

**Local Polynomial Regression Estimators
in Survey Sampling**

F. Jay Breidt
and
Jean D. Opsomer

Iowa State University

April 12, 1999

Outline

- Use of auxiliary information in surveys
 - operational considerations
 - review of estimation procedures
- Local polynomial regression estimators
 - asymptotic properties
 - variance estimation
 - Godambe-Joshi lower bound
- Example
 - soil mapping
 - Iowa pilot project
- Summary
 - future work

Auxiliary Information

- Finite population $U_N = \{1, 2, \dots, N\}$
- Draw sample $s \subset U_N$ (sample size n) via $p(\cdot)$
- Observe study variables, $y_i, i \in s$
- Obtain complete auxiliary information $x_i, i \in U_N$
 - household-specific rent or housing value on U.S. Census
 - individual-specific taxable income on Swedish population register
 - pixel-specific spectral value from Landsat image
- Modeling?

Modeling Environment

- Typical survey situation:
 - statistical agency collects data, auxiliary info x
 - data set is created and released to users
 - data set reflects knowledge of design and x
- Agency knows x , but may not know y :
 - many study variables
 - unlimited derived variables:

$$y, z, y^2, yz, \mathbf{1}_{\{y \leq t\}}, z\mathbf{1}_{\{y \in D\}}$$

- User knows y , but may not know x :
 - magnitude of $x_i, i \in U_N$
 - confidentiality of microdata

Modeling Constraints

- Limited time and other resources
- Potential controversy among end users:
 - Sierra Club vs. timber industry
 - Democrats vs. Republicans
- Estimation strategy:
 - should use information in $x_i, i \in U_N$
 - should *not* release $x_i, i \in U_N$
 - should handle any study variables (and be internally consistent)
 - should *not* require modeling efforts for every study variable
 - should be efficient if model is right
 - should *not* fail if model is wrong

Weighting

- Construct n weights $\{\omega_{is}\}$ for $i \in s$
 - reflect design properties
 - incorporate auxiliary information
 - do not depend on y_i

- Release data set

$$\begin{bmatrix} \omega_{1s} & y_{11} & \cdots & y_{1J} \\ \omega_{2s} & y_{21} & \cdots & y_{2J} \\ \vdots & \vdots & & \vdots \\ \omega_{ns} & y_{n1} & \cdots & y_{nJ} \end{bmatrix}$$

- For any study variable y , estimate $t_y = \sum_{i \in U_N} y_i$ via

$$\hat{t}_y = \sum_{i \in s} \omega_{is} y_i$$

- Internally consistent:

$$\hat{t}_{y+z} = \sum_{i \in s} \omega_{is} (y_i + z_i) = \sum_{i \in s} \omega_{is} y_i + \sum_{i \in s} \omega_{is} z_i = \hat{t}_y + \hat{t}_z$$

Horvitz-Thompson Estimator

- Goal: Estimate $t_y = \sum_{i \in U_N} y_i$
- Define $I_i = 1$ if $i \in s$, 0 otherwise:

$$E_p [I_i] =: \pi_i \text{ and } E_p [I_i I_j] =: \pi_{ij}$$

- Design-unbiased estimator of t_y is

$$\hat{t}_{\text{ht}} = \sum_{i \in U_N} y_i \frac{I_i}{\pi_i} = \sum_{i \in s} \frac{1}{\pi_i} y_i$$

- weights $\{1/\pi_i\}_{i \in s}$ work for any study variable
- do not incorporate x_i

- Variance is

$$\text{Var}_p(\hat{t}_{\text{ht}}) = \sum_{i,j \in U_N} y_i y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$$

- Unbiased variance estimator is

$$\hat{V}(\hat{t}_{\text{ht}}) = \sum_{i,j \in U_N} y_i y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}}$$

Generalized Difference Estimator

- Goal: Estimate $t_y = \sum_{i \in U_N} y_i$
- Suppose we could “guess” the value of y_i , $i \in U_N$
- Using the guesses $\{y_i^0\}$, form the **difference estimator**

$$\hat{t}_{\text{diff}} := \sum_{i \in U_N} y_i^0 + \sum_{i \in U_N} \frac{(y_i - y_i^0)I_i}{\pi_i}$$

Note that

- first term does not depend on sample
- second term is unbiased estimator of $\sum_{i \in U_N} (y_i - y_i^0)$

