

**Nonparametric Model-Assisted Estimation  
for Overlapping Samples**

F. Jay Breidt  
Colorado State University

*TIES 2001: Environmetrics for Decision Making  
Portland, Oregon*

August 14, 2001

## Outline

- Finite population
  - complete coverage of auxiliary information
- Two surveys with possible overlap
  - sampling on two occasions: pure panel, rotating panel
  - two-phase sample
  - independent samples
- Model-assisted nonparametric regression estimators
  - semiparametric additive model
  - operational features
- Empirical results
  - simulation experiments
  - three-phase forest inventory example
- Summary and future work

## Auxiliary Information

- Finite population  $U_N = \{1, 2, \dots, N\}$
- Draw sample  $s \subset U_N$  (sample size  $n_s$ )
- Observe study variables,  $z_i, i \in s$
- Obtain complete auxiliary information  $x_i, i \in U_N$ 
  - spatial coordinates
  - elevation, slope from digital elevation model
  - remotely-sensed imagery
- Also have  $y_i, i \in r \subset U_N$  (sample size  $n_r$ )
- Modeling?

## 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,  $x$ , and maybe  $y$
- Limited time and other modeling resources
  - many study variables, unlimited derived variables:

$$z_1, z_2, z_1^2, z_1 z_2, \mathbf{1}_{\{z_1 \leq a\}}, z_1 \mathbf{1}_{\{z_1 \in D\}}$$

- Estimation strategy:
  - 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  $n_s$  weights  $\{\omega_{irs}\}$  for  $i \in s$ 
  - reflect design properties
  - incorporate auxiliary information and  $y$ -data
  - do not depend on study variables  $z$

- Release data set

$$\begin{bmatrix} \omega_{1rs} & z_{11} & \cdots & z_{1J} \\ \omega_{2rs} & z_{21} & \cdots & z_{2J} \\ \vdots & \vdots & & \vdots \\ \omega_{n_srs} & z_{n_s1} & \cdots & z_{n_sJ} \end{bmatrix}$$

- For any study variable  $z$ , estimate  $t_z = \sum_{i \in U_N} z_i$  via

$$\hat{t}_z = \sum_{i \in s} \omega_{irs} z_i$$

- Internally consistent:

$$\hat{t}_{z_1+z_2} = \sum_{i \in s} \omega_{irs} (z_{i1} + z_{i2}) = \sum_{i \in s} \omega_{irs} z_{i1} + \sum_{i \in s} \omega_{irs} z_{i2} = \hat{t}_{z_1} + \hat{t}_{z_2}$$

- Start with  $z$  only, then  $z$  and  $x$ , then  $z$  and  $x$  and  $y$

## Horvitz-Thompson Estimator

- Goal: Estimate  $t_z = \sum_{i \in U_N} z_i$
- Define  $I_{\{i \in s\}} = 1$  if  $i \in s$ , 0 otherwise:

$$E_p [I_{\{i \in s\}}] =: \pi_i \text{ and } E_p [I_{\{i \in s\}} I_{\{j \in s\}}] =: \pi_{ij}$$

- Design-unbiased estimator of  $t_z$  is

$$\hat{t}_{HT} = \sum_{i \in U_N} z_i \frac{I_{\{i \in s\}}}{\pi_i} = \sum_{i \in s} \frac{1}{\pi_i} z_i$$

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

- Variance is

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

- Unbiased variance estimator is

$$\hat{V}(\hat{t}_{HT}) = \sum_{i,j \in s} z_i z_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{1}{\pi_{ij}}$$

## Motivation for Model-Assisted Approach

- Start with non-random “guesses”  $z_i^0$  and form the difference estimator:

$$\hat{t}_{DIFF} = \sum_{i \in U_N} z_i^0 + \sum_{i \in U_N} (z_i - z_i^0) \frac{I_{\{i \in s\}}}{\pi_i}$$

