

**Nonparametric Survey Regression
Estimation
in Two-Stage Spatial Sampling**

Siobhan Everson-Stewart

**F. Jay Breidt
Colorado State University**

**Jean D. Opsomer
Ji-Yeon Kim
Iowa State University**

**Statistical Survey Design and
Analysis for Aquatic Resources
September 21, 2002**

Research supported by EPA grants
R-82909501-0 to Colorado State University and
R-82909601-0 to Oregon State University

Outline

- Use of auxiliary information in surveys
 - operational considerations
 - review of estimators
- Nonparametric regression estimator
 - general case
 - two-dimensional extension
 - construction of weights for multiple status estimates
- Example: Northeastern lakes survey
- Further work

Auxiliary Information

- Finite population of clusters:
 $C = \{1, \dots, i, \dots, M\}$
- Draw a sample $s \subset C$ of size m
- For each sampled cluster, $i \in s$, a sample, s_i , is drawn from U_i
- Observe y_{ij}
- Obtain complete auxiliary information at the cluster level x_i , $i \in C$
 - elevation, slope, and aspect from digital elevation model
 - ecological indicators from GIS coverage
 - pixel-specific spectral values from Landsat image

Modeling Environment

- Common survey situation:
 - statistical agency collects data, auxiliary info (x)
 - data set is created and released to users
 - data set reflects knowledge of design and auxiliary information, x
 - agency is responsible for estimating status of many study variables

Modeling Constraints

- Like to use x to improve estimates for y
- Limited time and other resources
- Potential controversy among end users
- Estimation strategy
 - should use information in $x_i, i \in C$
 - should handle many study variables
 - should not require modeling efforts for every study variable
 - should be efficient if model is right
 - should not fail if model is wrong

Model-Assisted Estimators

$$\frac{1}{N} \left\{ \sum_{i \in C} \hat{\mu}_i + \sum_{i \in s} \frac{\bar{y}_i - \hat{\mu}_i}{\pi_i} \right\}$$

where

$$\bar{y}_i = \left(\sum_{j \in s_i} \frac{y_{ij}}{\pi_{j|i}} \right) \left(\sum_{j \in s_i} \frac{1}{\pi_{j|i}} \right)^{-1}$$

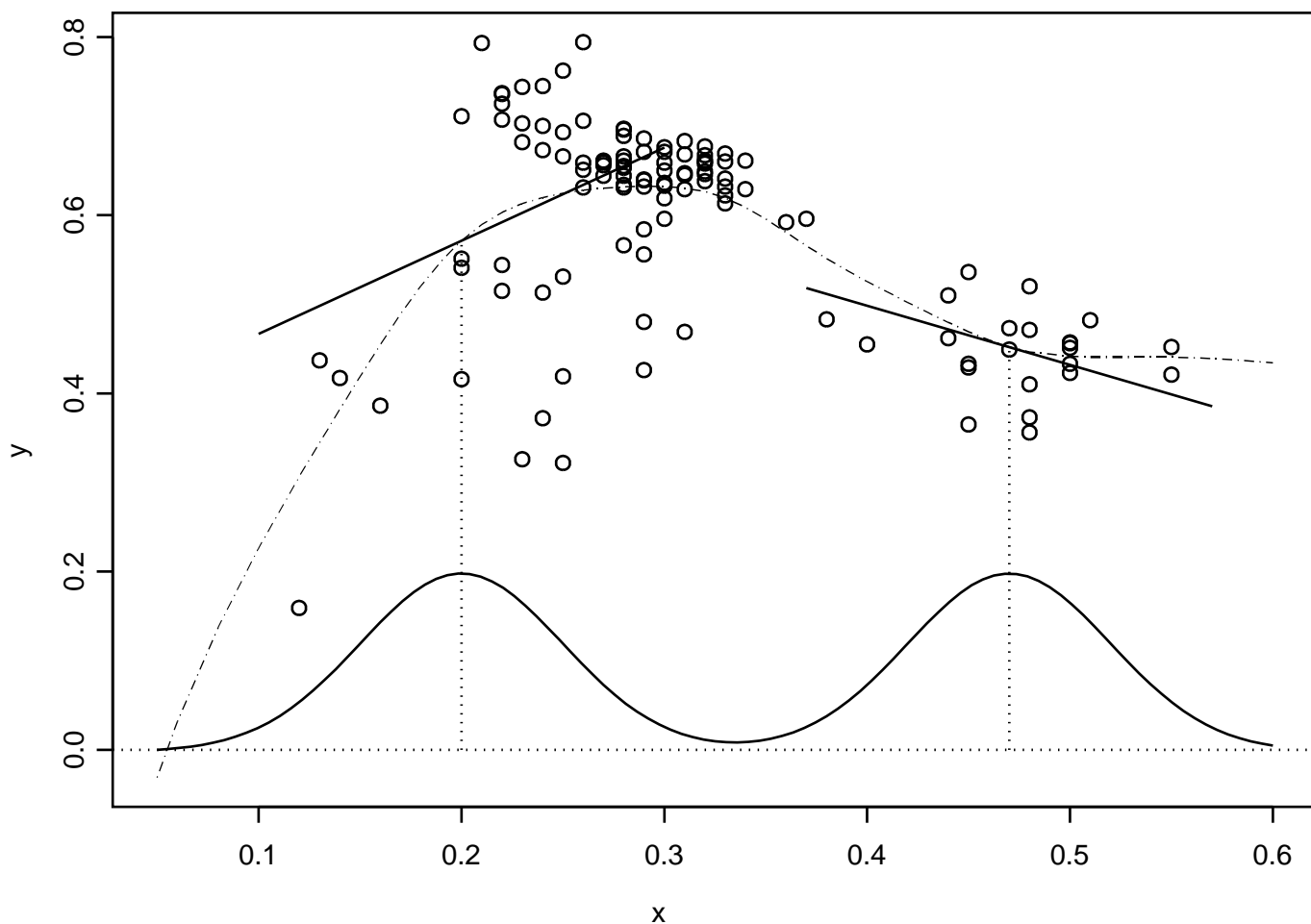
- Model-based prediction + design bias adjustment
- Approximately design-unbiased, with small variance if model is correct.
- Horvitz-Thompson: $\hat{\mu}_i \equiv 0$ since no auxiliary information is used
- Generalized Regression: $\hat{\mu}_i = x_i' \hat{\beta}$
- Local Polynomial Regression: $\hat{\mu}_i$ comes from a kernel smooth

