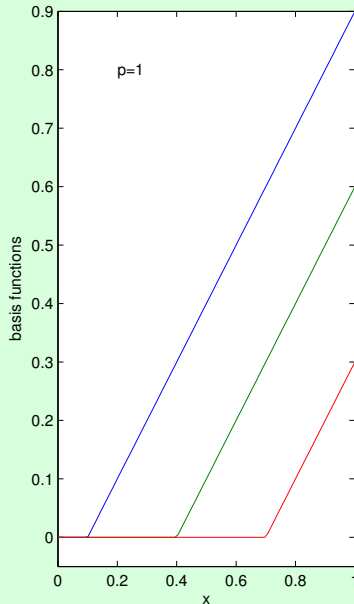


Back

Close

# Polynomial Spline Basis Functions

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

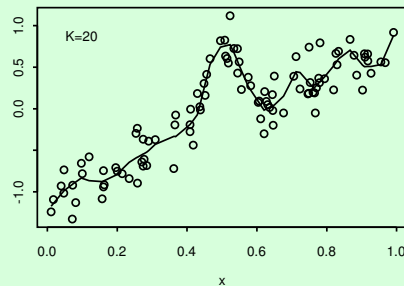
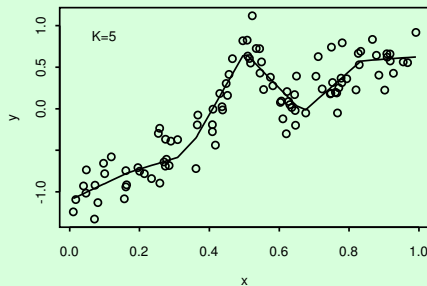
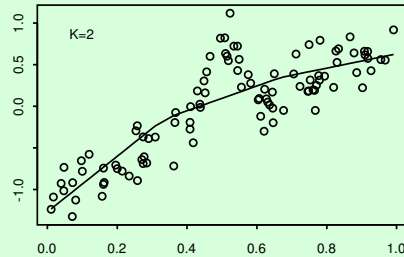
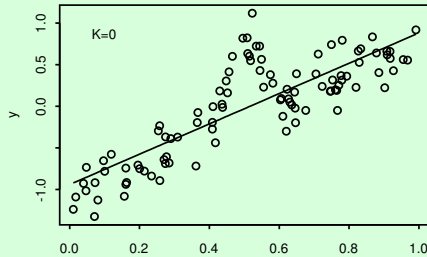


Other spline basis functions are possible (B-splines, radial splines)



# Choosing $K$

If  $K$  is sufficiently large,  $m_K(\cdot)$  can approximate large class of functions



$\Rightarrow$  Rule of thumb:  $K = \min(\#X/4, 35)$  (Ruppert, 2002)



# Expressing Spline Model as Parametric Model

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

with

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



Back

Close

# Fitting by Penalized Splines Regression

Minimize *penalized* sum of squares

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - m_K(x_i; \boldsymbol{\beta}))^2 + \lambda \sum_{k=1}^K \beta_{p+k}^2$$

$$\Rightarrow \hat{m}_{K,\lambda}(x) = \mathbf{x}^* \hat{\boldsymbol{\beta}}_{\lambda} = \mathbf{x}^* (\mathbf{X}^{*T} \mathbf{X}^* + \lambda \mathbf{A})^{-1} \mathbf{X}^{*T} \mathbf{Y}$$

$\lambda$  = smoothing penalty (fixed)

$\mathbf{A}$  =  $\text{diag}\{0, \dots, 0, 1, \dots, 1\}$

$\mathbf{X}^*$  = design matrix (including spline terms)

$\hat{\boldsymbol{\beta}}_{\lambda}$  is *ridge regression* estimator, with ridge penalty on nonlinear (spline) terms of model

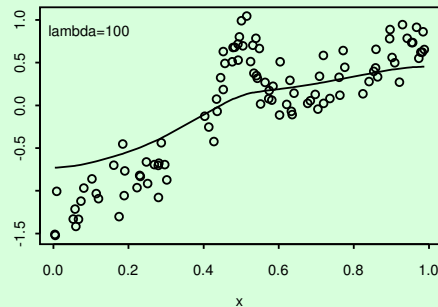
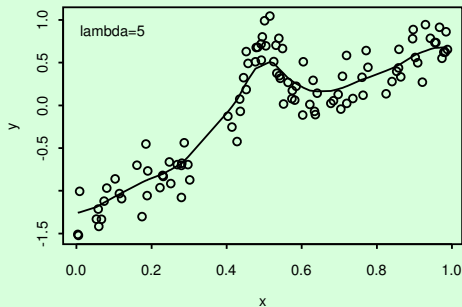
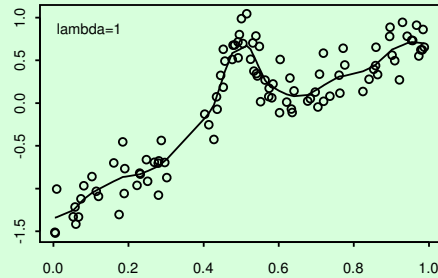
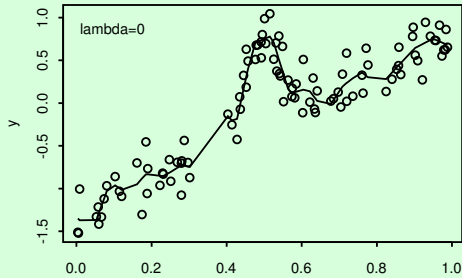


