



Extensions of Penalized Spline Regression for Natural Resource Monitoring Applications: Small Area Estimation of Soil Profiles

F. Jay Breidt
Colorado State University

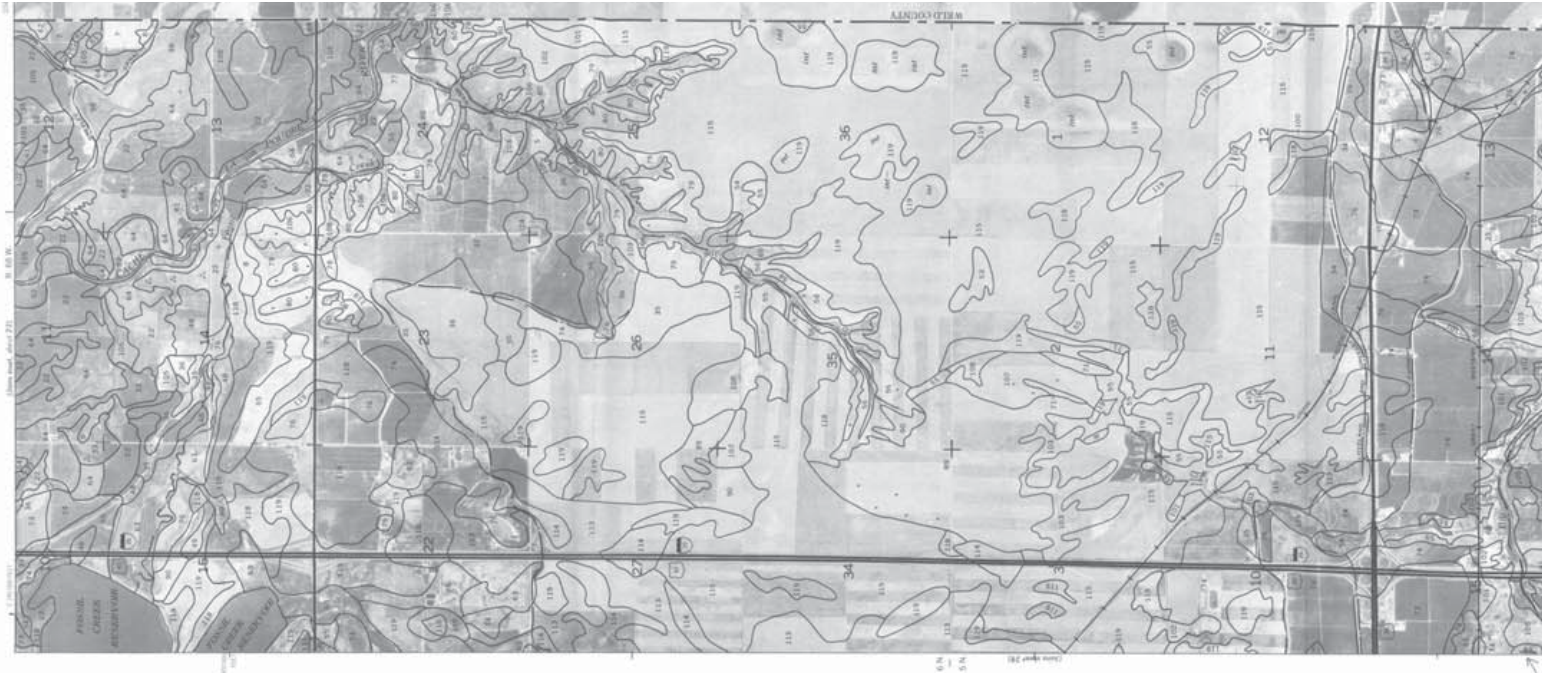
The work reported here was developed under STAR Research Assistance Agreements CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This presentation has not been formally reviewed by EPA. EPA does not endorse any products or commercial services mentioned in this report.

Environmental and Ecological Importance of Soils

- Soils support life
 - viruses, bacteria, fungi, algae, protozoa, mites, nematodes, worms, insects and insect larvae, larger animals, and plant roots
- Soils buffer environmental change
 - provide sink for greenhouse gases
 - regulate and partition water flow (infiltration vs. runoff)
 - decompose organic wastes (pesticides, sewage, etc.)
 - store and degrade nitrates, phosphorus, etc.