Properties of the Difference Estimator

$$\hat{t}_{\text{diff}} = \sum_{i \in U_N} y_i^0 + \sum_{i \in U_N} (y_i - y_i^0) \frac{I_i}{\pi_i}$$

- Design expectation is

$$E_p [\hat{t}_{\text{diff}}] = \sum_{i \in U_N} y_i^0 + \sum_{i \in U_N} (y_i - y_i^0) = t_y$$

- Design variance is

$$\text{Var}_p (\hat{t}_{\text{diff}}) = \sum_{i,j \in U_N} (y_i - y_i^0)(y_j - y_j^0) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$$

- Unbiased variance estimator is

$$\hat{V}(\hat{t}_{\text{diff}}) = \sum_{i,j \in U_N} (y_i - y_i^0)(y_j - y_j^0) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}}$$

Model-Assisted Approach

- Superpopulation model, ξ : y_i 's are independent r.v.'s with model mean

$$E_{\xi} [y_i] = m(x_i)$$

and model variance

$$\text{Var}_{\xi} (y_i) = v(x_i)$$

- Estimate parameters of model ξ and get model-based predictions \hat{m}_i
- Replace y_i^0 in GDE with \hat{m}_i :

$$\hat{t}_y = \sum_{i \in U_N} \hat{m}_i + \sum_{i \in U_N} (y_i - \hat{m}_i) \frac{I_i}{\pi_i}$$

- model-based prediction + design bias adjustment
- if model is good, should have small design variance
- With $E_{\xi} [y_i] = \mathbf{x}'_i \boldsymbol{\beta}$, get generalized regression (GREG) estimator
 - ADU and design-consistent (Robinson and Särndal, 1982)

Generalized Regression Estimator: Examples

- Regression through origin with $v(x_i) \propto x_i$: classic ratio estimator
- Heteroskedastic ANOVA: poststratification estimator
- Simple linear regression: $E_\xi[y_i] = \beta_0 + \beta_1 x_i$, $\text{Var}_\xi(y_i) = \sigma^2$

$$\hat{\mathbf{B}} = \begin{bmatrix} \tilde{y}_s - \hat{B}_1 \tilde{x}_s \\ \hat{B}_1 \end{bmatrix},$$

$$\tilde{y}_s = \frac{\sum_{i \in s} y_i \pi_i^{-1}}{\sum_{i \in s} \pi_i^{-1}}, \quad \hat{B}_1 = \frac{\sum_s x_i y_i \pi_i^{-1} - \sum_s x_i \pi_i^{-1} \sum_s y_i \pi_i^{-1} / \sum_s \pi_i^{-1}}{\sum_s x_i^2 \pi_i^{-1} - (\sum_s x_i \pi_i^{-1})^2 / \sum_s \pi_i^{-1}}$$

Then

$$\begin{aligned} \hat{t}_{\text{greg}} &= \sum_{i \in U_N} \mathbf{x}'_i \hat{\mathbf{B}} + \sum_{i \in U_N} (y_i - \mathbf{x}'_i \hat{\mathbf{B}}) \frac{I_i}{\pi_i} \\ &= N \left\{ \tilde{y}_s + \hat{B}_1 (\bar{x}_U - \tilde{x}_s) \right\}, \end{aligned}$$

the classic regression estimator

Polynomial Regression Estimator

- Model: $m(x_i) = \beta_0 + \beta_1 x_i + \cdots + \beta_q x_i^q$, $v(x_i) = \text{constant}$

- Define

$$\mathbf{X}_s = \left[1 \quad x_j - \bar{x}_U \quad \cdots \quad (x_j - \bar{x}_U)^q \right]_{j \in s}$$

and

$$\mathbf{W}_s = \text{diag} \left\{ \frac{1}{\pi_j} \right\}_{j \in s}$$

- Polynomial regression estimator of $m(x_i)$ based on s is

$$\hat{m}_i = \left[1 \quad x_i - \bar{x}_U \quad \cdots \quad (x_i - \bar{x}_U)^q \right] (\mathbf{X}'_s \mathbf{W}_s \mathbf{X}_s)^{-1} \mathbf{X}'_s \mathbf{W}_s \mathbf{y}_s$$