– guess for total + design bias adjustment

- Design expectation is

$$E_p [\hat{t}_{DIFF}] = \sum_{i \in U_N} z_i^0 + \sum_{i \in U_N} (z_i - z_i^0) = t_z$$

- Design variance is

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

– if guesses are good, should have small design variance

## Model-Assisted Approach

- Working model,  $\xi$ :  $z_i$ 's are independent r.v.'s with model mean

$$E_{\xi} [z_i] = \alpha(x_i)$$

- Estimate parameters of model  $\xi$  and get model-based predictions  $\hat{a}_i$ 
  - finite population parameter “estimators”  $a_i$
  - sample estimators  $\hat{a}_i$

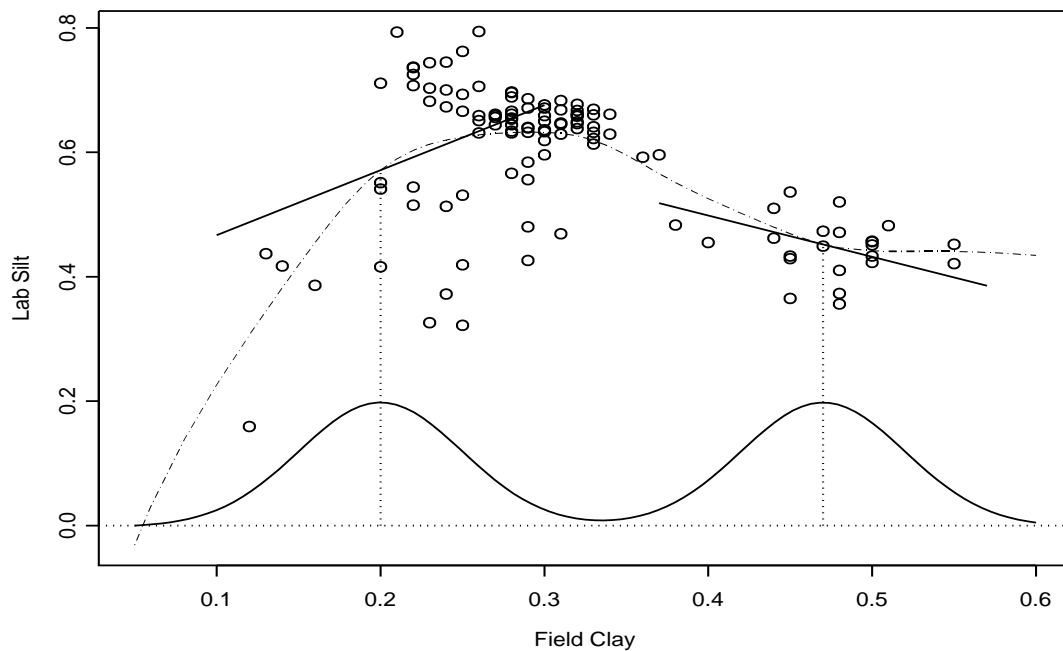
- Replace  $z_i^0$  in  $\hat{t}_{DIFF}$  with  $\hat{a}_i$ :

$$\hat{t}_z = \sum_{i \in U_N} \hat{a}_i + \sum_{i \in U_N} (z_i - \hat{a}_i) \frac{I_{\{i \in s\}}}{\pi_i}$$

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

## Local Polynomial Regression

- May want to avoid parametric modeling assumptions
  - nonparametric model,  $\xi: \alpha(x_i)$  is smooth
- Locally weighted least squares fits (Wand and Jones, 1995)



## Local Linear Regression Estimator

- Model:  $\alpha(x_i)$  is smooth,  $\text{Var}_\xi(z_i) = v(x_i)$  is positive and smooth
- Define

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

and

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

- Local linear regression estimator of  $\alpha(x_i)$  based on  $s$  is

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

- Local linear regression estimator of  $t_z$  is then

$$\hat{t}_{LLR} = \sum_{i \in U_N} \hat{a}_i + \sum_{i \in U_N} (z_i - \hat{a}_i) \frac{I_{\{i \in s\}}}{\pi_i}$$