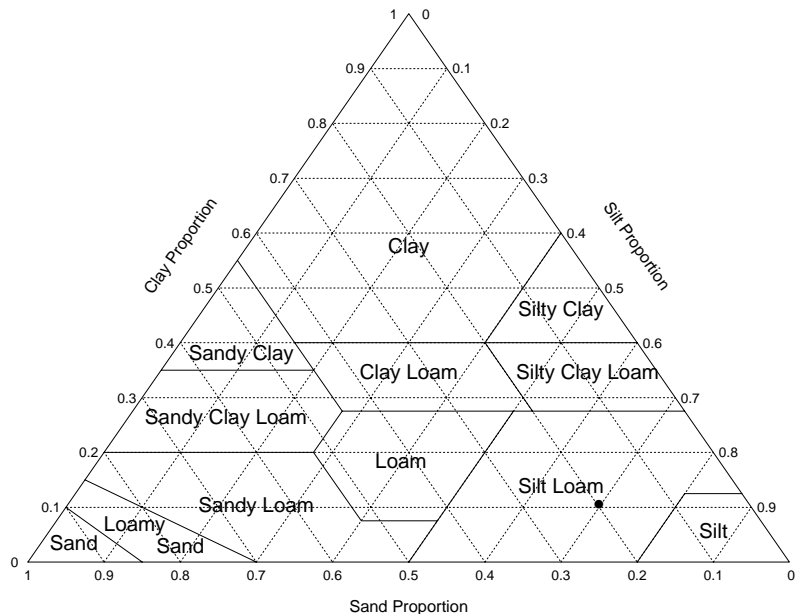
Soil Map and Map Unit Symbols

- Delineations superimposed on aerial photograph
 - map unit symbol 77 is Otero sandy loam, 0–3% slopes



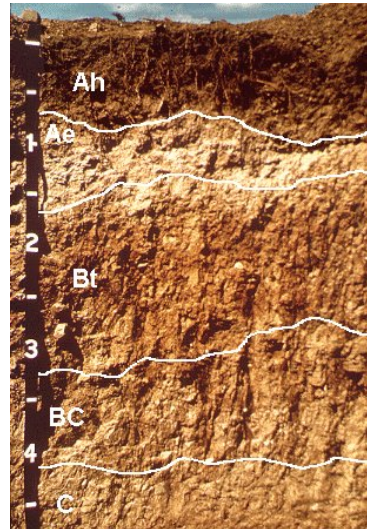
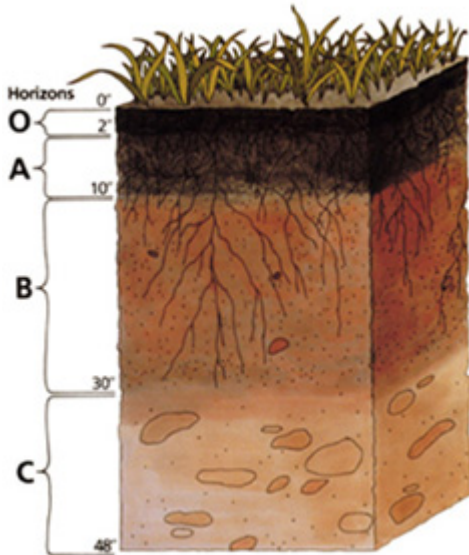
Map Unit Symbols Categorize Surface Information

- Otero sandy loam, 0–3% slopes: slope and texture
 - 70% sand, 20% silt, 10% clay



Soil Horizons and Soil Core Samples

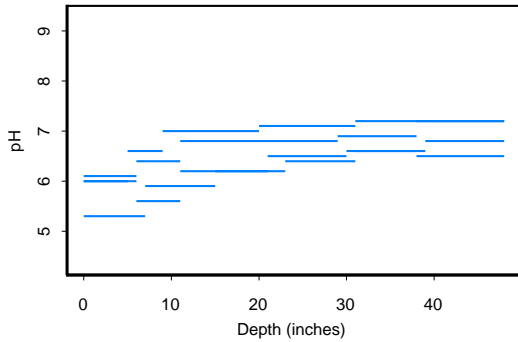
- Horizon thickness, number, and order vary from site to site
- Use probe to select core sample; separate by horizon
- Obtain lab analyses on cylindrical volume from each horizon