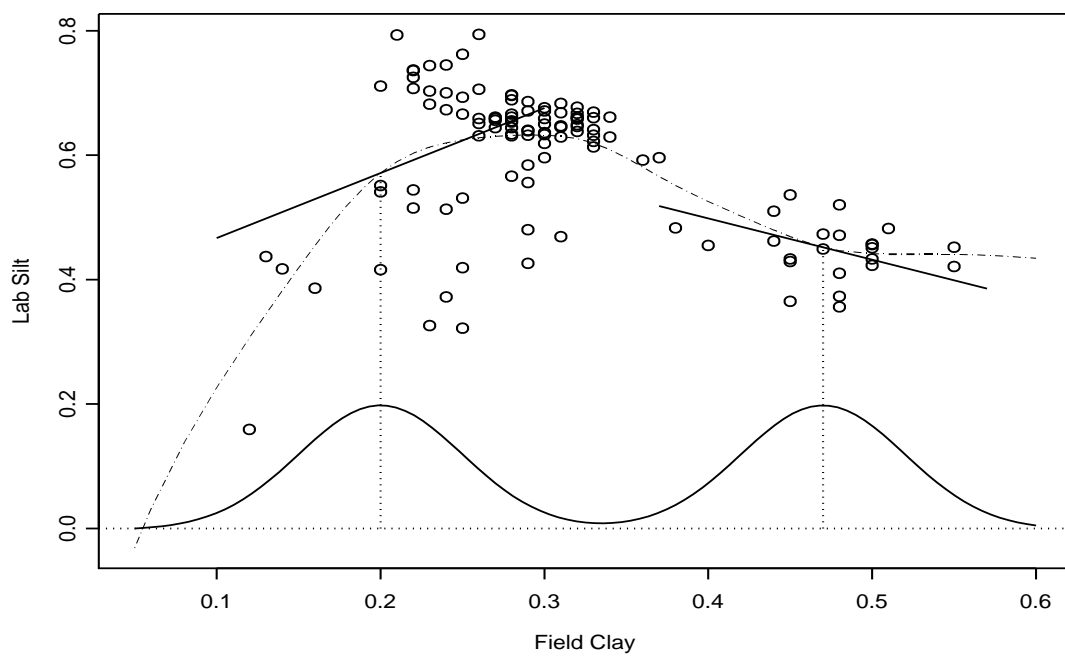
- GREG estimator of t_y is then

$$\hat{t}_{\text{greg}} = \sum_{i \in U_N} \hat{m}_i + \sum_{i \in U_N} (y_i - \hat{m}_i) \frac{I_i}{\pi_i}$$

- Strong modeling assumptions

Local Polynomial Regression

- Nonparametric model, ξ : $m(x_i)$ is smooth, $v(x_i)$ is positive and smooth
- Locally weighted least squares fits (Wand and Jones, 1995)



Local Polynomial Regression Estimator

- Model: $m(x_i)$ is smooth, $v(x_i)$ is positive and smooth
- Define

$$\mathbf{X}_{si} = \left[1 \quad x_j - x_i \quad \cdots \quad (x_j - x_i)^q \right]_{j \in s}$$

and

$$\mathbf{W}_{si} = \text{diag} \left\{ \frac{1}{\pi_j h} K \left(\frac{x_j - x_i}{h} \right) \right\}_{j \in s}$$

- Local polynomial regression estimator of $m(x_i)$ based on s is

$$\hat{m}_i = [1, 0, \dots, 0] (\mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{X}_{si})^{-1} \mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{y}_s$$

- LPR estimator of t_y is then

$$\hat{t}_{\text{lpr}} = \sum_{i \in U_N} \hat{m}_i + \sum_{i \in U_N} (y_i - \hat{m}_i) \frac{I_i}{\pi_i}$$

Weighting and Calibration

- Define

$$\mathbf{w}_{si} = [1, 0, \dots, 0] (\mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{X}_{si})^{-1} \mathbf{X}'_{si} \mathbf{W}_{si}$$

