

Theory and Methods of Nonparametric Survey Regression Estimation

Jean Opsomer
Iowa State University

Jay Breidt
Colorado State University

June 21, 2004



Back

Close

Outline

1. Introduction
2. Generic estimation for surveys
3. Nonparametric model-assisted estimation
4. From theory to applications
 - (a) National Resources Inventory (NRI)
 - (b) Forest Inventory and Analysis (FIA)
5. Smoothing parameter selection
6. Conclusion

[Back](#)[Close](#)

1. Introduction: Statistical Inference

- Specific Inference:
 - expensive, high quality, targeted
 - using “custom-built” method (or model) to achieve best possible estimator for particular variable(s)
 - willing to defend model



Back

Close

1. Introduction: Statistical Inference

- Specific Inference:
 - expensive, high quality, targeted
 - using “custom-built” method (or model) to achieve best possible estimator for particular variable(s)
 - willing to defend model

- Generic Inference:
 - cheap, reasonable quality, good for many purposes
 - using method appropriate for large number of variables that need to be estimated jointly



Corn
Flakes

NET WT. 12 OZ.



Back

Close

Statistical Inference in Surveys

Number of observations	Inference	Modelling
Large	generic	none
Moderate	{ generic specific	model-assisted model-based
Small	specific	small area estimation

- “Number of observations” depends on domain (subpopulation) size
- For moderate sample size, use of generic inference depends on model goodness-of-fit

Nonparametric methods can improve generic inference



Back

Close

2. Generic Estimation for Survey Data

- Population $U = \{1, \dots, i, \dots, N\}$ with unknown population “parameters”

$$\bar{y}_N = \frac{1}{N} \sum_U y_i \quad \text{and} \quad \bar{z}_N, \bar{x}_N, \dots$$

- Sample s selected from U according to known sampling design $p(s)$
 - stratification
 - clustering
 - multiple phases



Back

Close

2. Generic Estimation for Survey Data

- Population $U = \{1, \dots, i, \dots, N\}$ with unknown population “parameters”

$$\bar{y}_N = \frac{1}{N} \sum_U y_i \quad \text{and} \quad \bar{z}_N, \bar{x}_N, \dots$$

- Sample s selected from U according to known sampling design $p(s)$
 - stratification
 - clustering
 - multiple phases
- Generic estimator for the population mean

$$\hat{y}_s = \sum_s w_i y_i \quad \left[\hat{z}_s = \sum_s w_i z_i, \quad \hat{x}_s, \dots \right]$$



Back

Close

Generic Estimation for Surveys: Properties

The “ideal” generic estimator would have the following properties

1. easy to compute
2. applicable to large numbers of variables
3. local/scale invariant

$$z_i = a + by_i \Rightarrow \hat{z}_s = a + b\hat{y}_s$$

4. additive

$$U = \{U_1, U_2\} \Rightarrow N\hat{y}_s = N_1\hat{y}_{s1} + N_2\hat{y}_{s2}$$

5. calibrated

$$\hat{x}_s = \bar{x}_N \text{ for known population quantities } x$$

6. precise (low bias, low variance, consistent,...)



Back

Close

Simple Generic Estimation: Design-based

- Horvitz-Thompson estimator (1952)

$$\hat{y}_{HT} = \frac{1}{N} \sum_s \frac{1}{\pi_i} y_i$$

with *inclusion probabilities* $\pi_i = \Pr(i \in s)$

- Hájek estimator (1971)

$$\hat{y}_{HA} = \frac{\sum_s \frac{1}{\pi_i} y_i}{\sum_s \frac{1}{\pi_i}}$$



Back

Close

Better Generic Estimation: Model-assisted

- Superpopulation model ξ : y_i are iid with

- $E_{\xi}(y_i) = \beta_0 + \beta_1 x_i = \mathbf{x}_i^T \boldsymbol{\beta}$

- $\text{Var}_{\xi}(y_i) = \sigma^2$

- Least squares population fit for $\boldsymbol{\beta}$

$$\mathbf{B}_U = (\mathbf{X}_U^T \mathbf{X}_U)^{-1} \mathbf{X}_U \mathbf{Y}_U$$



Back

Close

Better Generic Estimation: Model-assisted

- Superpopulation model ξ : y_i are iid with

- $E_{\xi}(y_i) = \beta_0 + \beta_1 x_i = \mathbf{x}_i^T \boldsymbol{\beta}$

- $\text{Var}_{\xi}(y_i) = \sigma^2$

- Least squares population fit for $\boldsymbol{\beta}$

$$\mathbf{B}_U = (\mathbf{X}_U^T \mathbf{X}_U)^{-1} \mathbf{X}_U \mathbf{Y}_U$$

- \mathbf{B}_U is estimated by sample-based estimator

$$\hat{\mathbf{B}} = (\mathbf{X}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{Y}_s$$

with $\boldsymbol{\Pi}_s = \text{diag}\{\pi_i, i \in s\}$

- GREG: Model-assisted estimator (Cassel *et al.*, 1977)

$$\hat{y}_{REG} = \frac{1}{N} \sum_U \mathbf{x}_i^T \hat{\mathbf{B}} + \frac{1}{N} \sum_s \frac{y_i - \mathbf{x}_i^T \hat{\mathbf{B}}}{\pi_i}$$