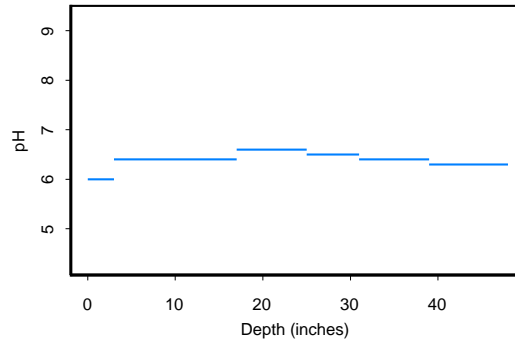
Horizon-Averaged Data: pH

- pH vs. depth by map unit symbol

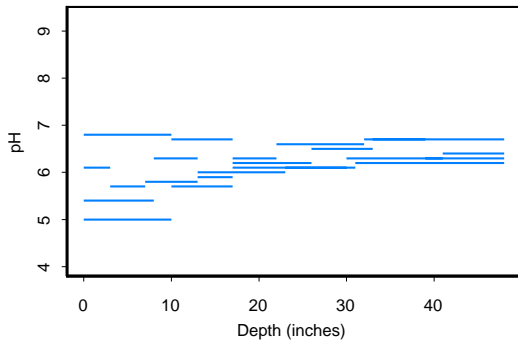
9D2



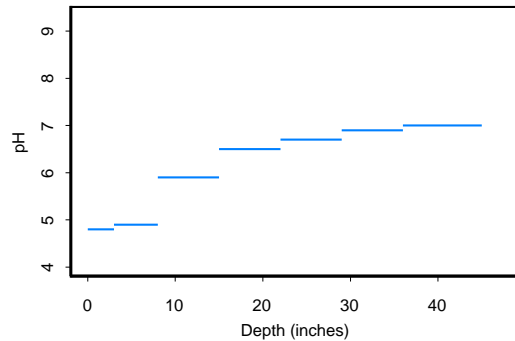
268F



9B



99D3



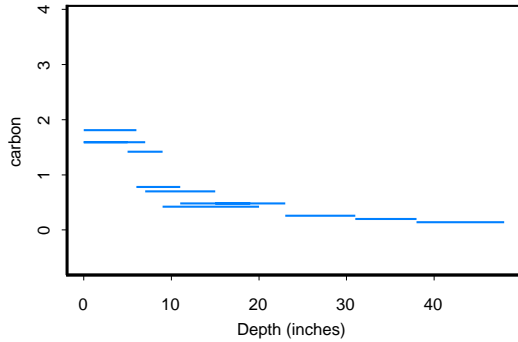
Sampled Cores from Soil Survey Update

- Major Land Resource Area 107 Pilot Project:
 - two counties in western Iowa
 - multiphase sampling design
 - focus on third phase: horizon-specific lab data
- Many study variables of interest
 - pH, carbon, sand, coarse silt, fine silt, clay, nitrogen, calcite, dolomite, CaCO_3
- Want area-specific profiles
 - properties of soil as a function of depth
 - standard is a text description

