

# Endogenous Post-Stratification in Surveys: Classifying with a Sample-Fitted Model

---

F. Jay Breidt  
Colorado State University  
Jean D. Opsomer  
Iowa State University

*Supported by Joint Venture Agreement 02-JV-11222007-004  
with USDA Forest Service Rocky Mountain Research Station*

## Overview

---

- **Motivation:** Use of remotely-sensed imagery to improve survey estimation
  - classified image stratifies survey data
  - survey data trains classification algorithm
- **Big question:** Is this legitimate?
  - introduce a modeling framework for classification
  - study estimators' asymptotic behavior
  - evaluate in finite samples via simulation
- **Cautious conclusion:** yes, it's legitimate

# Stratification and post-stratification

---

- **(Pre-)stratification**

- use stratum information at the **design** stage
- partition landscape into homogeneous sub-areas
- sample independently from each

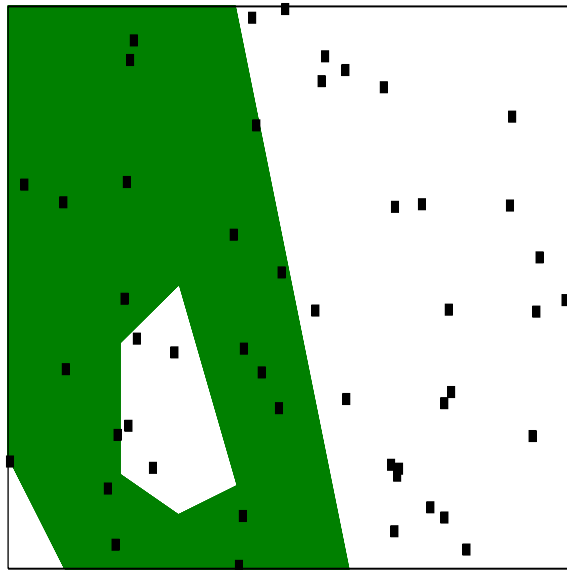
- **Post-stratification**

- use stratum information at the **estimation** stage
- sample the landscape, ignoring stratum information
- count sample elements falling into post-strata
- reweight sample to agree with landscape info

# How to use landscape-level auxiliary information?

---

- Remotely-sensed image of entire landscape
- Landsat 7 image of Hayman Springs fire scar: 7/18/2002



## Reweighting via post-stratification

---

- **Traditional post-stratification** with known classification:

- sample size is  $n = 50$ ; original weight =  $1/50 = 0.02$
- 0.445 of landscape is **known to be** green
- $19/50 = 0.380$  of sample is **known to be** green

- **Reweight:**

$$\begin{aligned} (\text{new weight}) &= (\text{original weight}) \frac{(\text{landscape proportion})}{(\text{sample proportion})} \\ &= \begin{cases} 0.02 \frac{0.445}{19/50} = 0.0234, & \text{for green sites;} \\ 0.02 \frac{0.555}{31/50} = 0.0179, & \text{for non-green sites} \end{cases} \end{aligned}$$