(Breidt and Opsomer, 2000)

– goes to classical survey regression estimator as  $h \rightarrow \infty$

## Asymptotic Properties of Local Linear Estimator

- Asymptotic framework described in Breidt and Opsomer (2000)
- $\hat{t}_{LLR}$  is asymptotically design unbiased (ADU) and consistent for  $t_z$
- Design mean squared error:

$$E_p \left( \frac{\hat{t}_{LLR} - t_z}{N} \right)^2 = \frac{1}{N^2} \sum_{i,j \in U_N} (z_i - a_i)(z_j - a_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} + o(n_N^{-1})$$

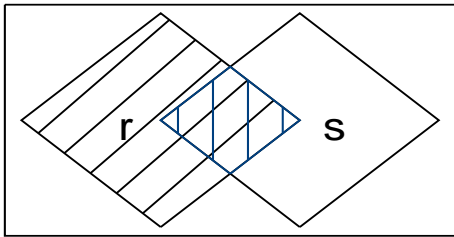
- ADU and design consistent variance estimator is

$$\hat{V}(N^{-1}\hat{t}_{LLR}) = \frac{1}{N^2} \sum_{i,j \in U_N} (z_i - \hat{a}_i)(z_j - \hat{a}_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{I_{\{i \in s\}} I_{\{j \in s\}}}{\pi_{ij}}$$

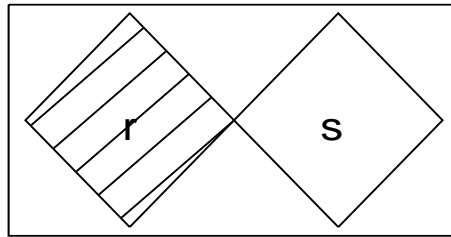
## Possibly Overlapping Samples

- Two samples  $s$  and  $r$ :

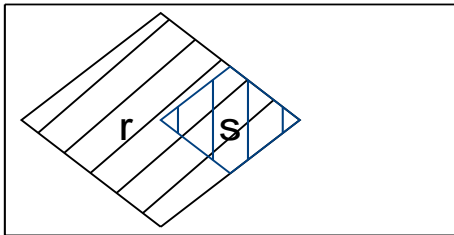
general overlapping



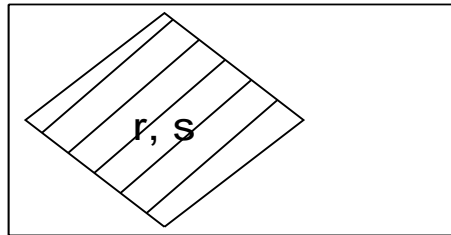
negatively coordinated



two-phase



pure panel



- Start with two-phase:  $\{x_j, j \in U\}$ ;  $\{y_j, j \in r\}$ ;  $\{z_j, j \in s \subset r\}$

## Semiparametric Additive Model

- Model for  $y$  on  $x$  is nonparametric:

$$y_j = \mu_0 + \mu(x_j) + \delta_j, \quad \{\delta_j\} \text{ iid } (0, \sigma_\delta^2)$$

where  $\mu(\cdot)$  is smooth

- Model for  $z$  on  $y$  is parametric,  $z$  on  $x$  is nonparametric:

$$\begin{aligned} z_j &= \alpha(x_j) + (y_j - \mu_0)\beta + \epsilon_j, \quad \{\epsilon_j\} \text{ iid } (0, \sigma_\epsilon^2) \\ &= \alpha(x_j) + \beta\mu(x_j) + (\beta\delta_j + \epsilon_j) \end{aligned}$$

where  $\alpha(\cdot)$  is smooth (Speckman, 1988; Opsomer and Ruppert, 1999)