Back

Close

# Fitting by Penalized Splines Regression (2)

$\lambda$  protects against overfitting and determines smoothness of fit



Back

Close

# Choosing the Penalty $\lambda$

- Cross-Validation: minimize CV sum of squares with respect to  $\lambda$
- Mixed model approach: treat spline parameters  $\beta_{p+1}, \dots, \beta_{p+K}$  as a random effect with common variance  $\sigma_{\beta}^2$  and fit regression using Maximum Likelihood approach (MLE, REML)



Back

Close

# P-spline: “Hybrid” Regression Method

- P-spline is a nonparametric regression method:
  - can fit very large classes of functions
  - adaptive to local features in the data
  - smoothness of function is determined by penalty parameter  $\lambda$

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

# P-spline: “Hybrid” Regression Method

- P-spline is a nonparametric regression method:
  - can fit very large classes of functions
  - adaptive to local features in the data
  - smoothness of function is determined by penalty parameter  $\lambda$
- P-spline is a parametric regression method:
  - model can be written as  $x^* \beta$
  - fitted by (global) least squares method
  - number of parameters  $p + K$  puts upper bound on flexibility of model



Back

Close

# P-Splines or Local Polynomials?

## Advantages of P-Splines

- Closely related to parametric modelling
- Model is easy to extend to multivariate, additive, semiparametric cases
- Handles data sparseness easily, very fast to compute
- Fits “look” better

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

# P-Splines or Local Polynomials?

## Advantages of P-Splines

- Closely related to parametric modelling
- Model is easy to extend to multivariate, additive, semiparametric cases
- Handles data sparseness easily, very fast to compute
- Fits “look” better

## Disadvantages of P-Splines

- Flexibility of model limited by number of parameters  $p + K$
- No “true” asymptotic theory



Back

Close

# Extending the model

- Semi-parametric regression

$$\text{Model } y_i = m(x_{1i}; \boldsymbol{\beta}_1) + \mathbf{x}_{2i}\boldsymbol{\beta}_2 + \varepsilon_i$$

$$\sum_{i=1}^n (y_i - m_K(x_{1i}; \boldsymbol{\beta}_1) - \mathbf{x}_{2i}\boldsymbol{\beta}_2)^2 + \lambda \sum_{k=1}^K \beta_{1,p+k}^2$$



Back

Close

# Extending the model

- Semi-parametric regression

$$\text{Model } y_i = m(x_{1i}; \boldsymbol{\beta}_1) + \mathbf{x}_{2i}\boldsymbol{\beta}_2 + \varepsilon_i$$

$$\sum_{i=1}^n (y_i - m_K(x_{1i}; \boldsymbol{\beta}_1) - \mathbf{x}_{2i}\boldsymbol{\beta}_2)^2 + \lambda \sum_{k=1}^K \beta_{1,p+k}^2$$