# Classifying the image into post-strata

---

- Need a mapping  $m(\cdot)$  from the image to the classes

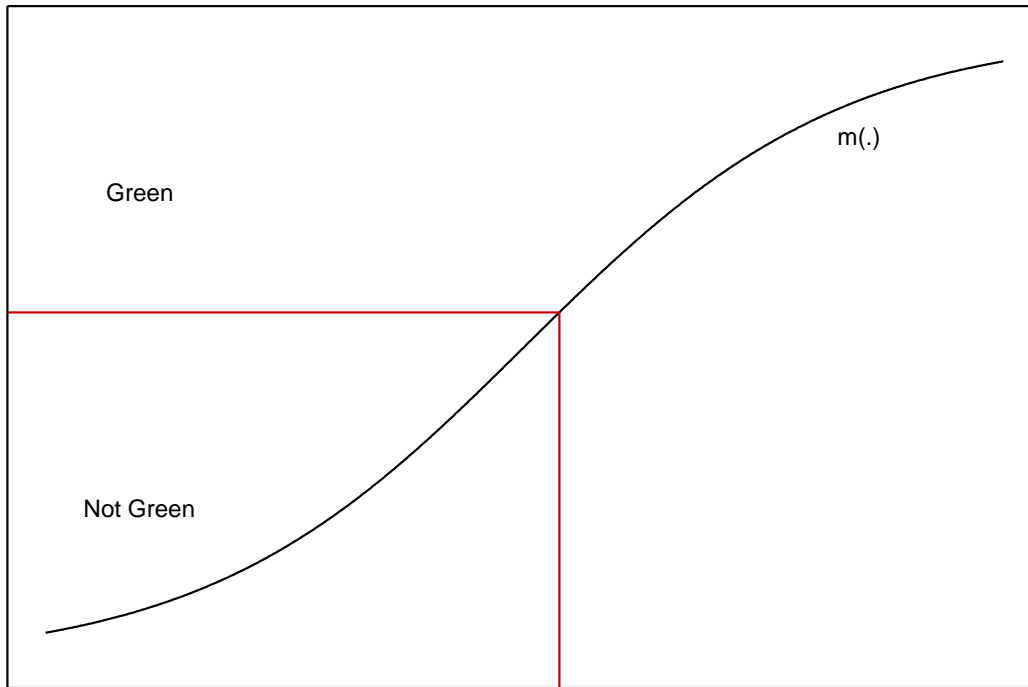


Image property, x

## More about the mapping from image to classes

---

- Covariate vector  $\mathbf{x}_i$  available for all  $i \in U$
- Index is created via  $m(\boldsymbol{\lambda}'\mathbf{x}_i)$  for some smooth  $m$
- Compare to stratum boundaries  $\{\tau_h\}_{h=0}^H$ : element  $i$  belongs to post-stratum  $h$  if

$$\tau_{h-1} \leq m(\boldsymbol{\lambda}'\mathbf{x}_i) < \tau_h$$

- Given  $\boldsymbol{\lambda}$ ,  $m$  maps image to post-strata **without error**

# Estimating the mapping from image to classes

---

- For  $\lambda$  unknown, assume we have sample data  $z_i$  from **generalized linear model** with  $E[z_i | \mathbf{x}_i] = m(\lambda' \mathbf{x}_i)$

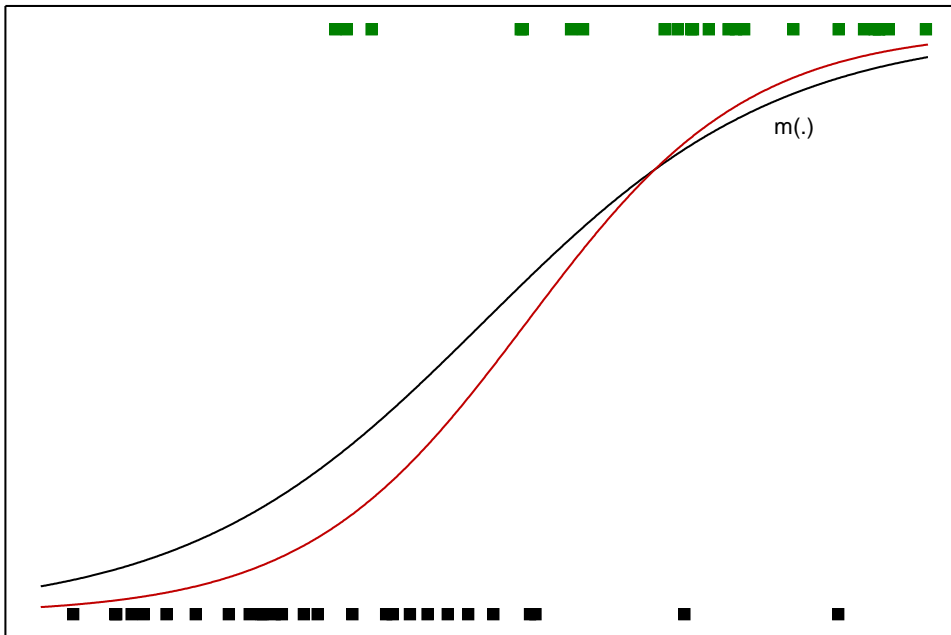


Image property, x

# Post-stratification versus endogenous post-stratification

---

- Traditional post-stratification:  $\text{PSE}(\boldsymbol{\lambda})$ 
  - $\boldsymbol{\lambda}$  in  $m(\boldsymbol{\lambda}'\mathbf{x}_i)$  is known
  - **known** landscape proportion, **known** sample proportion
- **Endogenous post-stratification**:  $\text{EPSE}(\hat{\boldsymbol{\lambda}})$ 
  - estimate  $\boldsymbol{\lambda}$  in  $m(\boldsymbol{\lambda}'\mathbf{x}_i)$  using sample data
$$(\text{EPSE weight}) = \frac{1 \text{ (estimated landscape proportion)}}{n \text{ (estimated sample proportion)}}$$

