Nonparametric Small Area Estimation for the Northeastern Lakes Survey

Jean Opsomer
Iowa State University

Jay Breidt and Siobhan Everson-Stewart
Colorado State University

August 11, 2003
Outline

1. Introduction

2. Nonparametric Regression Using Penalized Splines

3. Nonparametric Small Area Estimation
   (a) Penalized splines as mixed model regression
   (b) Small area estimation as mixed model regression
   (c) Nonparametric Small Area Estimation

4. Northeastern Lakes Model Fitting

5. Conclusions
1. Intro: Northeastern Lakes Survey

Ecological condition survey of 338 lakes in Northeastern US

Survey conducted by EPA as part of EMAP pilot project (1991-1996)
Intro: Northeastern Lakes Survey (2)

Region includes 21,384 lakes and 113 8-digit HUCs

Goal of research: produce small area estimates for HUCs
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 observation

⇒ Modelling required to construct reliable HUC-level estimates
HUC as Small Areas (2)

⇒ Predicting for “empty” HUCs relies exclusively on mean model

Penalized splines small area estimation is promising technique for this type of estimation problems
2. Nonparametric Regression Using Penalized Splines

Many nonparametric regression methods are available

- Kernel and local polynomial methods
- Smoothing splines, regression splines, ...
- Orthogonal decomposition (wavelet, Fourier series)

 Penalized splines or $P$-splines regression (Eilers and Marx, 1996) is simple, flexible and computationally attractive smoothing method
Definition of Penalized Splines Model

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

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

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

- \( p \): degree of spline (fixed)
- \( \kappa_1 < \ldots < \kappa_K \): set of \( K \) knots (fixed)
- \( \beta = (\beta_0, \ldots, \beta_{p+K}) \): vector of parameters (unknown)

(Ruppert, Wand and Carroll, 2003)
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)
Choosing $K$

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

$\Rightarrow$ Rule of thumb: $K = \min(\#X/4, 35)$
Expressing Spline Model as Parametric Model

\[ m(x; \beta) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k}(x - \kappa_k)^+_p \]

\[ \equiv x^* \beta \]

with

\[ x^* = (1, x, \ldots, x^p, (x - \kappa_1)^+_p, \ldots, (x - \kappa_K)^+_p) \]

\[ \beta = (\beta_0, \ldots, \beta_{p+K})^T \]
Fitting by Penalized Splines Regression

Minimize penalized sum of squares

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

$$\Rightarrow \hat{\beta}_\lambda = (X^*^T X^* + \lambda D)^{-1} X^*^T Y$$

$$\lambda = \text{smoothing penalty (fixed)}$$

$$D = \text{diag}\{0, \ldots, 0, 1, \ldots, 1\}$$

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

$$\hat{\beta}_\lambda$$ is ridge regression estimator, with ridge penalty on nonlinear (spline) terms of model.
Fitting by Penalized Splines Regression (2)

\( \lambda \) protects against overfitting and determines smoothness of fit
Choosing the Penalty $\lambda$

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

• Semi-parametric regression

Model \( y_i = m(x_{1i}; \beta_1) + x_{2i}\beta_2 + \varepsilon_i \)

\[
\min_{\beta_1, \beta_2} \sum_{i=1}^{n} (y_i - m(x_{1i}; \beta_1) - x_{2i}\beta_2)^2 + \lambda \sum_{k=1}^{K} \beta_{1,p+k}^2
\]
Extending the model

- Semi-parametric regression

Model \( y_i = m(x_{1i}; \beta_1) + x_{2i} \beta_2 + \varepsilon_i \)

\[
\min_{\beta_1, \beta_2} \sum_{i=1}^{n} (y_i - m(x_{1i}; \beta_1) - x_{2i} \beta_2)^2 + \lambda \sum_{k=1}^{K} \beta_{1,p+k}^2
\]

- Additive model

Model \( y_i = m_1(x_{1i}; \beta_1) + m_2(x_{2i}; \beta_2) + \varepsilon_i \)

\[
\min_{\beta_1, \beta_2} \sum_{i=1}^{n} (y_i - m_1(x_{1i}; \beta_1) - m_2(x_{2i}; \beta_2))^2 + \lambda_1 \sum_{k=1}^{K} \beta_{1,p+k}^2 + \lambda_2 \sum_{k=1}^{K} \beta_{2,p+k}^2
\]