- Additive model

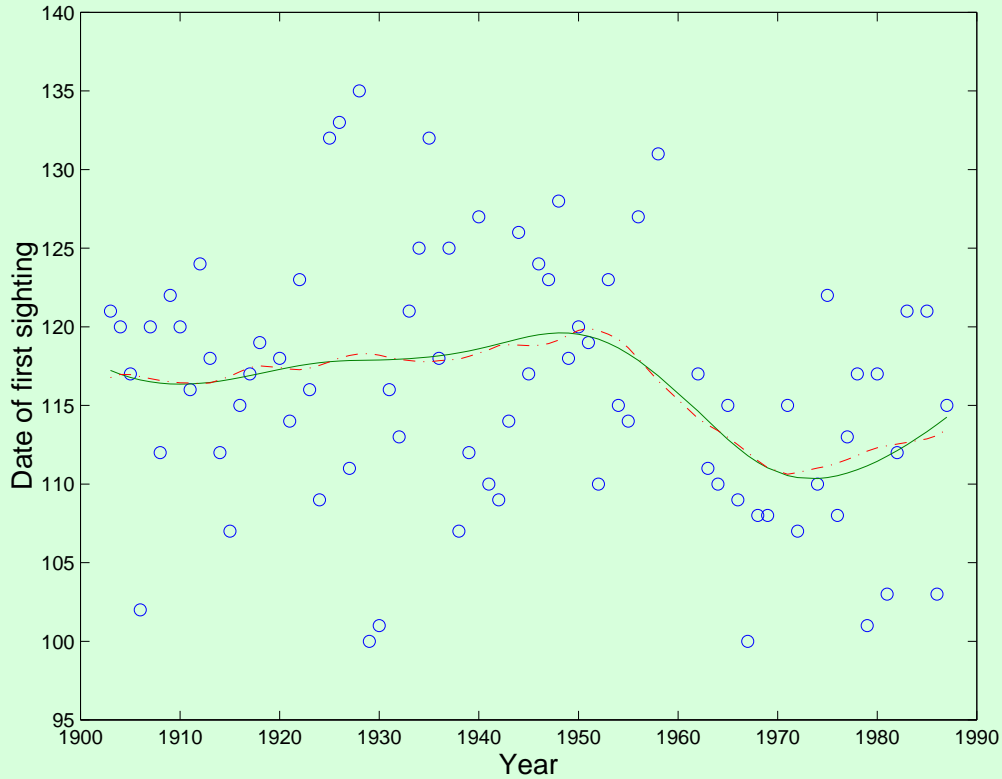
$$\text{Model } y_i = m_1(x_{1i}; \boldsymbol{\beta}_1) + m_2(x_{2i}; \boldsymbol{\beta}_2) + \varepsilon_i$$

$$\sum_{i=1}^n (y_i - m_{1,K}(x_{1i}; \boldsymbol{\beta}_1) - m_{2,K}(x_{2i}; \boldsymbol{\beta}_2))^2 + \lambda_1 \sum_{k=1}^K \beta_{1,p+k}^2 + \lambda_2 \sum_{k=1}^K \beta_{2,p+k}^2$$

- Other...



# Fits Often “Look” Better



Note: this is subjective...



# Theory for P-spline Regression?

$$\text{MSE} = E(\hat{m}_{K,\lambda}(x) - m(x))^2$$

- Regression splines ( $\lambda = 0, K \rightarrow \infty$ ):

$$\text{MSE} = O\left(K^{-2p} + \frac{1}{nK^{-1}}\right)$$

(Huang, 2001)

- Smoothing splines ( $\lambda \rightarrow 0, K = n$ ):

$$\text{MSE} = O\left(\lambda + \frac{1}{n\lambda^{1/2(p+1)}}\right)$$

(Cox, 1983)

- Wand (1999): asymptotic approximation to P-spline MSE for  $K$  fixed and  $\lambda \rightarrow 0$



Back

Close

# Theory for P-spline Regression (2)

Hall and Opsomer (2004): white-noise model ( $K = \infty$ )

- Penalized least squares criterion

$$\min_{\beta(\cdot)} \int \left\{ y_t - \int \beta(s) \phi(t | s) \rho(s) ds \right\}^2 f(t) dt + \lambda \int \beta(t)^2 dt$$

with  $\phi(t | s) = (t - s)_+^p$ , and  $\rho(\cdot)$  the density of the “knots”