Back

Close

Properties of Regression Estimator

- Generic estimator

$$\hat{y}_{REG} = \sum_s w_i(s) y_i$$

- Consistent, asymptotically design unbiased

$$E_p(\hat{y}_{REG}) \approx \bar{y}_N$$

- Approximate design variance

$$\text{Var}_p(\hat{y}_{REG}) \approx \frac{1}{N^2} \sum \sum_U \frac{y_i - \mathbf{x}_i^T \mathbf{B}_U}{\pi_i} \frac{y_j - \mathbf{x}_j^T \mathbf{B}_U}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$



Back

Close

Properties of Regression Estimator

- Generic estimator

$$\hat{y}_{REG} = \sum_s w_i(s) y_i$$

- Consistent, asymptotically design unbiased

$$E_p(\hat{y}_{REG}) \approx \bar{y}_N$$

- Approximate design variance

$$\text{Var}_p(\hat{y}_{REG}) \approx \frac{1}{N^2} \sum_U \sum_U \frac{y_i - \mathbf{x}_i^T \mathbf{B}_U}{\pi_i} \frac{y_j - \mathbf{x}_j^T \mathbf{B}_U}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$

- Calibration

$$\hat{\mathbf{x}}_{REG} = \sum_s w_i(s) \mathbf{x}_i = \bar{\mathbf{x}}_N$$

- Location/scale invariance, additivity,...



3. Nonparametric Regression Estimation?

- Superpopulation model ξ :

- $E_{\xi}(y_i) = \mathbf{x}_i^T \boldsymbol{\beta}$

- $\text{Var}_{\xi}(y_i) = \sigma^2$



Back

Close

3. Nonparametric Regression Estimation?

- Superpopulation model ξ :

- $E_{\xi}(y_i) = \mathbf{x}_i^T \boldsymbol{\beta}$

- $\text{Var}_{\xi}(y_i) = \sigma^2$

Replace by:

- Superpopulation model ξ :

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

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



Back

Close

Nonparametric Model-assisted Estimator

- Superpopulation model ξ :
 - $E_{\xi}(y_i) = m(x_i)$
 - $\text{Var}_{\xi}(y_i) = v(x_i)$
- Population fit for $m(\cdot)$ at $x_i, i \in U$

$$m_i = \mathbf{s}_{Ui} \mathbf{Y}_U$$



Back

Close

Nonparametric Model-assisted Estimator

- Superpopulation model ξ :
 - $E_{\xi}(y_i) = m(x_i)$
 - $\text{Var}_{\xi}(y_i) = v(x_i)$
- Population fit for $m(\cdot)$ at $x_i, i \in U$

$$m_i = \mathbf{s}_{Ui} \mathbf{Y}_U$$

- The $m_i, i \in U$ are estimated by design-weighted estimators

$$\hat{m}_i = \mathbf{s}_{si} \mathbf{Y}_s$$

- Model-assisted estimator

$$\hat{y}_{NP} = \frac{1}{N} \sum_U \hat{m}_i + \frac{1}{N} \sum_s \frac{y_i - \hat{m}_i}{\pi_i}$$



Back

Close

Nonparametric Model-assisted Estimator (2)

- Theoretical properties derived for
 - kernel-based methods (Breidt and Opsomer, 2000)
 - spline-based methods (Breidt, Claeskens and Opsomer, 2003)
- Nonparametric model-assisted estimator has same design properties as GREG
 - weighted (generic) form $\hat{y}_{NP} = \sum_s w_i(s)y_i$
 - design consistency, variance
 - calibration, invariance



Back

Close

Nonparametric Model-assisted Estimator (2)

- Theoretical properties derived for
 - kernel-based methods (Breidt and Opsomer, 2000)
 - spline-based methods (Breidt, Claeskens and Opsomer, 2003)
- Nonparametric model-assisted estimator has same design properties as GREG
 - weighted (generic) form $\hat{y}_{NP} = \sum_s w_i(s)y_i$
 - design consistency, variance
 - calibration, invariance
- Differences with GREG
 - requires continuous auxiliary variable, available for all $i \in U$
 - smoothing parameter selection



Back

Close

Efficiency Gains from Modelling

$$\text{Var}_p(\hat{y}_{HT}) = \frac{1}{N^2} \sum \sum_U \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$

$$\text{Var}_p(\hat{y}_{HA}) \approx \frac{1}{N^2} \sum \sum_U \frac{y_i - \bar{y}_N}{\pi_i} \frac{y_j - \bar{y}_N}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$



