

Nonparametric Survey Regression Estimation in Spatial Sampling

Siobhan Everson-Stewart
Colorado State University

Joint work with:
F. Jay Breidt
Colorado State University

Jean D. Opsomer
Ji-Yeon Kim
Iowa State University

ASA CO-WY Chapter Spring Meeting
April 11, 2003

The work reported here was developed under STAR Research Assistance Agreements CR-829095 and CR-829096 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University and Oregon State University. This presentation has not been formally reviewed by EPA. The views expressed here are solely those of the authors. EPA does not endorse any products or commercial services mentioned in this report.

Modeling Environment

- Many samples include population auxiliary information in addition to sample data.
 - EMAP Northeastern Lakes Survey
 - Auxiliary information available for each lake (e.g. latitude, longitude, elevation)
 - Data for sampled lakes (e.g. chemistry)
- Statistical agency collects data (z) and auxiliary information (\mathbf{x})
- Data set released to end users
- Users must estimate status of many study variables

Modeling Constraints

- Want to use \mathbf{x} to improve estimates for z
- Limited time and other resources
- Potential controversy among end users
- Estimation strategy
 - should use information in $\mathbf{x}_i, i \in U$
 - 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

Spatial Model

- Consider \mathbf{x}_i to be the spatial location of each population element.
- Model: $z_i = \mu(x_i, y_i) + \epsilon_i$
- Mean model: $\mu(x_i, y_i) \equiv \text{constant}$
- Regression: $\mu(x_i, y_i) = \beta_0 + \beta_x x_i + \beta_y y_i$
- Spatial: $\mu(x_i, y_i) = \text{deterministic trend} + \text{spatially correlated process}$

Model-Assisted Estimators

$$\sum_{i \in U} \hat{\mu}_i + \sum_{i \in s} \frac{z_i - \hat{\mu}_i}{\pi_i}$$

- 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: $\hat{\mu}_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}}$
- Local Polynomial Regression: $\hat{\mu}_i$ comes from a kernel smooth (Breidt and Opsomer, 2000)

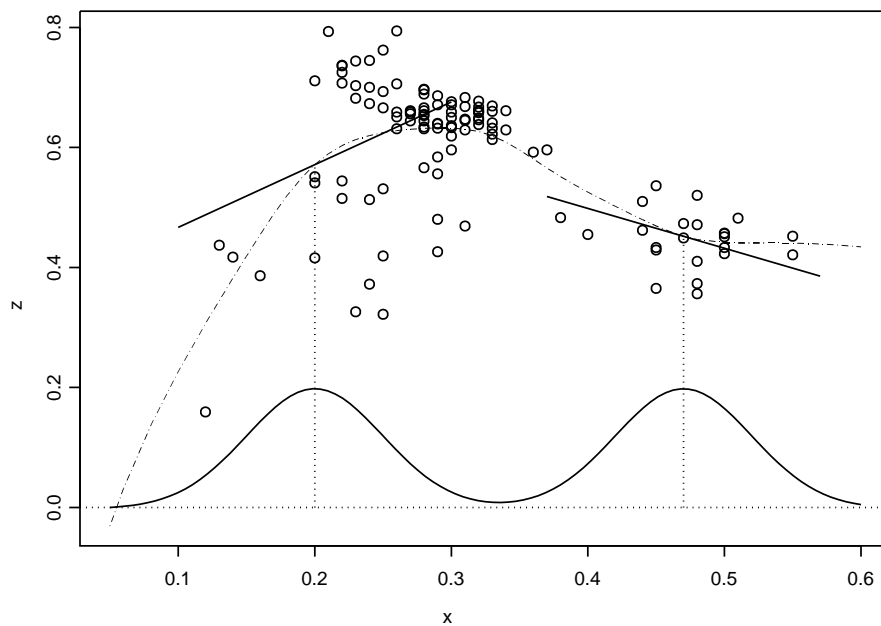
Kernel Smoothing

- Assumptions

$$z_i = \mu(\mathbf{x}_i) + v^{\frac{1}{2}}(\mathbf{x}_i)\epsilon_i$$

- $\mu(\mathbf{x}_i)$ is a smooth function of \mathbf{x}_i

- $v(\mathbf{x}_i)$, the variance of z_i , can vary with \mathbf{x}_i



Spatial Kernel Regression

- K is a two-dimensional function, e.g.
 - Product of one-dimensional kernel functions
 - Bivariate normal
- Bandwidth becomes a 2x2 matrix \mathbf{H} (symmetric, positive definite).
 - If \mathbf{H} is diagonal, h_{ii} is the bandwidth in the i th direction.
 - Non-diagonal matrices change the orientation of the smoothing weights.