- Estimator

$$\hat{m}(t) = \int \hat{\beta}(s) \phi(t | s) \rho(s) ds$$

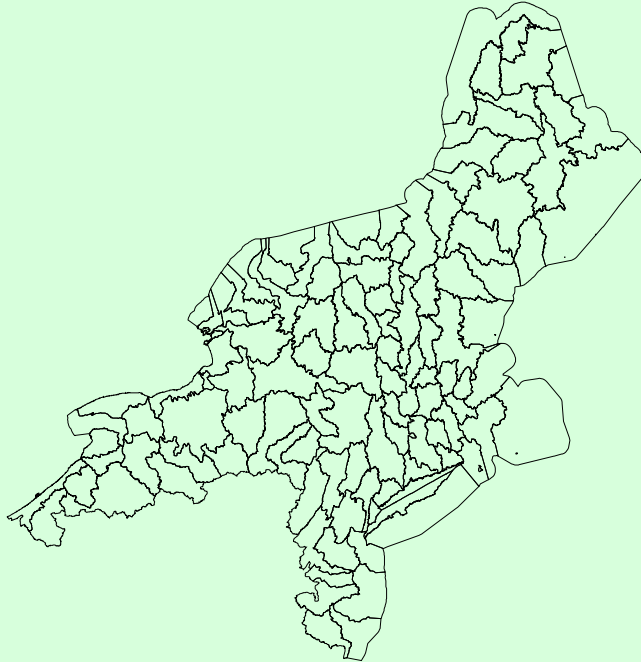
- Mean squared error

$$\text{MSE} = O \left( \lambda + \frac{1}{n \lambda^{1/2(p+1)}} \right)$$

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

### 3. Small Area Estimation

Data contain 557 observations over 113 HUCs



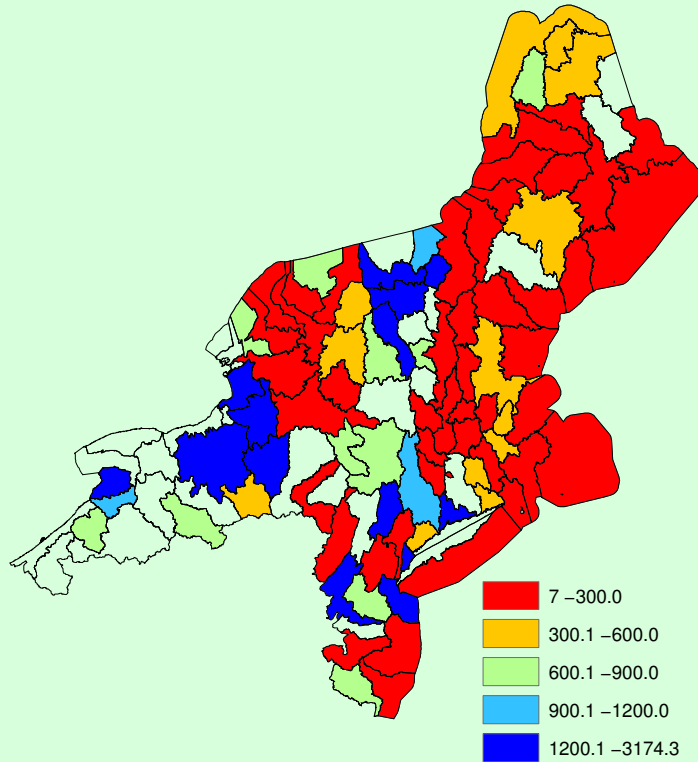
Goal: produce estimates of ANC for all HUCs



Back

Close

# HUC Sample Means



Problems: unreliable estimates, missing HUCs



Back

Close

# HUC 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

⇒ Modelling required to construct reliable HUC-level estimates

- model combines overall trend for region with random effect for small areas
- mixed model/prediction
- called *small area estimation* in surveys statistics



Back

Close

# Small Area Estimation as Mixed Model Regression

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

- Population of interest  $U$ , divided into small areas  $U_t, t = 1, \dots, T$
- Variable of interest  $y_i$  observed on sample,  $i \in s$
- Auxiliary variable  $x_i$  observed on sample,  $i \in s$ , with known small area means,  $\bar{x}_t = \sum_{i \in U_t} x_i / N_t$
- Assume linear relationship between  $y_i$  and  $x_i$  in population, with random effect  $u_t$  for small areas  $U_t, t = 1, \dots, T$

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

# Small Area Estimation Regression Model

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

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



Back

Close

# Small Area Estimation Regression Model

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

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

$$\begin{aligned}\mathbf{u} &= (u_1, \dots, u_T) \sim \text{iid } \mathcal{F}_u(0, \sigma_u^2) \\ \varepsilon_i &\sim \mathcal{F}_\varepsilon(0, \sigma_\varepsilon^2)\end{aligned}$$

In matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{u} + \boldsymbol{\varepsilon}$$



Back

Close

# Small Area Estimation: BLUP

- Small area estimation goal: predict

$$\bar{y}_t = \bar{\mathbf{x}}_t \boldsymbol{\beta} + u_t \quad t = 1, \dots, T$$

- Assuming  $\sigma_u^2, \sigma_\varepsilon^2$  known, Best Linear Unbiased Predictor (BLUP) of  $\bar{y}_t$  is

$$\hat{y}_t = \bar{\mathbf{x}}_t \hat{\boldsymbol{\beta}} + \hat{u}_t$$

with

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

$$\mathbf{V} = \text{Var}(\mathbf{Y}) = \sigma_\varepsilon^2 \mathbf{I}_n + \sigma_u^2 \mathbf{D}'\mathbf{D}$$

$$\hat{\mathbf{u}} = \sigma_u^2 \mathbf{D}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

(McCulloch and Searle, 2001)