Back

Close

Efficiency Gains from Modelling

$$\text{Var}_p(\hat{y}_{HT}) = \frac{1}{N^2} \sum \sum_U \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$

$$\text{Var}_p(\hat{y}_{HA}) \approx \frac{1}{N^2} \sum \sum_U \frac{y_i - \bar{y}_N}{\pi_i} \frac{y_j - \bar{y}_N}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$

$$\text{Var}_p(\hat{y}_{REG}) \approx \frac{1}{N^2} \sum \sum_U \frac{y_i - \mathbf{x}_i^T \mathbf{B}_U}{\pi_i} \frac{y_j - \mathbf{x}_j^T \mathbf{B}_U}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$

$$\text{Var}_p(\hat{y}_{NP}) \approx \frac{1}{N^2} \sum \sum_U \frac{y_i - m_i}{\pi_i} \frac{y_j - m_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$



Back

Close

4. From Theory to Applications...

- Adapt estimator to more complex designs
 - multi-stage
 - multi-phase
- ⇒ possible in model-assisted context



Back

Close

4. From Theory to Applications...

- Adapt estimator to more complex designs
 - multi-stage
 - multi-phase

⇒ possible in model-assisted context
- Extend model to incorporate different data types and multiple auxiliary variables
 - semiparametric models
 - multivariate smoothing techniques

⇒ wide range of nonparametric methods available



Back

Close

4. From Theory to Applications...

- Adapt estimator to more complex designs
 - multi-stage
 - multi-phase

⇒ possible in model-assisted context
- Extend model to incorporate different data types and multiple auxiliary variables
 - semiparametric models
 - multivariate smoothing techniques

⇒ wide range of nonparametric methods available
- Smoothing parameter selection
- Variance estimation

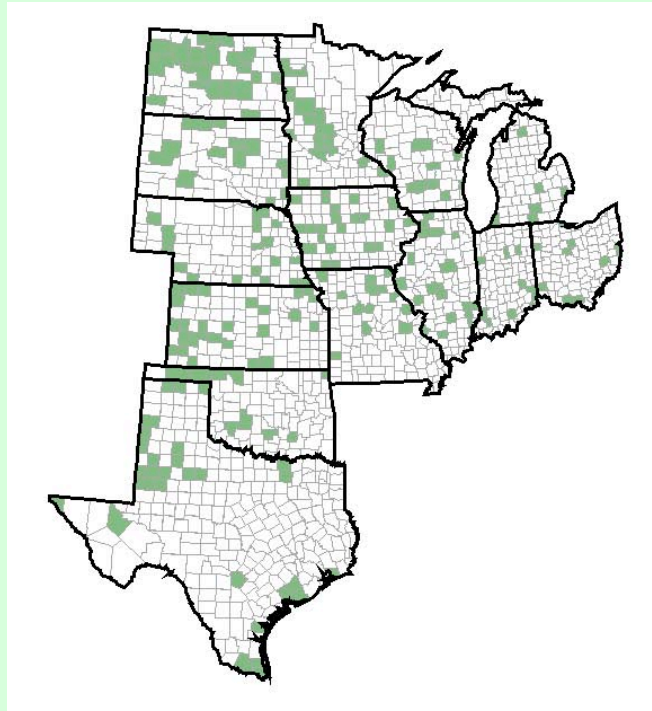


Back

Close

Application 1: 1995 NRI Special Study

Two-stage survey of agricultural lands with 1992 National Resources Inventory as sampling frame



Back

Close

NRI and 1995 Special Study

- National Resources Inventory is stratified longitudinal survey of non-federal land conducted by Natural Resources Conservation Service (USDA)
- sampling units are 160-acre plots of land, and points within plots
- 1992 NRI contains 300,000 plots



Back

Close

NRI and 1995 Special Study

- National Resources Inventory is stratified longitudinal survey of non-federal land conducted by Natural Resources Conservation Service (USDA)
- sampling units are 160-acre plots of land, and points within plots
- 1992 NRI contains 300,000 plots
- 1995 NRI Special Study is sample of 1900 plots obtained by stratified two-stage sampling
 - states are strata (14)
 - PSUs are counties (1357 total, 213 selected)
 - PSU selection probabilities are proportional to measure of erosion potential in county
 - variables of interest: water erosion (USLE), wind erosion (WEQ)



Back

Close

Estimator in two-stage sampling

- Usual case: auxiliary information x available for PSUs only
- Superpopulation model ξ for t_i (cluster total)
 - $E_{\xi}(t_i) = m(x_i)$
 - $\text{Var}_{\xi}(t_i) = v(x_i)$



Back

Close

Estimator in two-stage sampling

- Usual case: auxiliary information x available for PSUs only
- Superpopulation model ξ for t_i (cluster total)
 - $E_{\xi}(t_i) = m(x_i)$
 - $\text{Var}_{\xi}(t_i) = v(x_i)$
- x = square root of measure of erosion potential
- Model-assisted estimator

$$\hat{y}_{NP} = \frac{1}{N} \sum_{U_I} \hat{m}_i + \frac{1}{N} \sum_{S_I} \frac{\hat{t}_i - \hat{m}_i}{\pi_{Ii}}$$

with $\hat{t}_i = \sum_{s_i} y_{ki} / \pi_{k|i}$, and \hat{m}_i obtained from local linear regression of \hat{t}_i on $x_i, i \in S_I$