Local Linear Regression Estimator

- Define \mathbf{X}_{si} , the local design matrix as

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

- The local weighting matrix is

$$\mathbf{W}_{si} = \text{Diag} \left\{ \frac{1}{\pi_j h_x h_y} K \left(\frac{x_j - x_i}{h_x}, \frac{y_j - y_i}{h_y} \right) \right\}_{j \in s}.$$

- Local linear regression estimate of the i th mean response:

$$\hat{\mu}_i = \mathbf{e}'_1 \left(\mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{X}_{si} \right)^{-1} \mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{z} = \mathbf{w}'_{si} \mathbf{z}$$

- Estimator of the population total:

$$\hat{t} = \sum_{i \in U} \hat{\mu}_i + \sum_{i \in s} \frac{z_i - \hat{\mu}_i}{\pi_i}$$

Simulation Study

- Compare design properties of HT, REG, and LPR estimators
- Want to sample repeatedly from a realization of a continuous, spatially correlated surface.
- Create a trend over space which includes planar and smooth but arbitrary portions

Simulated Population

- Lay down an 11x11 grid of points over the unit square.
- Surface at each grid point is

$$g(x_i, y_i) = \beta_0 + \beta_x x_i + \beta_y y_i + G(x_i, y_i),$$

where G is Gaussian with

$$\text{Cov}(G(x_i, y_i), G(x_j, y_j)) = \sigma^2 \exp \left\{ \log(\rho) \left((x_i - x_j)^2 + (y_i - y_j)^2 \right)^{\frac{1}{2}} \right\}.$$

- Thin plate spline used to interpolate g to get a continuous surface $\mu(x_i, y_i)$ over the unit square.
- Finally, add measurement noise

$$z(x_i, y_i) = \mu(x_i, y_i) + \epsilon_i$$

Thin Plate Splines

- Use penalized least squares to create a smooth, twice-continuously differentiable surface (Green and Silverman, 1994).
- Form of the interpolating spline:

$$\mu(\mathbf{x}) = \sum_{j=1}^{121} \delta_j \eta(\|\mathbf{x} - \mathbf{x}_j\|) + a_1 + a_2 x + a_3 y$$

where

$$\eta(\|\mathbf{x} - \mathbf{x}_j\|) = \frac{1}{16\pi} \|\mathbf{x} - \mathbf{x}_j\|^2 \log \|\mathbf{x} - \mathbf{x}_j\|^2$$

and the values of δ and \mathbf{a} are found by solving

$$\begin{bmatrix} \mathbf{E} & \mathbf{T}' \\ \mathbf{T} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \delta \\ \mathbf{a} \end{pmatrix} = \begin{pmatrix} \mathbf{z} \\ \mathbf{0} \end{pmatrix}$$