Back

Close

# BLUP as Ridge Regression Estimator

$$\hat{y}_t = \bar{x}_t \hat{\beta} + \hat{u}_t$$

with

$$\begin{aligned} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} &= \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{D} \\ \mathbf{D}'\mathbf{X} & \mathbf{D}'\mathbf{D} + \frac{\sigma_\varepsilon^2}{\sigma_u^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{D}'\mathbf{Y} \end{bmatrix} \\ &= \left( \mathbf{X}^*{}'\mathbf{X}^* + \frac{\sigma_\varepsilon^2}{\sigma_u^2} \mathbf{A} \right)^{-1} \mathbf{X}^*{}'\mathbf{Y} \end{aligned}$$

where

$$\mathbf{X}^* = [\mathbf{X} \mathbf{D}]$$



Back

Close

# Small Area Estimation: EBLUP

When  $\sigma_u^2, \sigma_\varepsilon^2$  unknown, Empirical BLUP (EBLUP) of  $\bar{y}_t$  is found by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML)

$$\hat{y}_t = \bar{\mathbf{x}}_t \hat{\boldsymbol{\beta}} + \hat{u}_t$$

with

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{Y}$$

$$\hat{\mathbf{V}} = \hat{\sigma}_\varepsilon^2 \mathbf{I}_n + \hat{\sigma}_u^2 \mathbf{D}' \mathbf{D}$$

$$\hat{\mathbf{u}} = \hat{\sigma}_u^2 \mathbf{D} \hat{\mathbf{V}}^{-1} (\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}})$$

(e.g. Searle, Casella and McCulloch, 1992)



Back

Close

# Inference for Small Area Estimation

Target: prediction MSE =  $E(\hat{y}_t - \bar{y}_t)^2$

Asymptotic approximation of EBLUP available under linear mixed model specification

Can be estimated consistently



Back

Close

## 4. Nonparametric Small Area Estimation

- more flexible fixed component in model can improve prediction
- predicting for “empty” (no data) HUCs relies exclusively on fixed component



Back

Close

## 4. Nonparametric Small Area Estimation

- more flexible fixed component in model can improve prediction
- predicting for “empty” (no data) HUCs relies exclusively on fixed component

P-splines are ideally suited for small area estimation

- close relationship between “classical” small area estimation models and P-splines
- availability of existing software
- ability to evaluate need for nonlinearity in model and significance of small area effects



Back

Close

# P-splines as Random Effects

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

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

# P-splines as Random Effects

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

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

$$\begin{aligned}\mathbf{z}_i \boldsymbol{\gamma} &= z_{1i} \gamma_1 + \dots + z_{Ki} \gamma_K \\ &\equiv (x_i - \kappa_1)_+^p \beta_{p+1} + \dots + (x_i - \kappa_K)_+^p \beta_{p+K}\end{aligned}$$



Back

Close

# P-splines as Random Effects

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

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

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

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



Back

Close

# P-splines Estimator as BLUP

Assuming  $\sigma_\gamma^2, \sigma_\varepsilon^2$  known, BLUP/BLUE for 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)

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

with

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

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}^F \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \text{P-splines (ridge) regression estimator } \hat{\boldsymbol{\beta}}_\lambda \text{ with } \lambda = \sigma_\varepsilon^2 / \sigma_\gamma^2$$



Back

Close

# P-splines Estimator as EBLUP

If  $\sigma_\gamma^2, \sigma_\varepsilon^2$  are unknown, estimates can be obtained by ML/REML

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

$\Rightarrow \hat{\beta}_{\hat{\lambda}}$  is *Empirical BLUP* (EBLUP) for  $\beta$



Back

Close

# P-splines Estimator as EBLUP

If  $\sigma_\gamma^2, \sigma_\varepsilon^2$  are unknown, estimates can be obtained by ML/REML

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

$\Rightarrow \hat{\beta}_{\hat{\lambda}}$  is *Empirical BLUP* (EBLUP) for  $\beta$

- Smoothing penalty  $\lambda = \hat{\sigma}_\varepsilon^2 / \hat{\sigma}_\gamma^2$  is determined by data
- Automatically adjusts  $\lambda$  to “patterns” in data
  - small deviations from parametric shape  $\rightarrow \hat{\sigma}_\gamma^2$  small  $\rightarrow$  more smoothing
  - data exhibit significant deviations from parametric shape  $\rightarrow \hat{\sigma}_\gamma^2$  large  $\rightarrow$  less smoothing

(Wand, 2003)



Back

Close

# Nonparametric Small Area Model

Combine both random effects models

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

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

# Nonparametric Small Area Model

Combine both random effects models

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

Variance components

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

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

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

EBLUP can be computed by (RE)ML, and

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



Back

Close

# Inference for Nonparametric Small Area Estimation

- What is right target?