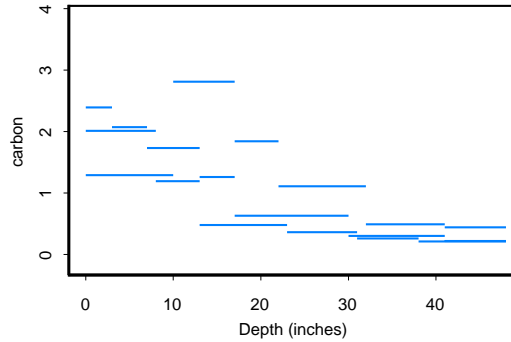
Horizon-Averaged Data: Carbon

- Carbon vs. depth by map unit symbol

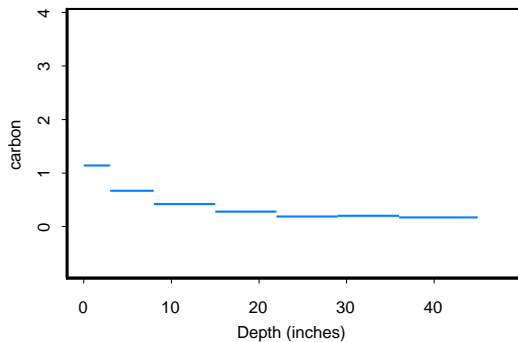
9D2



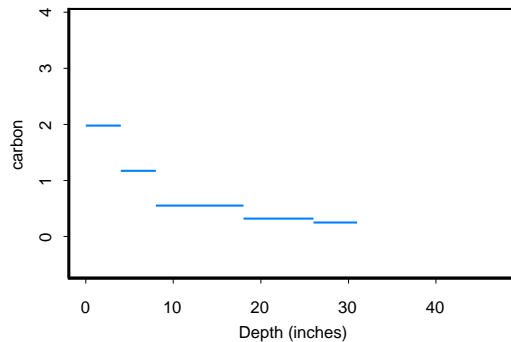
9B



99D3



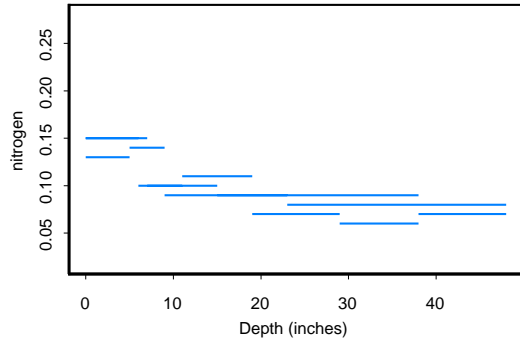
10D2



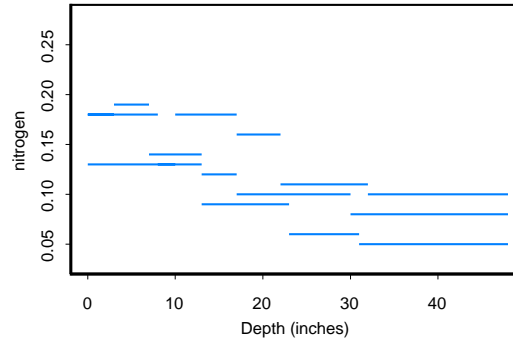
Horizon-Averaged Data: Nitrogen

- Nitrogen vs. depth by map unit symbol

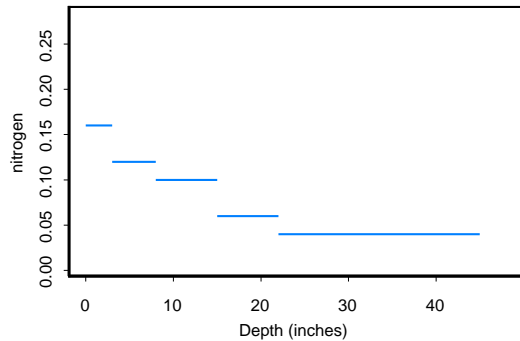
9D2



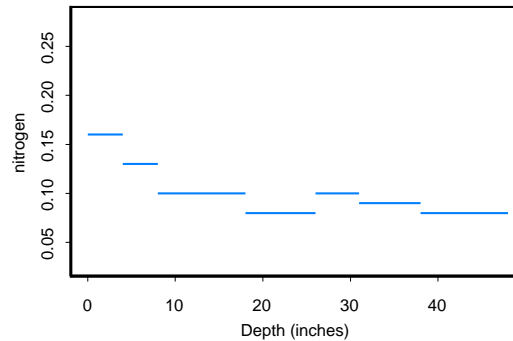
9B



99D3



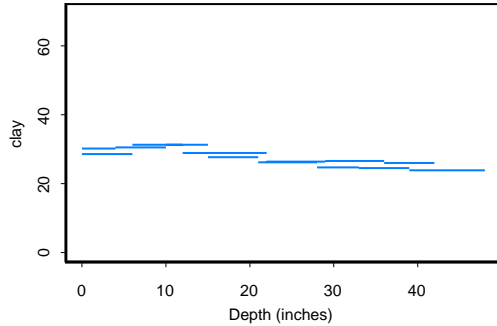
10D2



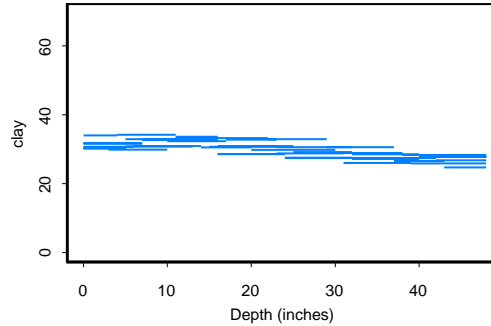
Horizon-Averaged Data: Percent Clay

- Clay vs. depth by map unit symbol

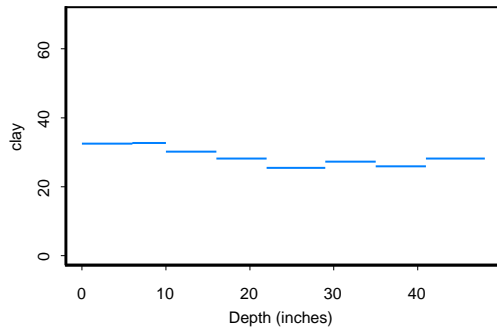
99D



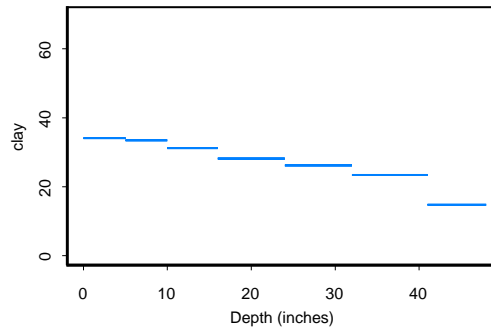
9D



99D2



9E3



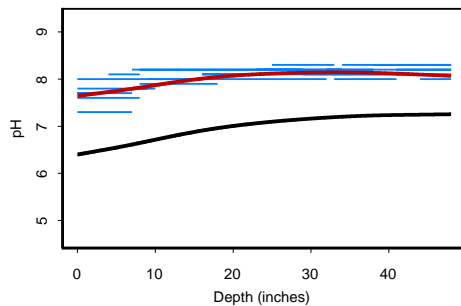
Key Data Features

- Horizon averaging
 - irregular, wide; measurements not depth-specific
 - difficult to specify parametric model
- Possible non-linear features (asymptotes)
- Within-core dependence
 - horizons within same core may be correlated
- Small sample sizes for many small areas
 - e.g., 97 map unit symbols, 193 soil cores
- Many study variables

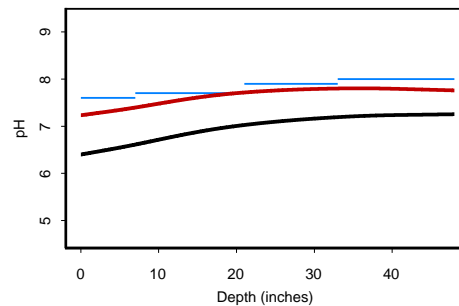
Small Area Estimation of Soil Profiles

- Goal: borrow strength across small areas to obtain profiles for each map unit symbol