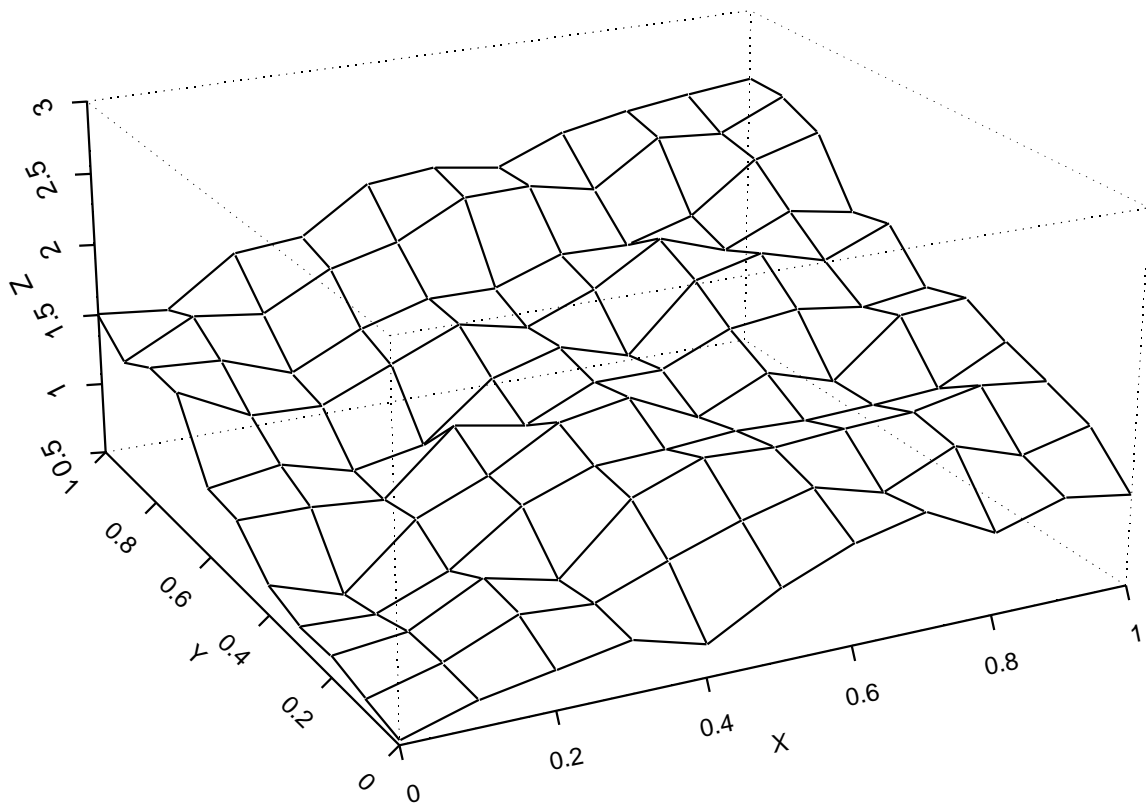
Also,

$$\begin{aligned} \mathbf{x}'_i &= (x_i, y_i) \\ \mathbf{T}' &= [1 \ x_i \ y_i]_{i=1}^{121} \\ \mathbf{E} &= [\eta(\|\mathbf{x}_i - \mathbf{x}_j\|)]_{i,j=1}^{121} \end{aligned}$$

Simulation Settings

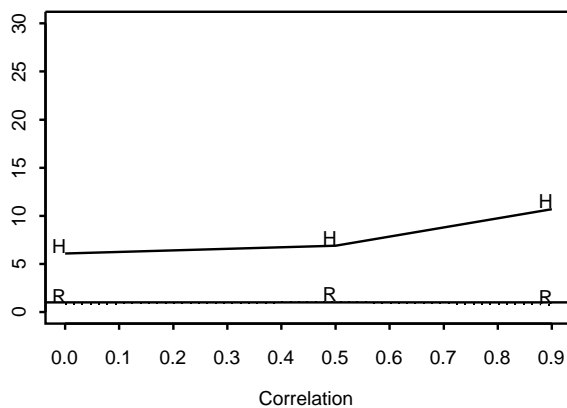
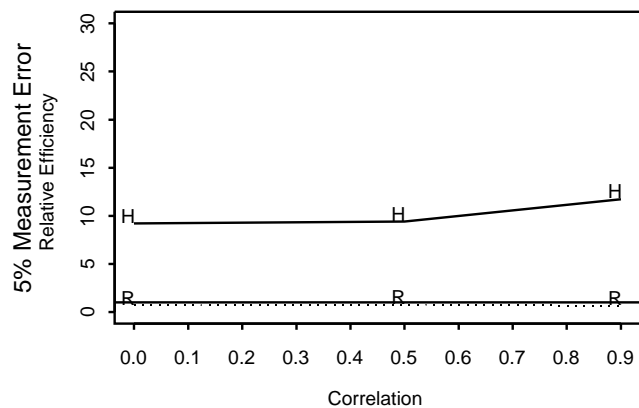
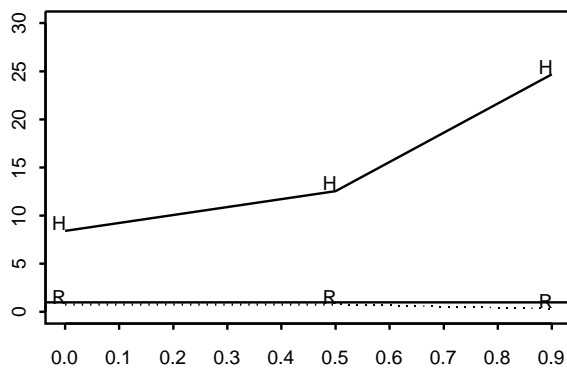
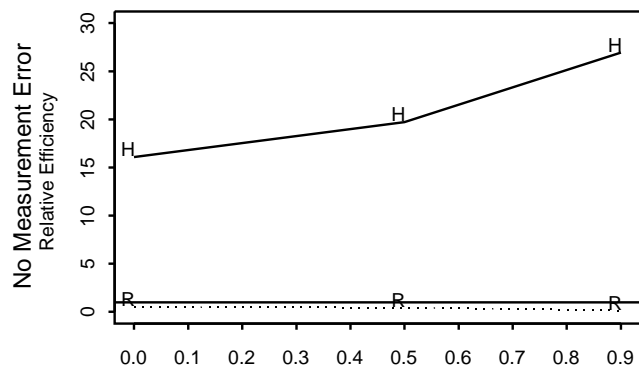
- Planar trend
 - $\beta_0 = 0.5$
 - $\beta_x = \beta_y = 1$ or $\beta_x = \beta_y = 0$
- Correlation (scaled to observations one grid unit apart)
 - $\rho = 0, 0.5, 0.9$
- Proportion of variation in $z(x_i, y_i)$ due to stochastic trend
$$\frac{\sigma^2}{\sigma^2 + \int_0^1 \int_0^1 (\mu(x, y))^2 dx dy - \left(\int_0^1 \int_0^1 \mu(x, y) dx dy\right)^2 + \nu^2} = .05 \text{ or } .15$$
- Measurement Error
 - none or 5% of total variation