1. full prediction MSE:  $E(\hat{y}_t - \bar{y}_t)^2$

2. full ridge regression:  $E(\hat{y}_t - \bar{y}_t | \boldsymbol{\gamma}, \mathbf{u})^2$

3. prediction MSE conditional on spline:  $E(\hat{y}_t - \bar{y}_t | \boldsymbol{\gamma})^2$

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

# Inference for Nonparametric Small Area Estimation

- What is right target?
  1. full prediction MSE:  $E(\hat{y}_t - \bar{y}_t)^2$
  2. full ridge regression:  $E(\hat{y}_t - \bar{y}_t | \boldsymbol{\gamma}, \mathbf{u})^2$
  3. prediction MSE conditional on spline:  $E(\hat{y}_t - \bar{y}_t | \boldsymbol{\gamma})^2$
- No clear winner:
  1. spline mean function is fixed, not random
  2. small areas too numerous to treat as fixed
  3. complicated (?)
- Asymptotic approximation can be derived for all 3 under mixed model specification



Back

Close

# Inference for Nonparametric Small Area Estimation (2)

- Inference about variance components
  1.  $H_0 : \sigma_\gamma^2 = 0$  versus  $H_a : \sigma_\gamma^2 > 0$
  2.  $H_0 : \sigma_u^2 = 0$  versus  $H_a : \sigma_u^2 > 0$
  3.  $H_0 : \sigma_\gamma^2 = \sigma_u^2 = 0$  versus  $H_a : \sigma_\gamma^2 > 0$  or  $\sigma_u^2 > 0$
- Existing results:
  - asymptotic distribution of likelihood ratio for parameter on boundary: Self and Liang (1987)
  - exact distribution for P-splines: Crainiceanu and Ruppert (2003)

⇒ neither one applies here...
- Develop parametric bootstrap approach



Back

Close

# Spatial Smoothing using P-splines

- NE Lakes auxiliary variable is location: requires 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

## 5. Small Area Estimation for Lakes Survey

- 557 measurements on 338 lakes
- Dependent variable

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

- Independent variables
  - Fixed effects

INT	intercept
ELEV	elevation

- Random effects

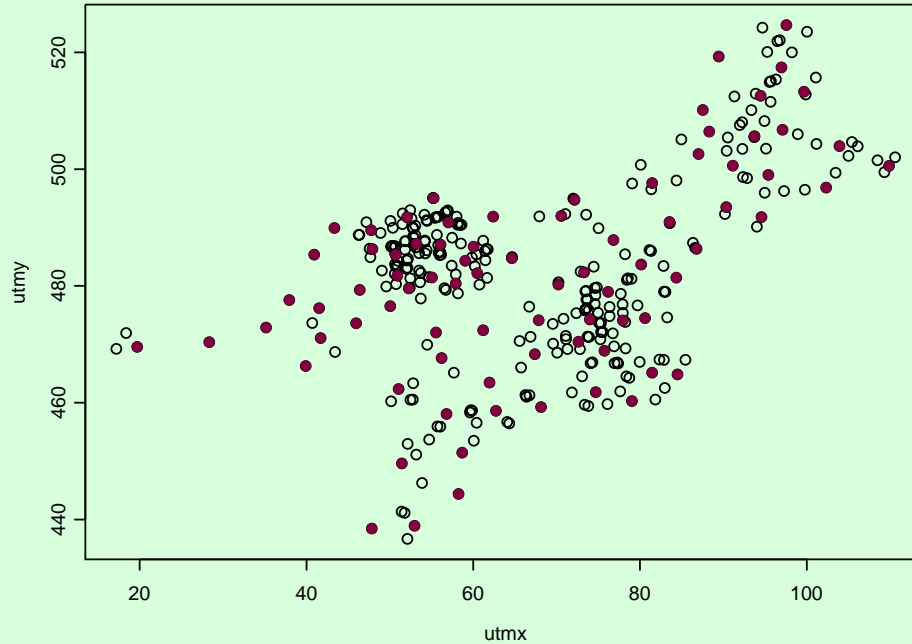
TPS	spatial thin-plate spline for $K = 81$
HUC	113 small areas