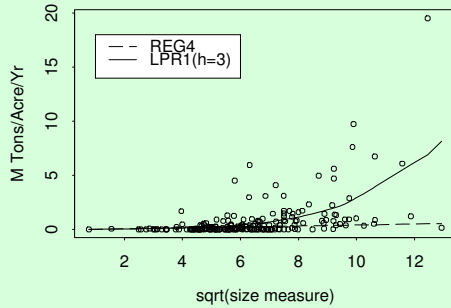


Back

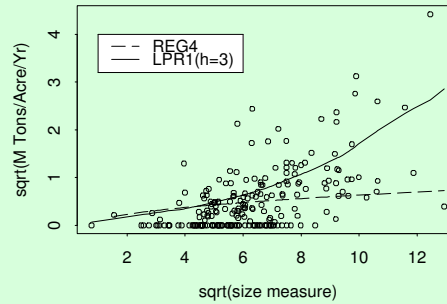
Close

Nonparametric fits

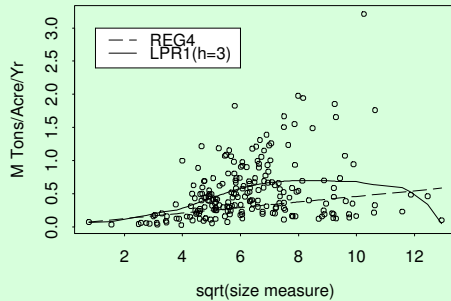
WEQ



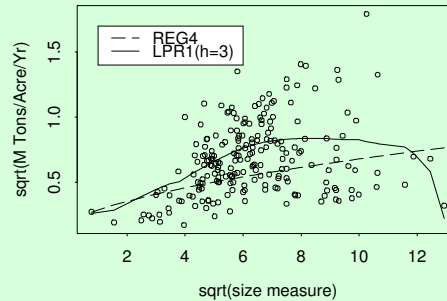
Transformed WEQ



USLE



Transformed USLE



Estimates

		WEQ	USLE
HT		443.6 (49.4)	551.5 (31.8)
REG2	$\nu(x) \propto x^2$	442.5 (50.7)	537.8 (26.5)
REG4	$\nu(x) \propto x^4$	442.1 (50.1)	537.7 (26.5)
REG8	$\nu(x) \propto x^8$	441.8 (50.3)	540.1 (27.6)
LPR1	h=1	434.1 (47.5)	529.0 (24.4)
LPR1	h=3	427.4 (48.9)	532.3 (25.3)
LPR1	h=5	430.5 (48.7)	541.2 (27.6)

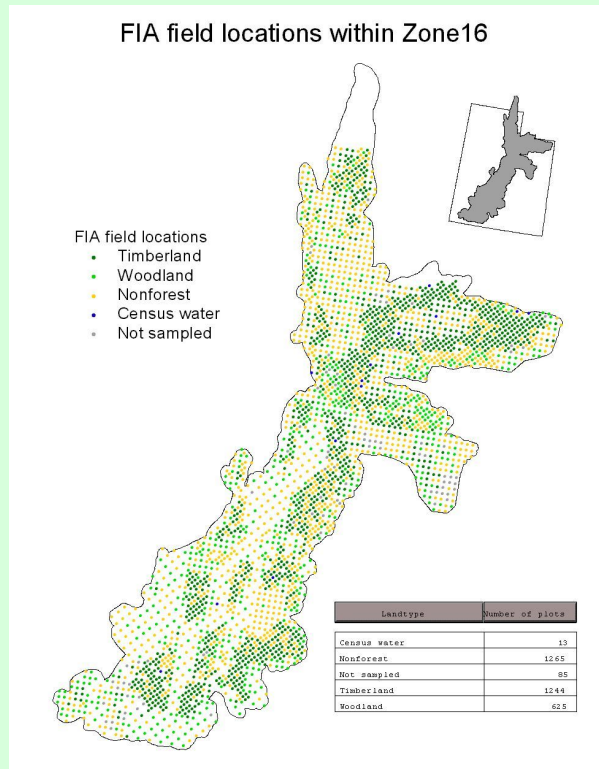


Back

Close

Application 2: Forest Health Monitoring

FHM: part of Forest Inventory and Analysis (FIA) of Forest Service



Back

Close

FIA and FHM data

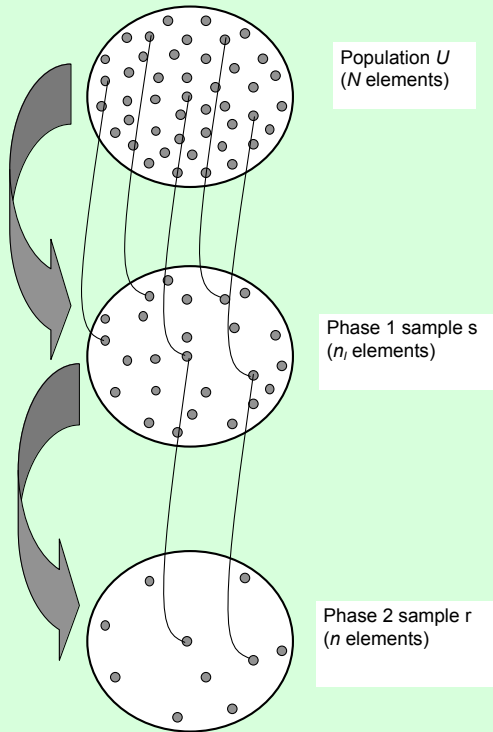
	Variables	
Geographic Information System (GIS)	X,Y	coordinates E-W, N-S
	elev	elevation (m)
	asp	aspect (deg)
	slope	slope (deg)
	hillshd	hillshade (solar radiation)
$N = 67,216$	nlcd	vegetation cover type (class)
<i>Forest Inventory and Analysis (FIA)</i>	fortyp	forest type (class)
	trees	number of trees
	agemax	max tree age (years)
	ageavg	avg tree age (years)
	bamax	max tree basal area (sq in)
	crcov	tree crown cover (%)
$n_I = 3,107$...	
<i>Forest Health Monitoring (FHM)</i>	lichen	lichen species present (count)
$n = 71$...	