- Two-phase semiparametric regression estimator

$$\hat{t}_1 = \sum_{j \in U_N} (\hat{a}_j + \hat{B}\hat{m}_j) + \hat{B} \sum_{j \in r} \frac{y_j - \hat{m}_0 - \hat{m}_j}{\rho_j} + \sum_{j \in s \cap r} \frac{z_j - \hat{a}_j - \hat{B}(y_j - \hat{m}_0)}{\rho_j \pi_{j|r}}$$

## Estimation Details

- Choose bandwidths  $h_y, h_z$  and compute smoothing vectors

$$\mathbf{s}'_{gj} = [1, 0] \left( \mathbf{X}'_{gj} \mathbf{W}_{gj} \mathbf{X}_{gj} \right)^{-1} \mathbf{X}'_{gj} \mathbf{W}_{gj}$$

for  $g = y, z$

- In the model  $y_j = \mu_0 + \mu(x_j) + \delta_j$ ,
  - estimate  $\mu_0$  by  $\hat{m}_0 = \bar{y}_r$
  - estimate  $\mu(x_j)$  by centering the smooths  $\mathbf{s}'_{yj} \mathbf{y}_r$
- In the model  $z_j = \alpha(x_j) + (y_j - \mu_0)\beta + \epsilon_j$ ,
  - estimate  $\beta$  using the smoother matrix  $\mathbf{S} = [\mathbf{s}'_{zj}]_{j \in s}$ :

$$\hat{B} = \{ \mathbf{d}'_s (\mathbf{I} - \mathbf{S}) \mathbf{d}_s \}^{-1} \mathbf{d}'_s (\mathbf{I} - \mathbf{S}) \mathbf{z}_s$$

where  $\mathbf{d}_s = [d_j]_{j \in s}$  and

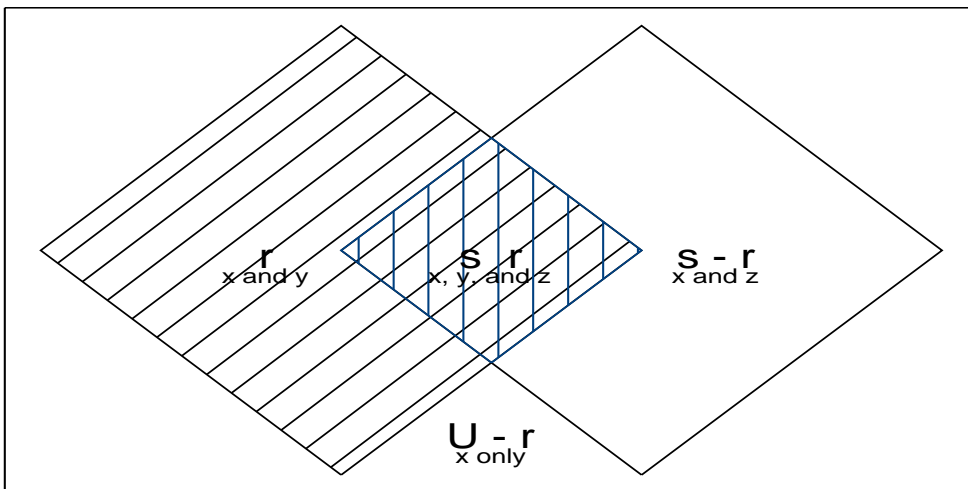
$$d_j = \begin{cases} y_j - \bar{y}_r, & j \in r \\ \hat{m}_j, & j \in s - r \end{cases}$$

- estimate  $\alpha(x_j)$  by  $\hat{a}_j = \mathbf{s}'_{zj} (\mathbf{z}_s - \mathbf{d}_s \hat{B})$

## General Overlapping Case

- From  $r$ , subsample  $s \cap r$  and use  $\hat{t}_1$
- From  $U - r$ , subsample  $s - r$  and use

$$\hat{t}_2 = \sum_{j \in U_N} (\hat{a}_j + \hat{B}\hat{m}_j) + \hat{B} \sum_{j \in U-r} \frac{\hat{m}_j - \hat{m}_j}{1 - \rho_j} + \sum_{j \in s-r} \frac{z_j - \hat{a}_j - \hat{B}\hat{m}_j}{(1 - \rho_j)\pi_{j|\bar{r}}}$$