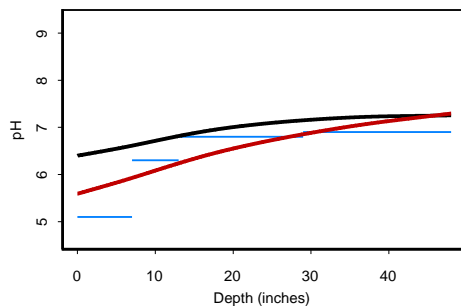
1D3



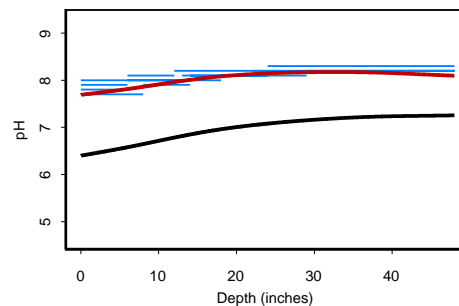
112D3



10C3



1E3



Small Area Estimation: Nested Error Regression Model

- Finite domain mean

$$\begin{aligned}\theta_g &= \frac{1}{N_g} \sum_{s \in D_g} \theta_{gs} \\ &= \frac{1}{N_g} \sum_{s \in D_g} (\mathbf{x}_{gs}^T \boldsymbol{\beta} + U_g + \eta_{gs})\end{aligned}$$

– observe $\{Y_{gs}\} = \{\theta_{gs}\}$ for sample of units

- Battese, Harter, and Fuller (1988)

actual crop acres = (satellite corn, soybean acres)^T $\boldsymbol{\beta}$
+ county-specific intercept
+ noise

Extension: Semiparametric Regression Model

- Generalize random effects (Datta and Ghosh 1991):

$$\begin{aligned}\theta_g &= \frac{1}{N_g} \sum_{s \in D_g} \theta_{gs} \\ &= \frac{1}{N_g} \sum_{s \in D_g} \left(\mathbf{x}_{gs}^T \boldsymbol{\beta} + \mathbf{z}_{gs}^T \mathbf{u} + \eta_{gs} \right)\end{aligned}$$

- Need auxiliary information $\sum_{s \in D_g} \mathbf{x}_{gs}$ and $\sum_{s \in D_g} \mathbf{z}_{gs}$
- Generalizes parametric linear regression to semiparametric regression, possibly area-specific
 - Opsomer, Claeskens, Ranalli, Kauermann and Breidt (2005)

Extension: Continuous Domain

- From finite population to infinite population

$$\begin{aligned}\theta_g &= \frac{1}{|D_g|} \int_{D_g} \theta_{gs} ds \\ &= \frac{1}{|D_g|} \int_{D_g} (\mathbf{x}_{gs}^T \boldsymbol{\beta} + \mathbf{z}_{gs}^T \mathbf{u} + 0) ds\end{aligned}$$

- Measure $Y_{gs} = \theta_{gs} + \epsilon_{gs}$
- Need auxiliary information

$$\int_{D_g} \mathbf{x}_{gs} ds \quad \text{and} \quad \int_{D_g} \mathbf{z}_{gs} ds$$

Extension: Continuous Parameter

- Soil profile indexed by depth

$$\begin{aligned}\theta_g(t) &= \frac{1}{|D_g|} \int_{D_g} \theta_{gs}(t) ds \\ &= \frac{1}{|D_g|} \int_{D_g} \left(\mathbf{x}_{gs}^T(t) \boldsymbol{\beta} + \mathbf{z}_{gs}^T(t) \mathbf{u} \right) ds\end{aligned}$$

- Need

$$\int_{D_g} \mathbf{x}_{gs}(t) ds \quad \text{and} \quad \int_{D_g} \mathbf{z}_{gs}(t) ds$$

- Measure horizon averages

$$Y_{gsj} = \frac{1}{d_{gsj} - d_{gs,j-1}} \int_{d_{gs,j-1}}^{d_{gsj}} \theta_{gs}(t) dt + \epsilon_{gsj}$$

Semiparametric Mixed Model: Longitudinal

- Observation on subject i at j th time point
- Subject-specific deviation is stochastic process

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + m(t_{ij}; \mathbf{w}_i) + \mathbf{z}_{ij}^T \mathbf{u}_i + U_i(t_{ij}) + \epsilon_{ij}$$

e.g., Zhang, Lin, Raz, Sowers (1998)

- Subject-specific deviation is smooth

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + m(t_{ij}; \mathbf{w}_i) + \mathbf{z}_{ij}^T \mathbf{u}_i + m_i(t_{ij}) + \epsilon_{ij}$$

e.g., Brumback and Rice (1998), Durbán, Harezlak, Wand, and Carroll 2005

- Need to adapt for small areas and horizon averages

Auxiliary Information?