Back

Close

Multi-phase Sampling



Geographic
Information
System
(GIS)

$$N = 67,216$$

*Forest
Inventory
and
Analysis*
(FIA)

$$n_I = 3,107$$

*Forest
Health
Monitoring*
(FHM)

$$n = 71$$



Back

Close

Model-assisted Estimation for Multi-phase Samples

Different information available at different phases

population $\mathbf{x}_{ai}, \mathbf{z}_{ai}, i \in U$ GIS variables

phase 1 $\mathbf{x}_{bi}, \mathbf{z}_{bi}, i \in s$ FIA measurements

phase 2 $\mathbf{y}_i, i \in r$ FHM measurements (lichen count)

$(\mathbf{x}_{bi}, \mathbf{z}_{bi}$ contains $\mathbf{x}_{ai}, \mathbf{z}_{ai}$)

Goal: estimate $\bar{\mathbf{y}}_N$ with generic but efficient estimator



Back

Close

Model-assisted Estimation for Multi-phase Samples

Different information available at different phases

population $\mathbf{x}_{ai}, \mathbf{z}_{ai}, i \in U$ GIS variables
phase 1 $\mathbf{x}_{bi}, \mathbf{z}_{bi}, i \in S$ FIA measurements
phase 2 $\mathbf{y}_i, i \in r$ FHM measurements (lichen count)
($\mathbf{x}_{bi}, \mathbf{z}_{bi}$ contains $\mathbf{x}_{ai}, \mathbf{z}_{ai}$)

Goal: estimate $\bar{\mathbf{y}}_N$ with generic but efficient estimator

Approach:

1. use penalized spline regression to fit semiparametric additive model for each “level” of auxiliary info
2. construct multi-phase model-assisted estimator



Back

Close

Model-assisted Estimation for Multi-phase Samples (2)

- Models

- Model **a**: using predictors available for U

$$E_{\xi}(y_i) = g_a(\mathbf{x}_{ai}, \mathbf{z}_{ai}) = m_a(\mathbf{x}_{ai}; \boldsymbol{\beta}_a) + \mathbf{z}_{ai}\boldsymbol{\gamma}_a$$

- Model **b**: using predictors available for s

$$E_{\xi}(y_i) = g_b(\mathbf{x}_{bi}, \mathbf{z}_{bi}) = m_b(\mathbf{x}_{bi}; \boldsymbol{\beta}_b) + \mathbf{z}_{bi}\boldsymbol{\gamma}_b$$



Back

Close

Model-assisted Estimation for Multi-phase Samples (2)

- Models

- Model a : using predictors available for U

$$E_{\xi}(y_i) = g_a(\mathbf{x}_{ai}, \mathbf{z}_{ai}) = m_a(\mathbf{x}_{ai}; \boldsymbol{\beta}_a) + \mathbf{z}_{ai}\boldsymbol{\gamma}_a$$

- Model b : using predictors available for s

$$E_{\xi}(y_i) = g_b(\mathbf{x}_{bi}, \mathbf{z}_{bi}) = m_b(\mathbf{x}_{bi}; \boldsymbol{\beta}_b) + \mathbf{z}_{bi}\boldsymbol{\gamma}_b$$

- Fit both models on data from r

- Model-assisted estimator

$$\hat{y}_{NP} = \frac{1}{N} \sum_U \hat{g}_{ai} + \frac{1}{N} \sum_s \frac{\hat{g}_{bi} - \hat{g}_{ai}}{\pi_{i(s)}} + \frac{1}{N} \sum_r \frac{y_i - \hat{g}_{bi}}{\pi_{i(s)} \pi_{i(r|s)}}$$