- Other...
3. Nonparametric Small Area Estimation

- Part of research on local inference suitable for natural resource surveys
- P-splines are ideally suited for small area estimation when the mean function is difficult to specify a priori
  - 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
3.a) P-splines as Random Effects

\[ y_i = m(x_i; \beta) + \varepsilon_i = x_i^* \beta + \varepsilon_i \]
3.a) P-splines as Random Effects

\[ y_i = m(x_i; \beta) + \varepsilon_i \]
\[ = x_i^* \beta + \varepsilon_i \]
\[ = x_i^F \beta^F + z_i \gamma + \varepsilon_i \]

\( x_i^F \beta^F \equiv \beta_0 + x_i \beta_1 + \ldots + x_i^p \beta_p \) (parametric, fixed component)

\[ z_i \gamma = z_{1i} \gamma_1 + \ldots + z_{Ki} \gamma_K \]
\[ \equiv (x_i - \kappa_1)^p \beta_{p+1} + \ldots + (x_i - \kappa_K)^p \beta_{p+K} \]
3.a) P-splines as Random Effects

\[ y_i = m(x_i; \beta) + \epsilon_i \]
\[ = x_i^* \beta + \epsilon_i \]
\[ = x_i^F \beta^F + z_i \gamma + \epsilon_i \]

\[ x_i^F \beta^F \equiv \beta_0 + x_i \beta_1 + \ldots + x_i^p \beta_p \quad \text{(parametric, fixed component)} \]

\[ z_i \gamma = z_1 i \gamma_1 + \ldots + z_K i \gamma_K \]
\[ \equiv (x_i - \kappa_1)^p \beta_{p+1} + \ldots + (x_i - \kappa_K)^p \beta_{p+K} \]

(deviations from parametric, treated as random effect)

\[ \gamma = (\gamma_1, \ldots, \gamma_K) \sim \text{iid } \mathcal{F}_\gamma(0, \sigma_\gamma^2) \]
\[ \epsilon_i \sim \mathcal{F}_\epsilon(0, \sigma_\epsilon^2) \]
P-splines Estimator as BLUP

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

$$
\min_{\beta^F, \gamma} \sum_{i=1}^{n} (y_i - x_i^F \beta^F + z_i \gamma)^2 + \frac{\sigma^2_\varepsilon}{\sigma^2_\gamma} \sum_{k=1}^{K} \gamma^2_k
$$

(Henderson et al., 1959)

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

with

$$
D = \text{diag}\{0, \ldots, 0, 1, \ldots, 1\}
$$

$$
X^* = [x^F \ z]
$$

$$
\begin{bmatrix}
\hat{\beta}^F \\
\hat{\gamma}
\end{bmatrix} = \text{P-splines (ridge) regression estimator} \ \hat{\beta}_\lambda \ \text{with} \ \lambda = \frac{\sigma^2_\varepsilon}{\sigma^2_\gamma}
$$
P-splines Estimator as EBLUP

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

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

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

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

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

- Smoothing penalty $\lambda = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\gamma^2}$ 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
3.b) 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 \( t = 1, \ldots, T \)
- Model

\[
y_i = x_i \beta + u_t + \varepsilon_i = x_i \beta + d_i u + \varepsilon_i
\]

\[
d_i = (d_{i1}, \ldots, d_{iT}) \quad d_{it} = \begin{cases} 1 & \text{if } i \in \text{small area } t \\ 0 & \text{otherwise} \end{cases}
\]

\[
u = (u_1, \ldots, u_T) \sim \text{iid } F_u(0, \sigma_u^2)
\]

\[
\varepsilon_i \sim F_\varepsilon(0, \sigma_\varepsilon^2)
\]

- Target \( \bar{Y}_t = \bar{X}_t \beta + u_t \) estimated by (E)BLUP \( \bar{y}_t = \bar{X}_t \hat{\beta} + \hat{u}_t \)
3.c) Nonparametric Small Area Model

Combine both random effects models

\[ y_i = m(x_i; \beta) + d_i u + \varepsilon_i \]
\[ = x_i^F \beta^F + z_i \gamma + d_i u + \varepsilon_i \]

Variance components

\[ \gamma \sim \text{iid } \mathcal{F}_\gamma(0, \sigma^2_\gamma) \]
\[ u \sim \text{iid } \mathcal{F}_u(0, \sigma^2_u) \]
\[ \varepsilon_i \sim \mathcal{F}_\varepsilon(0, \sigma^2_\varepsilon) \]
3.c) Nonparametric Small Area Model

Combine both random effects models

\[ y_i = m(x_i; \beta) + d_i u + \varepsilon_i \]
\[ = x_i^F \beta^F + z_i \gamma + d_i u + \varepsilon_i \]