Back

Close

# Knot Locations for Spatial Spline



Algorithm: `funfits()` in R/S-Plus



# Full Model: Model Fit

- Estimates
  - Fixed effects

	$\hat{\beta}^F$
INT	586
ELEV	-0.74

- Random effects

	$\hat{\sigma}$
TPS	79
HUC	420
Error	173

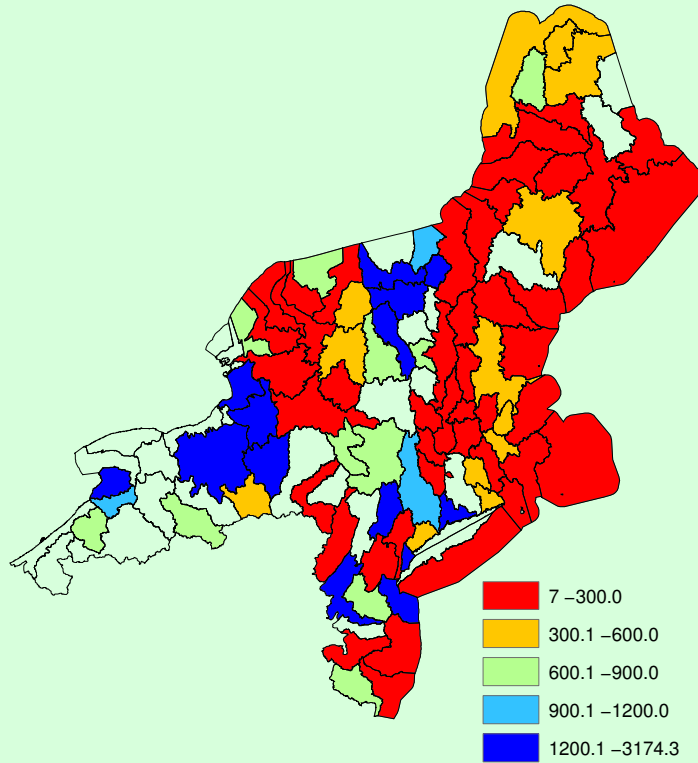
- Correlation between model predictions and ANC: 0.96