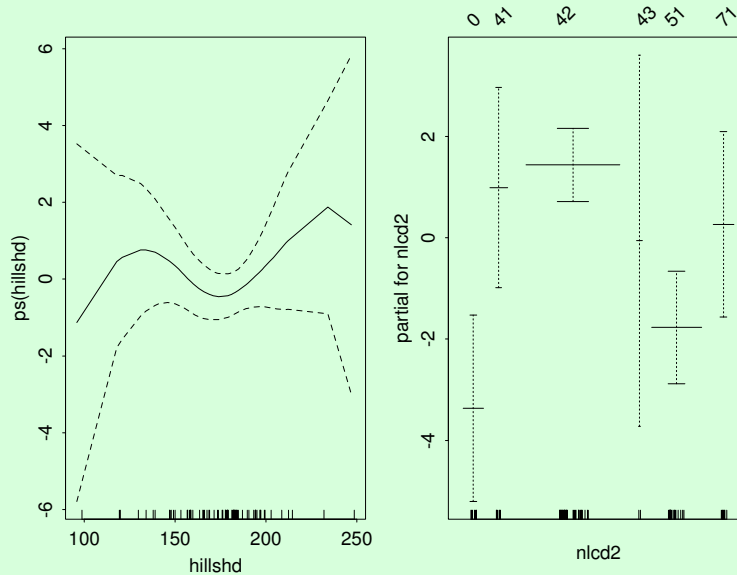


Back

Close

Semiparametric Model a

$$E_{\xi}(\text{LICHEN}) = m(\text{HILLSHD}; \beta) + z_{\text{NLCD}}\gamma$$

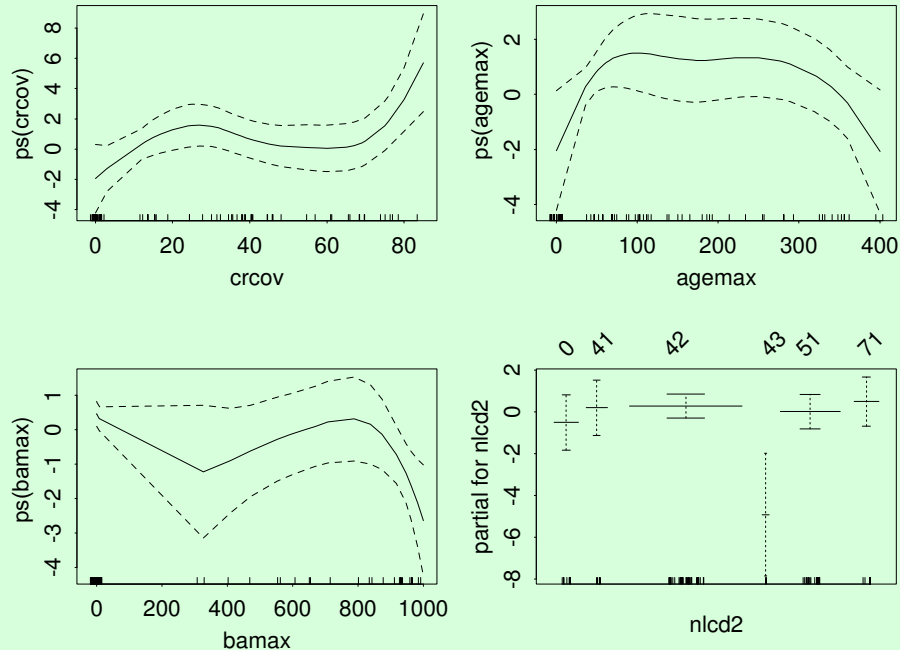


Back

Close

Semiparametric Model b

$$E_{\xi}(\text{LICHEN}) = m_1(\text{CRCOV}; \beta_1) + m_2(\text{AGEMAX}; \beta_2) \\ + m_3(\text{BAMAX}; \beta_3) + z_{\text{NLCD}}\gamma$$



Back

Close

Forest Health Monitoring Estimates

- HT = Generic estimator, ignores all auxiliary information
- Linear = Generic estimator, all models are fitted by linear regression
- Semiparametric = Generic estimator, semiparametric models

	HT	Linear	Semiparametric
Estimate	3.62	2.92	2.67
Est. St. Dev.	0.36	0.25	0.16
		(69%)	(44%)

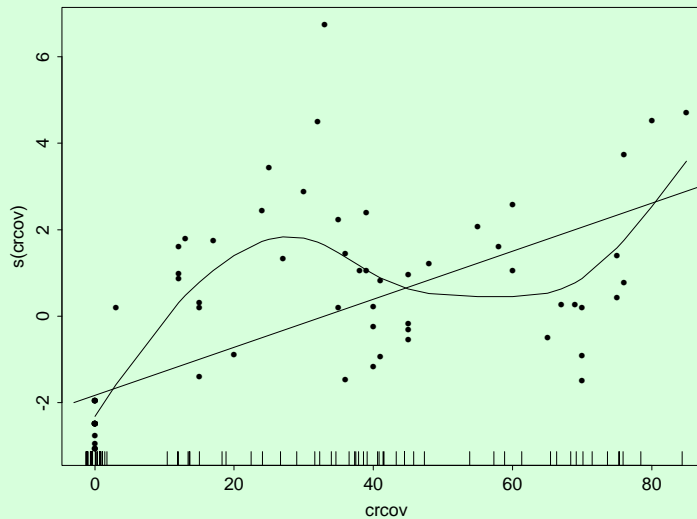


Back

Close

Estimators for Domains

- “Domain”: subpopulation for which separate estimator is needed
- Models can improve precision of domain estimators, if they have good local properties
- Nonparametric model better able to adapt to local features of data