## Combined Estimator

- Estimate  $t_z$  by

$$\hat{t}_{SAM} = \lambda \hat{t}_1 + (1 - \lambda) \hat{t}_2$$

- Optimal linear combination is given by

$$\lambda_{\text{opt}}^* = \frac{\text{Var}(\hat{t}_2) - \text{Cov}(\hat{t}_1, \hat{t}_2)}{\text{Var}(\hat{t}_1) + \text{Var}(\hat{t}_2) - 2\text{Cov}(\hat{t}_1, \hat{t}_2)}$$

- minimal variance is then

$$\frac{\text{Var}(\hat{t}_1) \text{Var}(\hat{t}_2) - \text{Cov}(\hat{t}_1, \hat{t}_2)^2}{\text{Var}(\hat{t}_1) + \text{Var}(\hat{t}_2) - 2\text{Cov}(\hat{t}_1, \hat{t}_2)}$$

- Ad hoc linear combination:  $\lambda = |s \cap r|/|s|$  = overlap fraction
  - does not depend on  $z$
  - goes to two-phase estimator as  $|s \cap r| \rightarrow |s|$ , one-phase as  $|s \cap r| \rightarrow 0$

## Variance Estimation

- Under simple random sampling, with  $n_s = n_r = n$ ,  $f = n/N$ , and  $\phi =$  matched fraction:

$$\text{Var}(\hat{t}_{SAM}) = \lambda^2 \text{Var}(\hat{t}_1) + (1 - \lambda)^2 \text{Var}(\hat{t}_2) + 2\lambda(1 - \lambda) \text{Cov}(\hat{t}_1, \hat{t}_2)$$

where

$$\text{Var}(\hat{t}_1) \simeq \frac{N^2}{n}(1 - f)S_{e^*U}^2 + \frac{N^2}{\phi n}(1 - \phi)S_{eU}^2$$

$$\text{Var}(\hat{t}_2) \simeq \frac{N^2}{(1 - f)(1 - \phi)n} \{fn(1 - \phi) + (1 - 2f + \phi f)\} S_{e^*U}^2$$

$$\text{Cov}(\hat{t}_1, \hat{t}_2) \simeq -NS_{e^*U}^2$$

$S_{e^*U}^2$  is the finite population variance of  $\{e_j^*\} = \{z_j - a_j - Bm_j\}$ , and  $S_{eU}^2$  is the finite population variance of  $\{e_j\} = \{z_j - a_j - B(y_j - m_0)\}$

- Estimate by substituting sample variances of  $\{\hat{e}_j^*\}$ ,  $\{\hat{e}_j\}$

## Weighting and Calibration

- Easy to show that

$$\hat{t}_{SAM} = \sum_{j \in s} \omega_{jrs} z_j$$

where  $\omega_{jrs}$  do not depend on  $z$  if  $\lambda$  does not

- **one set of weights** can be applied to any study variable

- Calibration: weighted sample  $x$ -sums equal population  $x$ -sums

$$\sum_{j \in s} \omega_{jrs} 1 = N, \quad \sum_{j \in s} \omega_{jrs} x_j = \sum_{j \in U_N} x_j$$

- desirable property for survey weights (Deville and Särndal, 1992)

- if  $z_j = \alpha_0 + \alpha_1 x_j$ , then  $\hat{t}_{SAM} = t_z$  for every possible sample

- Weighted sample “ $d$ ”-sum produces shrunken version of  $\hat{t}_{LLR}$ :