- Need

$$\int_{D_g} \mathbf{x}_{gs}(t) ds \quad \text{and} \quad \int_{D_g} \mathbf{z}_{gs}(t) ds$$

- Depth*Site-specific covariates? Unlikely!
- Depth-specific covariates?
 - functions of t only: polynomials, spline basis functions
- Site-specific covariates?
 - can be remotely sensed
 - map unit symbol already includes land cover, topography, geology, formation/deposition

Small Area Soil Profile Model

- Model for profile of a single core

$$\begin{aligned}\theta_{gs}(t) &= m(t) + m_g(t) + U_{gs}(t) \\ &= \text{global profile} \\ &\quad + \text{area-specific deviation} \\ &\quad + \text{core-specific deviation} \\ &= \beta_0 + \beta_1 t + \sum_{k=1}^K a_k (t - \kappa_k)_+ \\ &\quad + b_{0g} + b_{1g} t + \sum_{k=1}^{K^*} b_{k+1,g} (t - \kappa_k^*)_+ \\ &\quad + U_{gs}(t)\end{aligned}$$

- Fixed, known knots $\{\kappa_k\}_{k=1}^K$ and $\{\kappa_k^*\}_{k=1}^{K^*}$

Semiparametric Small Area Mixed Model: Horizons

- j th horizon average in g th map unit symbol, s th core

$$\begin{aligned} Y_{gsj} &= \frac{1}{d_{gsj} - d_{gs,j-1}} \int_{d_{gs,j-1}}^{d_{gsj}} \theta_{gs}(t) dt + \epsilon_{gsj} \\ &= \beta_0 + \beta_1 \frac{d_{gs,j-1} + d_{gsj}}{2} \\ &\quad + \sum_{k=1}^K \frac{a_k}{2} \left\{ (d_{gsj} - \kappa_k)_+^2 - (d_{gs,j-1} - \kappa_k)_+^2 \right\} \\ &\quad + b_{0g} + b_{1g} \frac{d_{gs,j-1} + d_{gsj}}{2} \\ &\quad + \sum_{k=1}^{K^*} \frac{b_{k+1,g}}{2} \left\{ (d_{gsj} - \kappa_k^*)_+^2 - (d_{gs,j-1} - \kappa_k^*)_+^2 \right\} \\ &\quad + \frac{1}{d_{gsj} - d_{gs,j-1}} \int_{d_{gs,j-1}}^{d_{gsj}} U_{gs}(t) dt + \epsilon_{gsj} \end{aligned}$$

Mixed Model Formulation of Penalized Splines

- Regard semiparametric profile model as linear mixed model (Ruppert, Wand and Carroll 2003; Durbán et al. 2005)

$$\begin{aligned} \{a_k\} &\text{ iid } \mathbf{N}(0, \sigma_a^2), & \left\{ \begin{bmatrix} b_{0g} \\ b_{1g} \end{bmatrix} \right\} &\text{ iid } \mathbf{N}(\mathbf{0}, \Sigma), \\ \{b_{k+1,g}\} &\text{ iid } \mathbf{N}(0, \sigma_b^2), & \{\epsilon_{gsj}\} &\text{ iid } \mathbf{N}(0, \sigma_\epsilon^2) \end{aligned}$$

- For now, assume simple within-core dependence model:

$$\{U_{gs}(t)\} \equiv \{U_{gs}\} \text{ iid } \mathbf{N}(0, \sigma_U^2)$$

Model Fitting: Global Spline + Area Line

- Straightforward estimation with standard software: S-Plus

```
#
# Set up global linear splines, then integrate over horizons.
#
Z1 <- outer(d1, knots, "-")
Z1 <- Z1 * (Z1 > 0)
Z2 <- outer(d2, knots, "-")
Z2 <- Z2 * (Z2 > 0)
Z <- 0.5 * diag(1/(d2 - d1)) %*% (Z2^2 - Z1^2)
#
# Fit global spline plus area-specific line using lme. Midd is horizon midpoint.
# Adapted from Durban, Harezlak, Wand and Carroll 2005.
#
SplineCoeff <- factor(rep(1, length(y)))
#
fit <- lme(y ~ Midd, random = list(SplineCoeff = pdIdent( ~ Z - 1),
  MUS = pdSymm( ~ Midd), Core = pdIdent( ~ 1)))
```

Model Comparisons

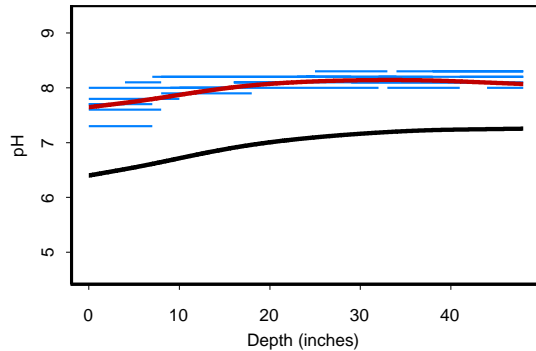
- Likelihood ratio tests to determine if variance parameter is zero or positive
 - best to use percentiles based on simulation
(Crainiceanu and Ruppert 2003, Crainiceanu, Ruppert, Claeskens, and Wand 2004)
 - rough comparison based on 90th percentile of $\frac{1}{2}\chi_2^2 + \frac{1}{2}\chi_3^2$: 5.528
(Durbán, Harezlak, Wand, and Carroll 2005)

	carbon	nitrogen	pH	clay
# horizons	393	432	1192	1192
# cores	75	75	193	193
# map unit symbols	46	46	97	97
line+intercept → spline+intercept	40.6	61.5	49.1	0.0
$-2 \log RLR$ spline+intercept → spline+line	80.9	40.6	286.0	129.3
spline+line → spline+spline	20.7	0.0	2.4	5.0
best-core → best	24.0	20.7	393.2	66.7