Local Polynomial Regression

- Nonparametric model:

$$\frac{1}{N_i} \sum_{j \in U_i} y_{ij} = m(x_i) + v^{1/2}(x_i) \epsilon_i$$

- Locally weighted least squares fits (Wand and Jones, 1995)



Form of the Estimate for Two Dimensional Case

- Nonparametric Mean Estimator

$$\hat{\mu}_i = e_1' (X_{si}' W_{si} X_{si})^{-1} X_{si}' W_{si} \bar{y}_s = w_{si}' \bar{y}_s$$

- Local Design Matrix

$$X_{si} = \left[1 \quad x_j - x_i \quad y_j - y_i \right]_{j \in s}$$

- Local Weighting Matrix

$$W_{si} = \text{Diag} \left\{ \frac{1}{\pi_j h^2} K \left(\frac{x_j - x_i}{h} \right) K \left(\frac{y_j - y_i}{h} \right) \right\}_{j \in s}$$

Weighting

- LPR is a linear estimator; construct n weights $\{\omega_{ij}\}$ for $i \in s, j \in s_i$
 - reflect design properties
 - incorporate auxiliary information
 - do not depend on a particular study variable
- For any study variable y , estimate

$$\theta_y = \frac{1}{N} \sum_{i \in C} \sum_{j \in U_i} \frac{y_{ij}}{N_i}$$

via

$$\hat{\theta}_y = \frac{1}{N} \sum_{i \in s} \sum_{j \in s_i} \omega_{ij} y_{ij}$$

Ex: Northeastern Lakes

- Survey of 20,000+ lakes in 8 Northeastern states
- More than 300 individual lakes were visited between one and six times.
- Many study variables stored in 32 data sets.
- x = longitude was used as the auxiliary information
- Each lake was treated as a cluster.
- Nonparametric regression estimates calculated for various chemistry variables.

Northeastern Lakes Lake Chemistry Means and Coefficients of Variation

Chemistry Measure	HT	REG	LPR
Log K	2.845 (8.33)	2.795 (3.04)	2.797 (2.79)
Log SO ₄	4.828 (7.37)	4.739 (1.75)	4.727 (1.69)
Log Ca	5.835 (7.76)	5.721 (1.73)	5.725 (1.69)
Log Cl	4.531 (8.63)	4.447 (3.67)	4.461 (3.43)
HCO ₃	522.8 (16.20)	520.1 (11.75)	518.1 (11.58)

- Average mean, with coefficients of variation underneath(%)
- Standard error estimates made using with replacement approximation.

Further Work

- Variance estimation (unequal probability EMAP samples)
- Efficient computation (binning, interpolation)
- Simulation studies
- Bandwidth selection