- Weights

$$\begin{aligned} \hat{t}_{\text{lpr}} &= \sum_{i \in s} \left\{ \frac{1}{\pi_i} + \sum_{j \in U_N} \left(1 - \frac{I_j}{\pi_j} \right) \mathbf{w}'_{sj} \mathbf{e}_i \right\} y_i \\ &= \sum_{i \in s} \omega_{is} y_i \end{aligned}$$

where \mathbf{e}_i is the i th column of the $n \times n$ identity matrix

– can be applied to any study variable

- Calibration (Deville and Särndal, 1992)

$$\sum_{i \in s} \omega_{is} x_i^\ell = \sum_{i \in U_N} x_i^\ell$$

for $\ell = 0, 1, \dots, q$

– if $y_i = \beta_0 + \beta_1 x_i + \dots + \beta_q x_i^q$, then $\hat{t}_{\text{lpr}} = t_y$ for every possible sample

Asymptotic Framework

- Population assumptions:

- population size $N \rightarrow \infty$
- mean function $m(x)$ smooth, variance function $v(x)$ smooth and positive
- population “moments”: for $\epsilon_i := y_i - m(x_i)$,

$$\limsup_{N \rightarrow \infty} N^{-1} \sum_{i \in U_N} |\epsilon_i|^4 < \infty$$

with ξ -probability one

- Smoothing assumptions:

- kernel $K(\cdot)$ has compact support, is positive, symmetric, continuous
- bandwidth $h_N \rightarrow 0$ and $Nh_N \rightarrow \infty$

Asymptotic Framework, Continued

- Design assumptions:

- sampling rate $n_N N^{-1} \rightarrow \pi \in (0, 1)$
- probability sampling design: for all N , $\min_{i \in U_N} \pi_i \geq \lambda > 0$
- measurability: for all N , $\min_{i, j \in U_N} \pi_{ij} \geq \lambda^* > 0$
- limited dependence:

$$\limsup_{N \rightarrow \infty} n_N \max_{i, j \in U_N: i \neq j} |\pi_{ij} - \pi_i \pi_j| < \infty$$

$$\limsup_{N \rightarrow \infty} N^2 \max_{(i, j, k, \ell) \text{ distinct}} |\mathbb{E}_p [(I_i - \pi_i)(I_j - \pi_j)(I_k - \pi_k)(I_\ell - \pi_\ell)]| < \infty$$

$$\lim_{N \rightarrow \infty} \max_{(i, j, k, \ell) \text{ distinct}} |\mathbb{E}_p [(I_i I_j - \pi_{ij})(I_k I_\ell - \pi_{k\ell})]| = 0$$

and

$$\limsup_{N \rightarrow \infty} n_N \max_{(i, j, k) \text{ distinct}} |\mathbb{E}_p [(I_i - \pi_i)^2 (I_j - \pi_j)(I_k - \pi_k)]| < \infty$$

- (Holds for simple random sampling without replacement)

Asymptotic Properties of LPR Estimator

- Asymptotically design unbiased (ADU):

$$\lim_{N \rightarrow \infty} E_p \left[\frac{\hat{t}_{\text{lpr}} - t_y}{N} \right] = 0 \text{ with } \xi\text{-probability one}$$

- Design consistent:

$$\lim_{N \rightarrow \infty} E_p \left[\mathbf{1}_{\left\{ N^{-1} |\hat{t}_{\text{lpr}} - t_y| > \eta \right\}} \right] = 0 \text{ with } \xi\text{-probability one}$$

for all $\eta > 0$

- Design mean squared error:

$$E_p \left(\frac{\hat{t}_{\text{lpr}} - t_y}{N} \right)^2 = \frac{1}{N^2} \sum_{i,j \in U_N} (y_i - m_i)(y_j - m_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} + o(n_N^{-1})$$

Variance Estimation

- ADU and design consistent variance estimation:

$$\lim_{N \rightarrow \infty} n_N E_p |\hat{V}(N^{-1}\hat{t}_{\text{1pr}}) - \text{AMSE}(N^{-1}\hat{t}_{\text{1pr}})| = 0$$

with ξ -probability one, where

$$\hat{V}(N^{-1}\hat{t}_{\text{1pr}}) = \frac{1}{N^2} \sum_{i,j \in U_N} (y_i - \hat{m}_i)(y_j - \hat{m}_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}}$$

and

$$\text{AMSE}(N^{-1}\hat{t}_{\text{1pr}}) = \frac{1}{N^2} \sum_{i,j \in U_N} (y_i - m_i)(y_j - m_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$$