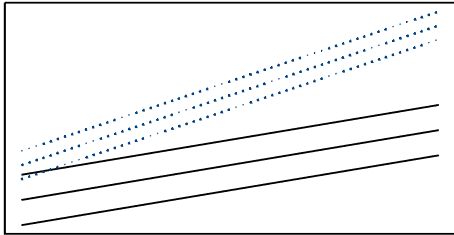
$$\hat{t}_1 = \hat{t}_{LLR}, \quad \hat{t}_2 = \sum_{j \in U_N} (\hat{m}_0 + \hat{m}_j),$$

$$\hat{t}_{SAM} = \sum_{j \in U_N} (\hat{m}_0 + \hat{m}_j) + \lambda \sum_{j \in r} \frac{y_j - \hat{m}_0 - \hat{m}_j}{\rho_j}$$

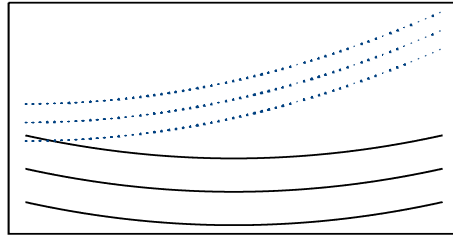
## Simulation Experiment

- $N = 1000$ ,  $\sigma_\delta = 0.3$ ,  $\sigma_\epsilon = 0.2$ ,  $n = 100$ , 50% overlap

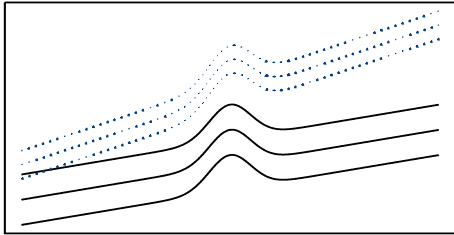
line



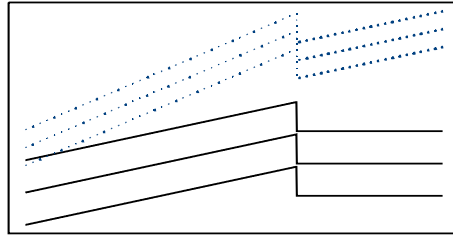
quad



bump



jump



## Simulation Results: Correct Parametric Specification

- For  $h_y = h_z = 0.1$  and 1000 reps of simple random sampling:

	% Relative Bias			Relative Efficiency	
	HT	REG	SAM	HT	REG
$y$	0.05	0.00	-0.02	4.10	0.92
line	0.01	-0.02	-0.03	8.15	0.81
quad	0.01	-0.08	0.02	2.91	1.08
bump	0.02	0.08	-0.01	8.12	1.48
jump	-0.04	0.09	-0.07	4.18	1.50
line- $y$	-0.08	-0.05	-0.06	10.82	0.68
quad- $y$	-0.06	-0.22	0.10	5.08	1.14
bump- $y$	-0.04	0.22	0.01	10.51	1.61
jump- $y$	-0.30	0.38	-0.23	6.40	1.70

– Percent relative design bias,  $(E_p [\hat{t} - t_y] / t_y) \times 100\%$

– SAM relative efficiency,  $\text{Var}_p(\hat{t}) / \text{Var}_p(\hat{t}_{SAM})$