Back

Close

Model-Assisted Estimators for Domains

(Särndal, 1984)

- Obtain sample-weighted model fit for complete sample s
- Estimator for domain $U_d \subset U$ with realized sample $s_d \subset s$ is

$$\hat{y}_{NP} = \frac{1}{N_d} \sum_{U_d} \hat{g}_i + \frac{1}{N_d} \sum_{s_d} \frac{y_i - \hat{g}_i}{\pi_i}$$

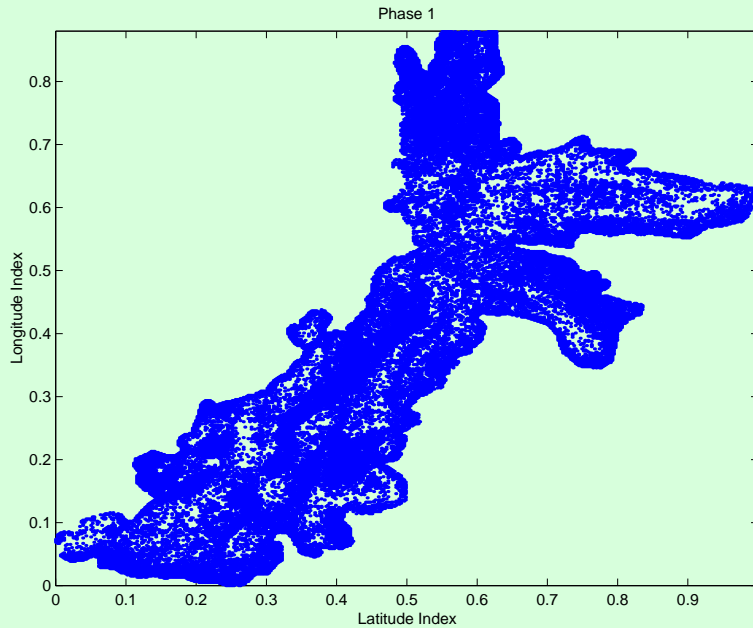
- Variance follows from model-assisted estimation theory
- Approach maintains additivity of domain estimates



Back

Close

Example: Estimation for Domain with NLCD > 50



Only $n = 27$ observations in domain



Back

Close

Example (2)

All Data ($n = 71$)

	HT	Linear	Semiparametric
Estimate	3.62	2.92	2.67
Est. St. Dev.	0.36	0.25	0.16

Domain ($n = 27$)

	HT	Linear	Semiparametric
Estimate	2.00	1.78	1.72
Est. St. Dev.	0.57 (1.58)	0.39 (1.56)	0.17 (1.06)

Nonparametric regression makes it possible to maintain generic approach at smaller “scales”



Back

Close

5. Smoothing Parameter Selection

- Smoothing parameter selection less important in generic estimation
 - ⇒ optimal value depends on variable being estimated
 - ⇒ but: single set of survey weights, many variables!



Back

Close

5. Smoothing Parameter Selection

- Smoothing parameter selection less important in generic estimation
 ⇒ optimal value depends on variable being estimated
 ⇒ but: single set of survey weights, many variables!

- Minimizing estimate of asymptotic design variance is poor choice

$$\hat{V}(\hat{y}_{NP}; h) = \frac{1}{N^2} \sum \sum_s \frac{y_i - \hat{m}_i(h)}{\pi_i} \frac{y_j - \hat{m}_j(h)}{\pi_j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}$$

- Proposed approach based on “design-based cross-validation”

$$CV(\hat{y}_{NP}; h) = \frac{1}{N^2} \sum \sum_s \frac{y_i - \hat{m}_i^{-i}(h)}{\pi_i} \frac{y_j - \hat{m}_j^{-j}(h)}{\pi_j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}$$