Central Limit Theorem

- Under simple random sampling without replacement,

$$\frac{N^{-1}(\hat{t}_{\text{lpr}} - t_y)}{\hat{V}^{1/2}(N^{-1}\hat{t}_{\text{lpr}})} \xrightarrow{\mathcal{L}} N(0, 1)$$

as $N \rightarrow \infty$, with ξ -probability one, where

$$\begin{aligned}\hat{V}(N^{-1}\hat{t}_{\text{lpr}}) &= \left(1 - \frac{n_N}{N}\right) \frac{\sum_{i \in s} (y_i - \hat{m}_i)^2 - n_N^{-1} [\sum_{i \in s} (y_i - \hat{m}_i)]^2}{n_N(n_N - 1)} \\ &= \left(1 - \frac{n_N}{N}\right) \frac{S_{\text{resid}}^2}{n_N}\end{aligned}$$

Godambe-Joshi Lower Bound

- Godambe and Joshi (1965): for any estimator \hat{t}_y satisfying

$$E_p \left[N^{-1}(\hat{t}_y - t_y) \right] = 0,$$

the following inequality holds:

$$E_\xi E_p \left(\frac{\hat{t}_y - t_y}{N} \right)^2 \geq \frac{1}{N^2} \sum_{i \in U_N} v(x_i) \frac{1 - \pi_i}{\pi_i}$$

- LHS is *anticipated variance*
- RHS minimized for $\pi_i \propto v^{1/2}(x_i)$

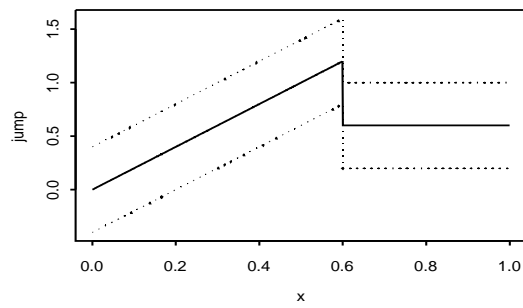
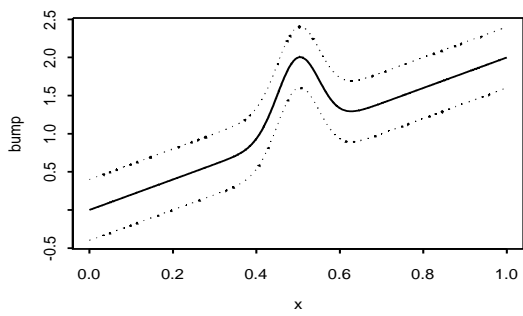
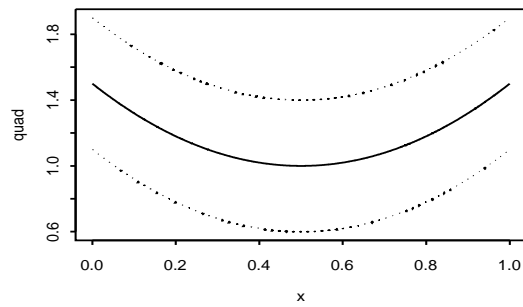
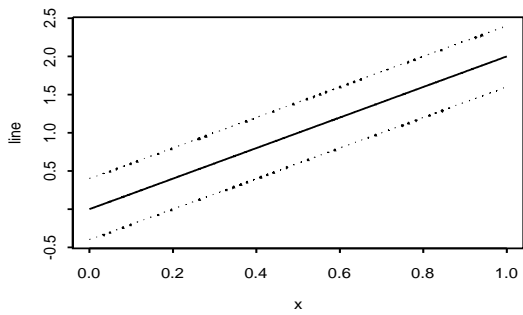
- LPR asymptotically attains GJLB:

$$E_\xi E_p \left(\frac{\hat{t}_{\text{lpr}} - t_y}{N} \right)^2 = \frac{1}{N^2} \sum_{i \in U_N} v(x_i) \frac{1 - \pi_i}{\pi_i} + o(n_N^{-1})$$

- GREG attains GJLB if model is linear (Wright, 1983)

Simulation Experiment

- $N = 1000$, $\{\epsilon_i\}$ iid $N(0, 0.2^2)$, $n = 100$



Simulation Results

For $h = 0.1$ and 200 replications of simple random sampling,

- Relative design bias, $E_p [\hat{t} - t_y] / t_y$:

	line	quad	bump	jump
HT	0.0028	-0.0003	0.0031	-0.0005
GREG	-0.0008	-0.0010	0.0007	0.0010
LPR	-0.0006	-0.0001	0.0003	-0.0002