## Simulation Results: Incorrect Parametric Specification

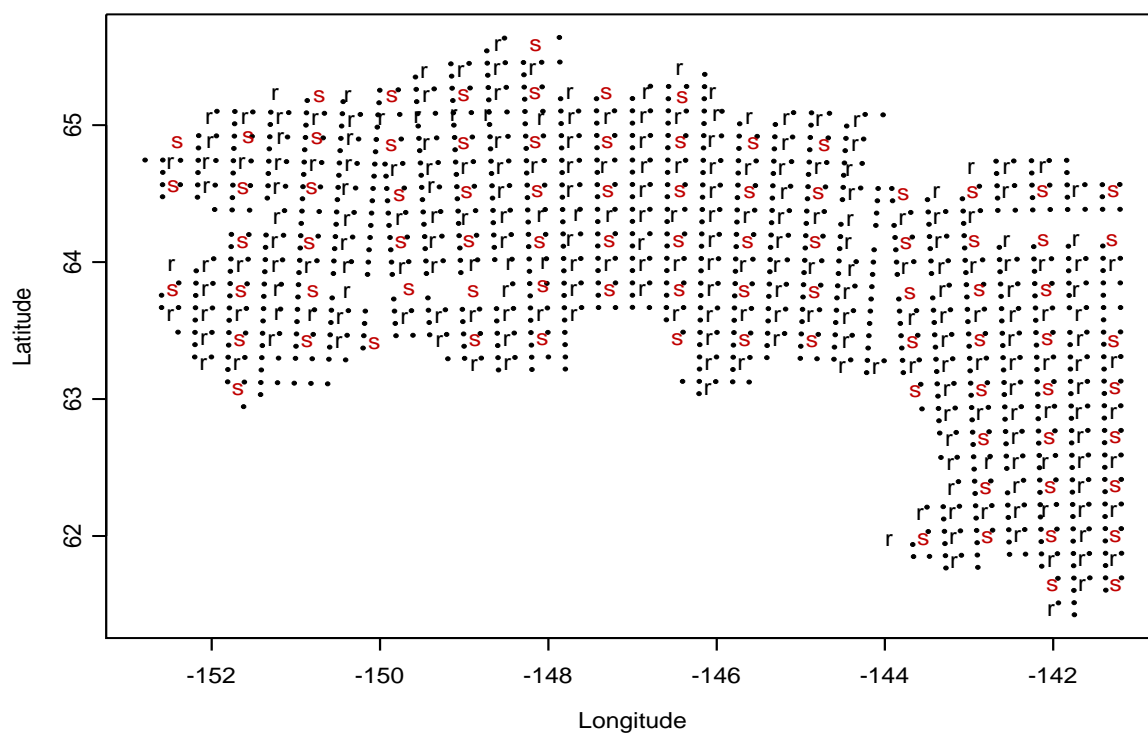
- $E[z_j]$  contains interaction term  $3(x_j - 0.5)^2 y_j$ .
- For  $h_y = h_z = 0.1$  and 1000 reps of simple random sampling:

	% Relative Bias			Relative Efficiency	
	HT	REG	SAM	HT	REG
$y$	0.05	0.00	-0.02	4.10	0.92
line	0.06	-0.19	0.08	9.12	2.41
quad	0.06	-0.23	0.13	5.00	3.62
bump	0.07	-0.10	0.10	9.07	1.98
jump	0.03	-0.11	0.06	4.81	1.37
line- $y$	0.07	-0.43	0.22	13.32	3.48
quad- $y$	0.07	-0.51	0.30	8.03	5.54
bump- $y$	0.09	-0.21	0.24	13.28	2.70
jump- $y$	-0.01	-0.31	0.20	7.18	1.68

- Percent relative design bias,  $(E_p [\hat{t} - t_y] / t_y) \times 100\%$
- SAM relative efficiency,  $\text{Var}_p(\hat{t}) / \text{Var}_p(\hat{t}_{SAM})$