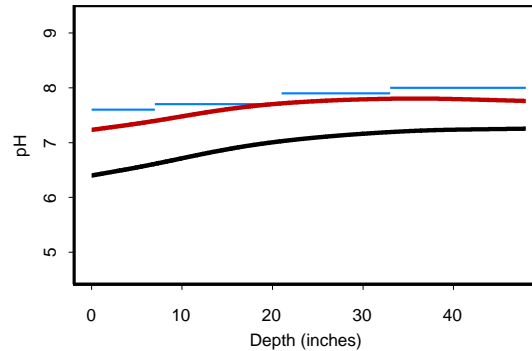
Small Area Profile Estimates: Global Spline, Local Line

- pH vs. depth profiles by map unit symbol

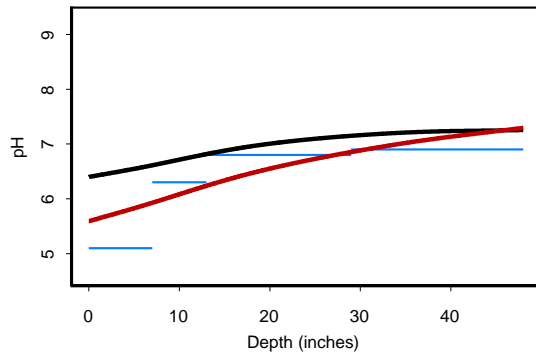
1D3



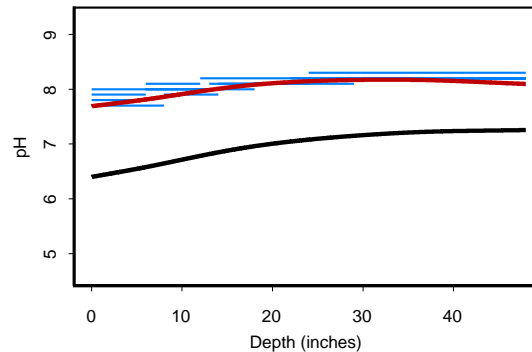
112D3



10C3



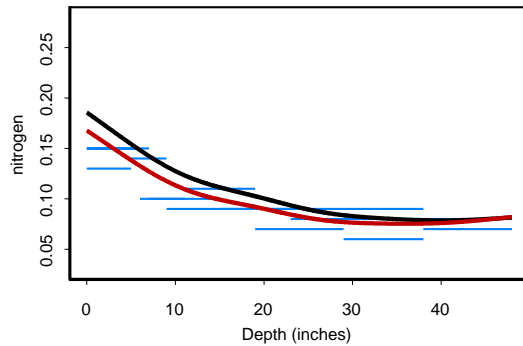
1E3



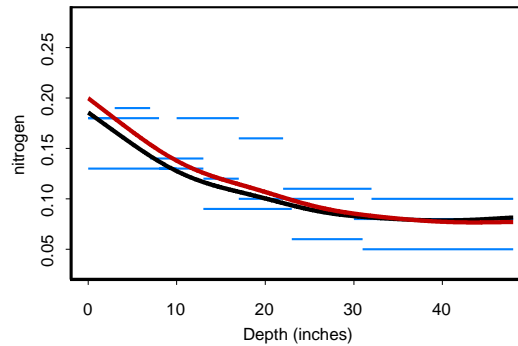
Small Area Profile Estimates: Global Spline, Local Line