- LPR efficiency, $\text{Var}_p(\hat{t}) / \text{Var}_p(\hat{t}_{\text{lpr}})$:

	line	quad	bump	jump
HT	7.8819	1.2991	5.8143	2.2429
GREG	0.9787	1.3113	1.7279	2.0369

- Mean of estimated standard error over MC standard error:

	line	quad	bump	jump
HT	1.0990	1.0660	1.0668	1.0250
GREG	0.9891	1.0516	1.0030	0.9934
LPR	0.9572	0.9623	0.9677	0.9342

Illustration with Soil Mapping Data

- Soil mapping
 - traditional approach
 - western Iowa pilot project
- Design
 - multiphase, stratified
- Estimation
 - local linear regression for texture data
 - finite population cdf

Soil Mapping

- Definitions:
 - map unit symbol: alpha-numeric soil label
 - soil map: collection of disjoint, labeled polygons
 - soil map unit: collection of polygons with same map unit symbol
- National Cooperative Soil Survey:
 - USDA + state agency, often Agricultural Experiment Station
 - county-level soil surveys: maps and soil series descriptions
 - extensive field work
- Uses:
 - civil engineering
 - agriculture
 - scientific modeling

Traditional Soil Mapping

- Map constructed from:
 - informal model
 - subjective evaluation of topography, vegetation, etc.
 - occasional observation of soil properties at purposively-located sites
- Numerical summaries:
 - midpoint
 - range
- Problems:
 - no consistent measurement protocol
 - poor distributional information

Western Iowa Pilot Project

- Soil survey update in two counties
 - Crawford and Woodbury
- Goals:
 - statistically defensible design
 - field work commitment comparable to traditional update
 - consistent measurement protocols

Technological Innovations

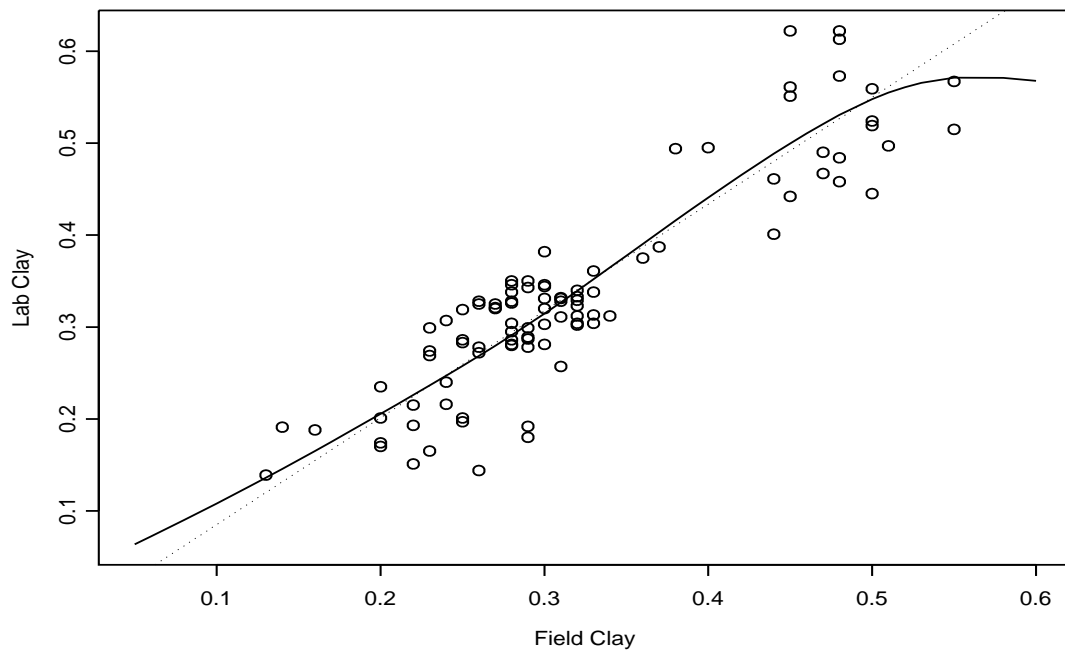
- Geographic Information System (GIS)
 - digitized soil maps from previous surveys
 - allows detailed stratification
- Global Positioning System (GPS)
 - allows precise navigation to randomly selected points
- Personal Digital Assistants (PDA)
 - simplify collection of field data
 - enforce consistent protocol
 - perform edits in the field