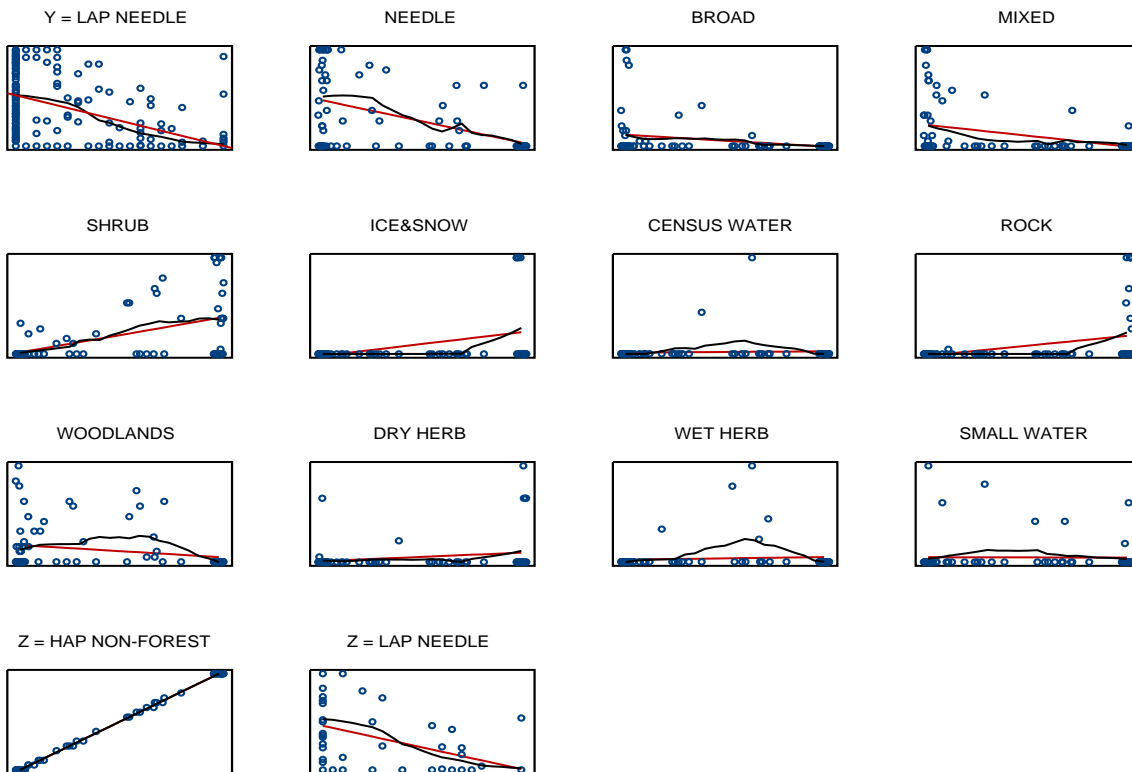
## Example: Three-phase Forest Inventory

- High-altitude photos ( $N = 1290$ ), low-altitude photos ( $n_r = 323$ ), ground visits ( $n_s = 86$ ) [Breidt and Fuller, 1993]



## Example: Three-phase Forest Inventory, Continued

- Model:  $z_j = \alpha(x_j) + \beta(y_j - \mu_0) + \epsilon_j$ ,  $y_j = \mu_0 + \mu(x_j) + \delta_j$   
–  $x_j$  = HAP proportion of non-forest;  $y_j$  = LAP needle-leaf forest



### Example: Three-phase Forest Inventory, Continued

- $h_y = 0.2, h_z = 0.3$
- Systematic sampling treated as simple random.

	Percent Cover			Relative Efficiency	
	HT	REG	SAM	HT	REG
NEEDLE	29.29	32.72	32.81	2.29	1.01
BROAD	6.62	6.33	6.31	1.12	1.01
MIXED	12.56	11.71	11.40	1.35	1.02
SHRUB	16.49	15.83	15.82	1.39	1.01
ICE & SNOW	8.14	7.77	7.51	1.25	1.06
CENSUS WATER	0.62	0.56	0.63	1.11	1.10
ROCK	6.85	6.54	6.32	1.25	1.06
WOODLAND	12.33	11.81	12.33	1.26	1.13
DRY HERB	4.34	4.08	4.00	1.08	1.02
WET HERB	1.53	1.48	1.63	1.35	1.34
SMALL WATER	1.23	1.17	1.25	1.05	1.03

## Summary

- Overlapping samples with complete auxiliary data
  - two-phase, three-phase, independent samples
  - sampling on two occasions: pure panel, rotating panel
- Semiparametric additive model-assisted survey estimation
  - parametric model:  $z$  on  $y$ ; nonparametric models:  $y$  on  $x$ ,  $z$  on  $x$
  - variance approximation, weighting and calibration
- Good performance in simulations and example
  - some efficiency loss when linear specification is correct
  - potentially large efficiency gains when linear specification is incorrect
- Further work
  - asymptotics, bandwidth selection
  - additional samples, additional nonparametric regressors
  - optimal matching fractions