- Nitrogen vs. depth profiles by map unit symbol

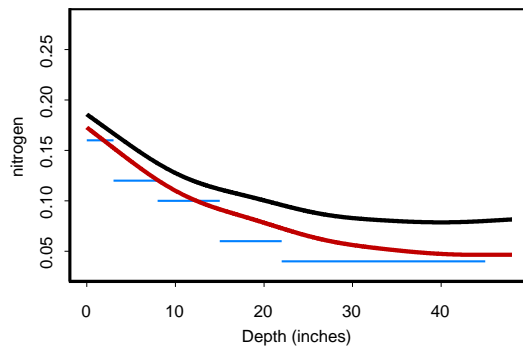
9D2



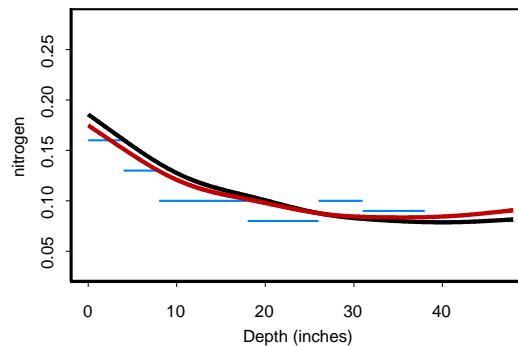
9B



99D3



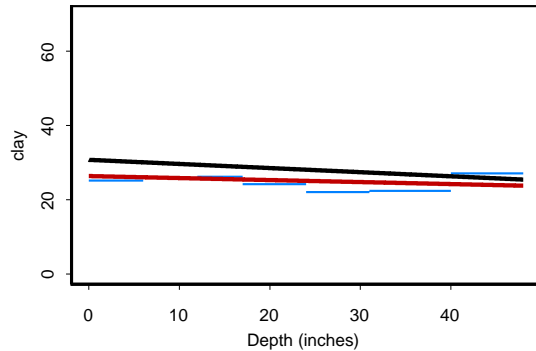
10D2



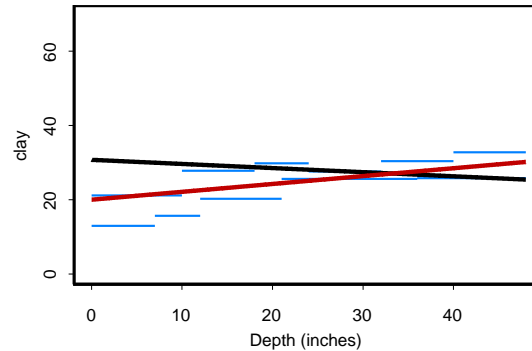
Small Area Profile Estimates: Global Line, Local Line

- Clay vs. depth profiles by map unit symbol

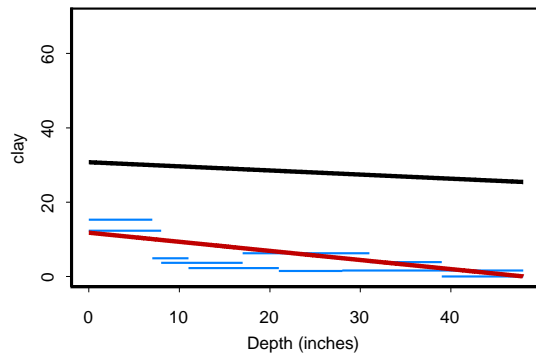
527B



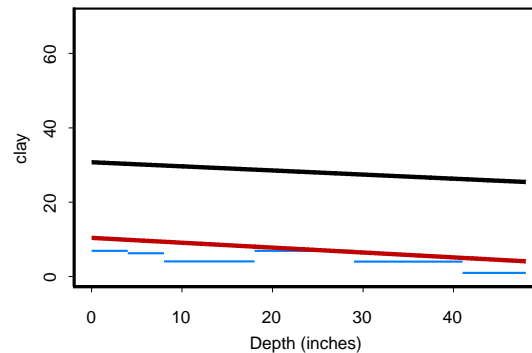
212+



237



237C



Types of Small Area Estimates for Soil Profiles

- Point: classified on site as a given map unit symbol

$$\hat{\theta}_g(t) \pm 2\hat{\sigma}_\epsilon \sqrt{C_{gt}(C^T C + \hat{\Lambda})^{-1} C_{gt}^T}$$

- Field: small unit with detailed soil map
 - field proportions by map unit symbol:

$$\boldsymbol{\psi} = (\psi_1, \dots, \psi_G)^T$$

$$\boldsymbol{\psi}^T (\hat{\theta}_1(t), \dots, \hat{\theta}_G(t))^T \pm 2\hat{\sigma}_\epsilon \sqrt{\boldsymbol{\psi}^T C_{*t} (C^T C + \hat{\Lambda})^{-1} C_{*t}^T \boldsymbol{\psi}}$$

- Soil map unit: all delineations with same map unit symbol
 - replace $\boldsymbol{\psi}$ by $\hat{\boldsymbol{\psi}}$; adjust uncertainty

Spatial Correlation

- Diagnostic F -test for horizon averages (Breidt, Hsu, and Coar 2005)
- Low rank formulation: project horizon averages of $U_{gs}(t)$ onto L orthogonal rv's (Ruppert, Wand and Carroll 2003)

$$\mathbf{u}_{gs} = \sigma_U \mathbf{\Omega}_L^{-1/2} \left[\frac{\int_{\tau_{\ell-1}}^{\tau_{\ell}} U_{gs}(t) dt}{\tau_{\ell} - \tau_{\ell-1}} \right]_{1 \leq \ell \leq L}$$

– $0 < \tau_1 < \tau_2 < \dots < \tau_L$ are fixed, known knots

- Choices of $U_{gs}(t)$ (Breidt, Hsu, and Ogle 2004):

– $U_{gs}(t) \equiv U_{gs}$ iid $\mathbf{N}(0, \sigma_U^2)$

– $U_{gs}(t)$ non-homogeneous Ornstein-Uhlenbeck process

$$\text{corr}(U_{gs}(r), U_{gs}(t)) = \exp\{-\gamma|r - t|\}$$

Summary

- Semiparametric mixed model for small area profiles
 - flexibly models horizon averages
 - handles fixed and random effects
 - fits in standard linear mixed model framework
- Further work
 - look for useful auxiliary information
 - investigate other dependence structures for noise
 - generalized mixed models