Two-Phase Sampling Design

- Surface horizon points ($N = 665$)
 - stratified by soil map unit
 - field observations on top soil horizon
 - let x_i denote field clay
- Laboratory points ($n = 97$)
 - subsample (roughly one in eight) of surface horizon points
 - horizon-specific lab measurements
- Condition on phase one for illustration purposes

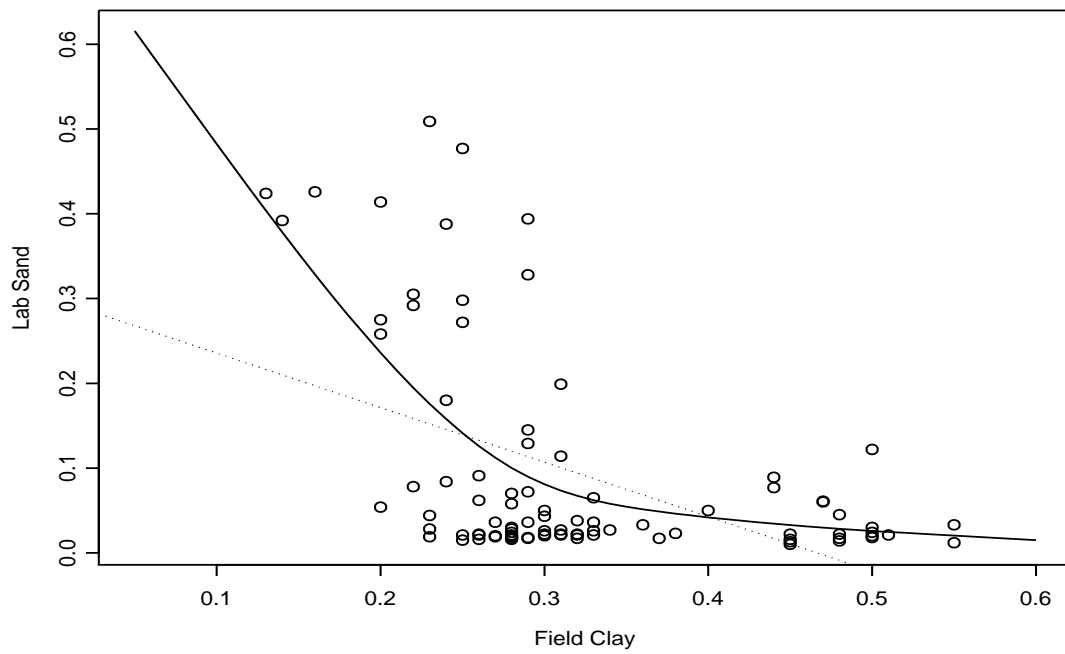
Lab Clay Data

- Local linear with $h = 0.08$:



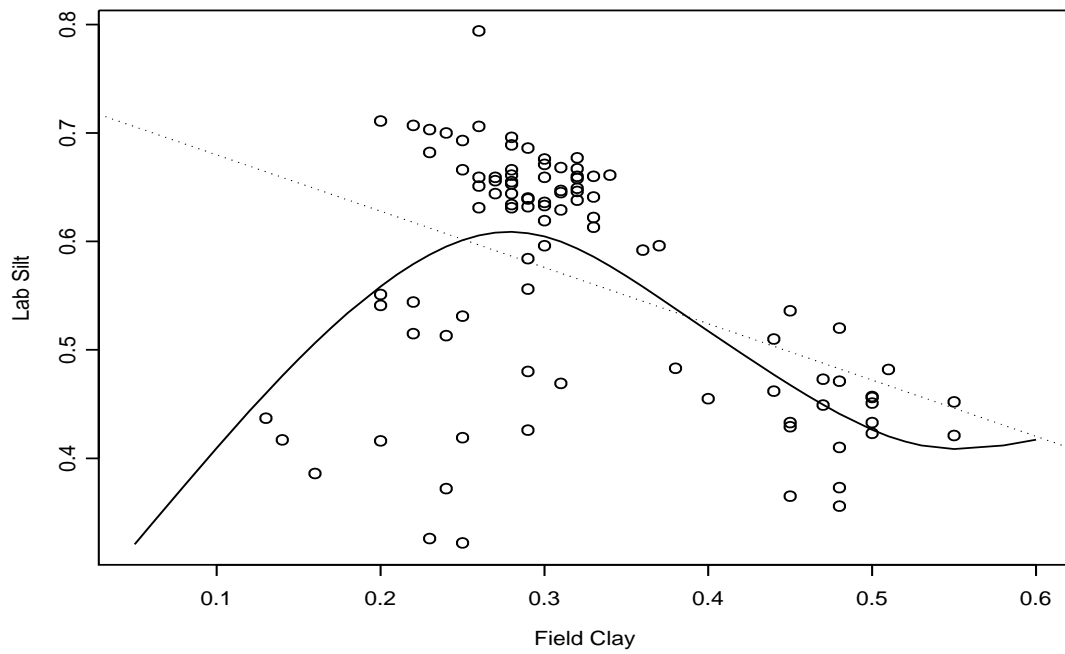
Lab Sand Data

- Local linear with $h = 0.08$:



Lab Silt Data

- Local linear with $h = 0.08$:



Compositional Data

- Local linear with $h = 0.08$:

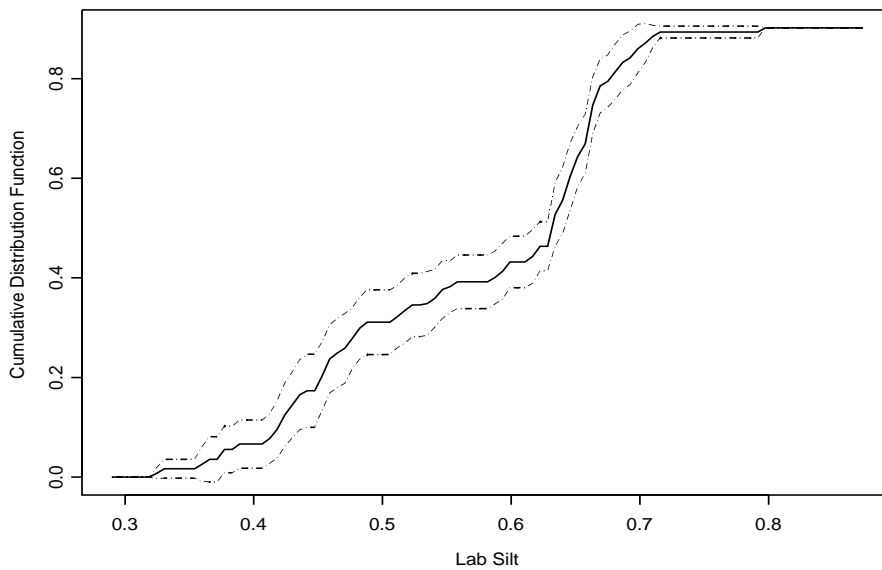
Component	HT	GREG	LPR
Clay	0.346 (0.0129)	0.351 (0.0051)	0.351 (0.0049)
Sand	0.091 (0.0109)	0.088 (0.0067)	0.088 (0.0064)
Silt	0.563 (0.0075)	0.561 (0.0072)	0.562 (0.0059)
Total	1.000	1.000	1.000

Finite Population CDF Estimation

- Finite population cdf:

$$F(t) = N^{-1} \sum_{i \in U_N} \mathbf{1}_{\{y_i \leq t\}} = N^{-1} \sum_{i \in U_N} y_i^*$$

- model mean $m^*(x_i) = G\left(\frac{t - m(x_i)}{v^{1/2}(x_i)}\right)$
- model variance $v^*(x_i) = m^*(x_i)(1 - m^*(x_i))$



Summary

- Use of auxiliary information
 - operational constraints
 - weighting and calibration
- Estimation strategies
 - Horvitz-Thompson estimator
 - generalized difference estimator
 - generalized regression estimator
- Local polynomial regression estimator
 - ADU and design consistency
 - variance estimation
 - asymptotic normality
 - robustness
- Application to soil mapping
 - finite population cdf

Further Work

- Simulation studies
 - mean function, variance function, error distribution
 - kernel function, bandwidth selection
 - sampling design
- Variance estimation alternatives
- Extensions
 - multiple auxiliary variables
 - multi-phase designs