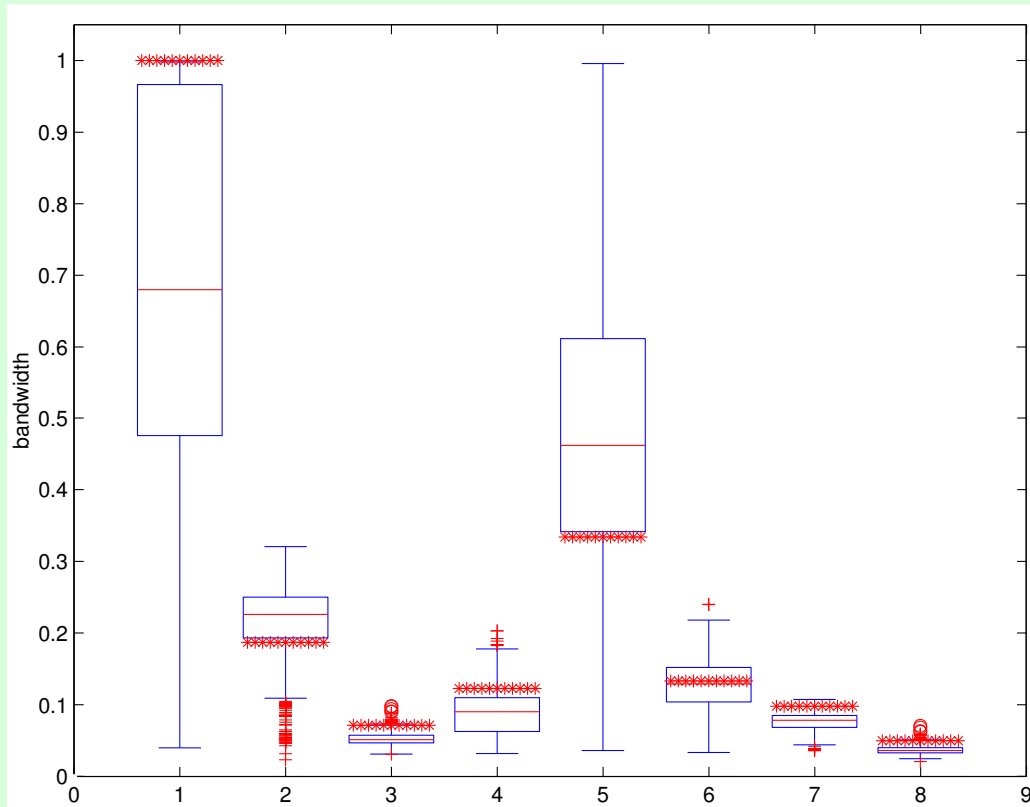
- “Leave-one-out” estimator \hat{m}_i^{-i} is easy to compute for most non-parametric regression techniques



Back

Close

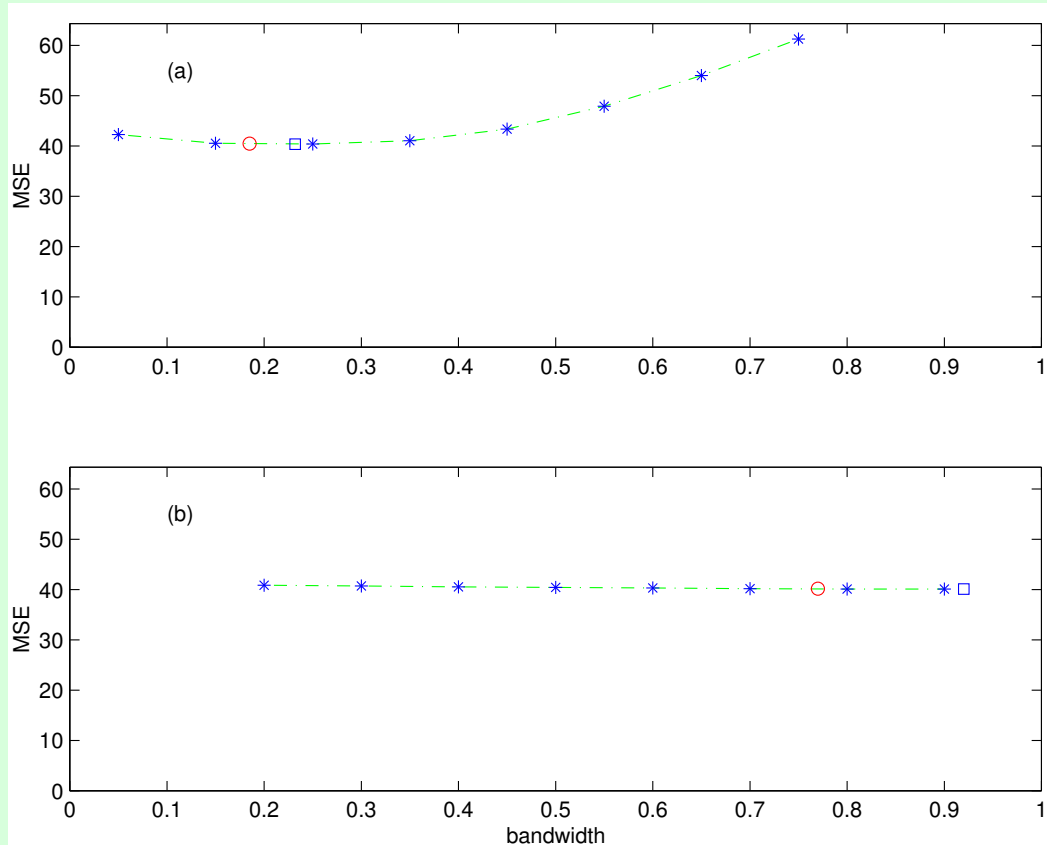
Smoothing parameter selection (2)



Back

Close

Smoothing parameter selection (3)



Back

Close

6. Conclusions

- Generic estimation can be improved with nonparametric methods
 - more efficient when relationship exists but parametric model not appropriate
 - almost as efficient when parametric model is correct
- Nonparametric model-assisted estimation
 - fits in current survey estimation paradigm
 - shares properties of parametric methods
 - complementary with parametric approaches
 - easy to implement with currently available software
- Requires unit-level “frame” information

Contact: – jopsomer@iastate.edu

– <http://www.public.iastate.edu/~jopsomer/home.html>



Back

Close