Variance components

\[ \gamma \sim \text{iid } F_\gamma(0, \sigma^2_\gamma) \]
\[ u \sim \text{iid } F_u(0, \sigma^2_u) \]
\[ \varepsilon_i \sim F_\varepsilon(0, \sigma^2_\varepsilon) \]

EBLUP can be computed by REML, and

\[ \bar{y}_t = \bar{X}_t^F \hat{\beta}^F + \bar{Z}_t \hat{\gamma} + \hat{u}_t \]
Fitting Nonparametric Small Area Model

Fitting procedure implemented in SAS

- PROC TRANSREG: construct basis functions $x_i, \ldots, x_i^p$, $(x_i - \kappa_1)_+^p, \ldots, (x_i - \kappa_K)_+^p$ for sample and for population
- PROC MIXED: compute parameter estimates $\hat{\beta}^F, \hat{\gamma}, \hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2$ and predictions $\hat{\gamma}, \hat{u}$
- PROC IML: combine estimators/predictions into EBLUP

$$\bar{y}_t = \bar{X}_t^F \hat{\beta}^F + \bar{Z}_t \hat{\gamma} + \hat{u}_t$$
Fitting Nonparametric Small Area Model

Fitting procedure implemented in SAS

- PROC TRANSREG: construct basis functions $x_i, \ldots, x_i^p$, $(x_i - \kappa_1)_+^p, \ldots, (x_i - \kappa_K)_+^p$ for sample and for population
- PROC MIXED: compute parameter estimates $\hat{\beta}^F, \hat{\sigma}_\gamma^2, \hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2$ and predictions $\hat{\gamma}, \hat{u}$
- PROC IML: combine estimators/predictions into EBLUP

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

But:

- Testing for departure from linearity in mean $\rightarrow$ testing $\sigma_\gamma^2 = 0$
- Testing for significance of small areas $\rightarrow$ testing $\sigma_u^2 = 0$
- Variance estimation
4. Small Area Estimation for NE Lakes

- 557 measurements on 338 lakes
- Dependent variable
  - ANC
  - Acid Neutralizing Capacity
- Independent variables
  - Fixed effects
    - INT intercept
    - ECO 11 “eco-regions”
    - ELEV elevation
    - LONG longitude
    - LAT latitude linear component
  - Random effects
    - LATSPL LAT nonlinear components for $K = 20, p = 1$
    - HUC 113 small areas
Full Model: Model Fit

• Significant variables
  – Fixed effects
    
    |     | $\hat{\beta}^F$ |
    |-----|----------------|
    | ECO 4 | 1340           |
    | ECO 5 | 1324           |
    | ECO 10| 1377           |
    | ELEV  | -0.61          |
    | LAT   | 1028           |
  – Random effects
    LATSPL and HUC (but: formal test?)

• Correlation between model predictions and ANC: 0.93

• Above model selected as best fit by AIC, BIC criterion
Full Model: HUC Predictions

Correlation between HUC means and model predictions: 0.97
Usefulness of Splines in Small Area Model

- AIC model selection criterion

<table>
<thead>
<tr>
<th>LATSPL</th>
<th>HUC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yes</td>
<td>7790</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>7817</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>8153</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>8120</td>
</tr>
</tbody>
</table>

- Correlation between ANC and model prediction

<table>
<thead>
<tr>
<th>LATSPL</th>
<th>HUC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yes</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>0.60</td>
</tr>
</tbody>
</table>
HUCs with No Observations

HUC random effect cannot improve prediction in “empty” HUCs, but spline random effect can.
HUC Predictions, No Random Effects

Correlation between HUC means and model predictions: 0.54
HUC Predictions, Spline Random Effect

Correlation between HUC means and model predictions: 0.61
5. Conclusions

- Development of nonparametric small area estimation method that extends existing mixed model approach
- For ANC prediction: spline small area model appears to provide a minor overall improvement over linear small area model
- Potential improvement for “empty” HUCs
- To do:
  - variance estimation
  - formal test for significance of random effects
The End

Contact information:
- jopsomer@iastate.edu
- http://www.public.iastate.edu/~jopsomer/home.html

Disclaimer:
The research described in this presentation has been funded by the U.S. Environmental Protection Agency through the STAR Cooperative Agreement CR82-9096-01 National Research Program on Design-Based/Model-Assisted Survey Methodology for Aquatic Resources at Oregon State University. It has not been subjected to the Agency’s review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.