Realization of Trend Surface



Simulation Results for Planar Trend

Efficiency of LPR Relative to HT and REG

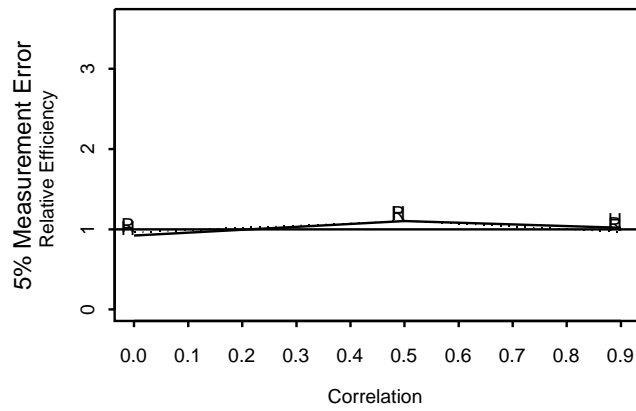
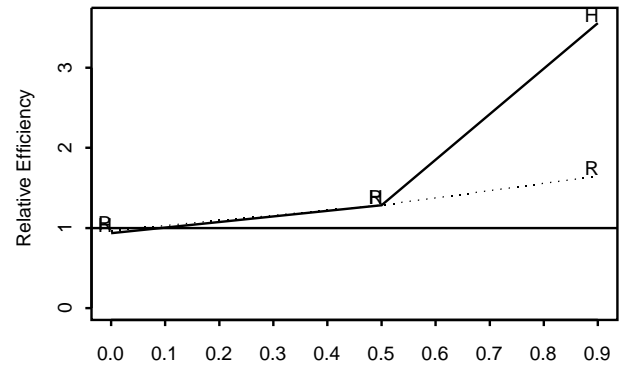
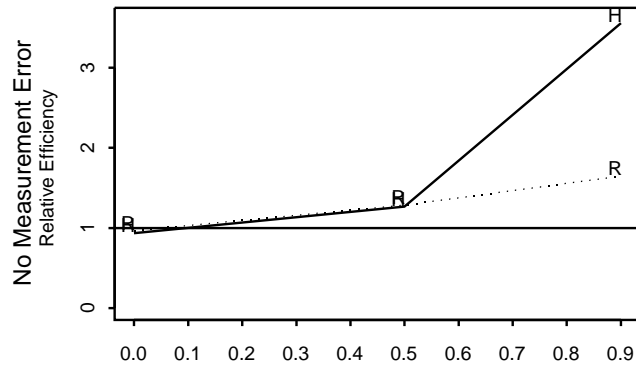


5% Stochastic Trend

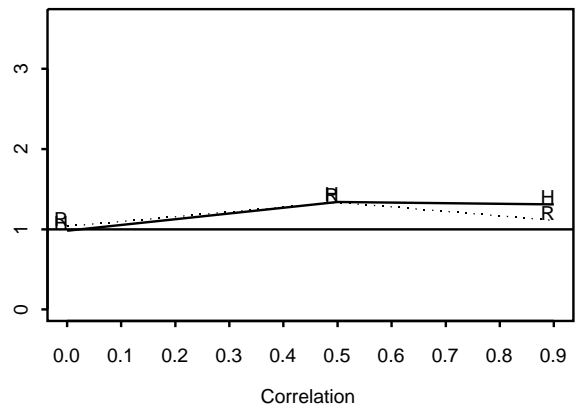
15% Stochastic Trend

Simulation Results for No Planar Trend

Efficiency of LPR Relative to HT and REG



5% Stochastic Trend



15% Stochastic Trend

Conclusions

- Local linear regression has many desirable qualities:
 - Incorporates auxiliary information into population status estimation
 - Few assumptions required; generally applicable
 - Easily applied to many study variables
- If data has planar trend, regression estimator does well, even with a stochastic component.
- If trend is nonlinear, local linear regression is more efficient than established estimators.