Back

Close

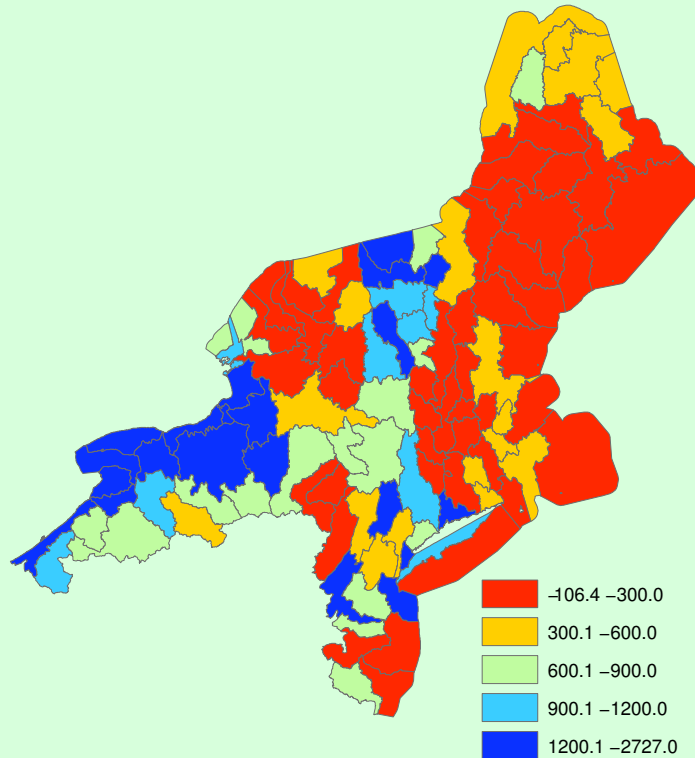
# HUC Sample Means



Back

Close

# Full Model: HUC Predictions



Correlation between HUC means and model predictions: 0.97



# Are both random effects needed?

- AIC model selection criterion

		HUC	
		yes	no
TPS	yes	7755	7933
	no	7968	8497

- Correlation between ANC and model prediction

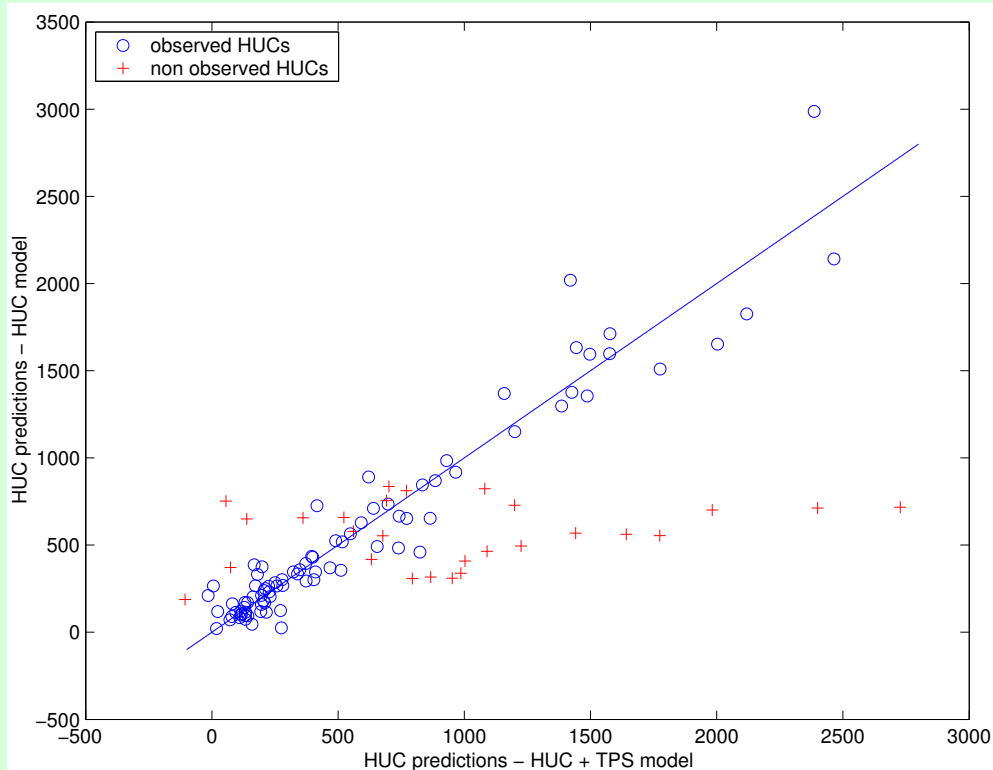
		HUC	
		yes	no
TPS	yes	0.96	0.90
	no	0.90	0.19



Back

Close

# Spline in Small Area Model?



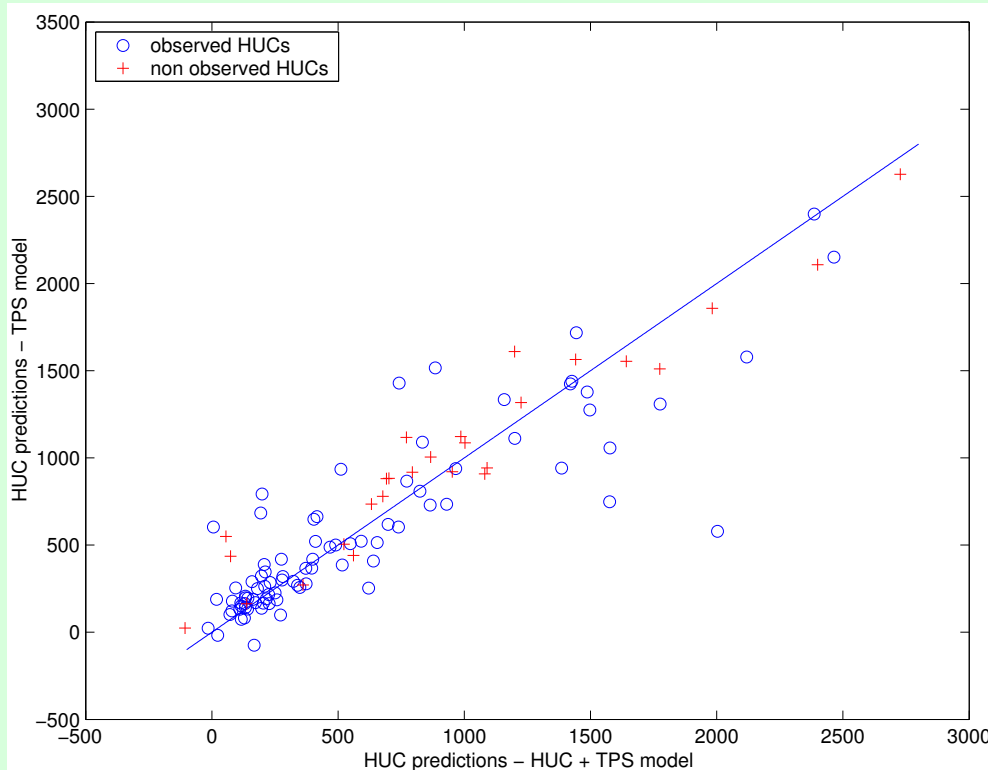
Spline model provides better predictions for “empty” HUCs



Back

Close

# Small Area Random Effect in Spatial Model?



HUC effect results in predictions that are closer to observed data



Back

Close

## 6. Conclusions

- P-spline regression is promising and flexible new tool in smoothing applications
  - full theoretical development still lacking
- Mixed model formulation allows easy incorporation into existing small area estimation techniques
- To do:
  - tests for significance of random effects

### Contact information:

- [jopsomer@iastate.edu](mailto:jopsomer@iastate.edu)
- <http://www.public.iastate.edu/~jopsomer/home.html>



Back

Close