## Is endogenous post-stratification legitimate?

---

- Consider asymptotic approximation to statistical estimator:

$$\begin{aligned}\bar{y}_s &= \mu_y + (\bar{y}_s - \mu_y) = \mathbf{fixed} + \mathbf{small} \\ \bar{y}_s + \hat{\beta}(\mu_x - \bar{x}_s) &= \mu_y + \{(\bar{y}_s - \mu_y) + \beta(\mu_x - \bar{x}_s)\} \\ &\quad + (\hat{\beta} - \beta)(\mu_x - \bar{x}_s) \\ &= \mathbf{fixed} + \mathbf{small} + \mathbf{tiny}\end{aligned}$$

- Ignore the tiny stuff, but **not** the small stuff:
  - **small** randomly varies about zero in finite samples
  - $\sqrt{\text{Var}(\mathbf{small})}$  = asymptotic std. error of estimator
- Is effect of estimating  $\lambda$  **small** or **tiny**?

## Taylor linearization for nonlinear functions

---

- Expand non-linear statistical estimator for  $f$  smooth:

$$\begin{aligned} f(\hat{\boldsymbol{\lambda}}) &= f(\boldsymbol{\lambda}) + \frac{\partial f(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}'} (\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) + \text{remainder} \\ &= \mathbf{fixed} + \mathbf{small} + \mathbf{tiny} \end{aligned}$$

- **Problem:** EPSE( $\hat{\boldsymbol{\lambda}}$ ) is not smooth

– e.g., sample proportion is

$$n^{-1} \sum_{i \in s} I \left\{ \tau_{h-1} \leq m(\hat{\boldsymbol{\lambda}}' \mathbf{x}_i) < \tau_h \right\}$$

- **Solution:** Use Randles (1982, *Ann. Stat.*)

## Asymptotic results for EPSE

---

- For any study variable  $y$ , **EPSE variance estimator** is analogue of PSE's:

$$\hat{V} = \frac{1}{n} \left(1 - \frac{n}{N}\right) \sum_{h=1}^H \frac{A_h^2(\hat{\boldsymbol{\lambda}})}{n_h(\hat{\boldsymbol{\lambda}})/n} S_{yh}^2(\hat{\boldsymbol{\lambda}})$$

- Under generalized linear model assumptions on  $z$ ,

$$\text{EPSE}(\hat{\boldsymbol{\lambda}}) \pm 1.96\hat{V}^{1/2}$$

is an **asymptotic 95% confidence interval** for the mean  $\mu_y$  of any study variable  $y$

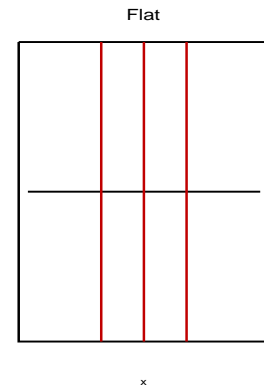
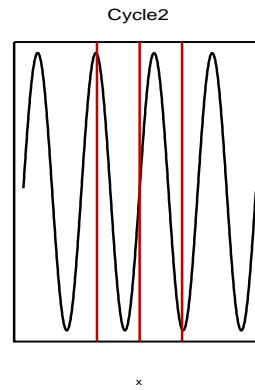
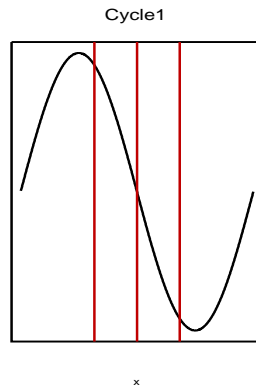
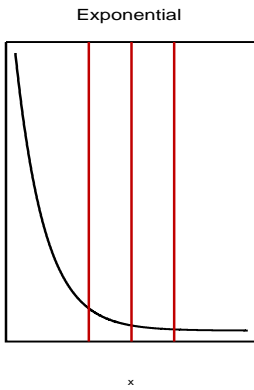
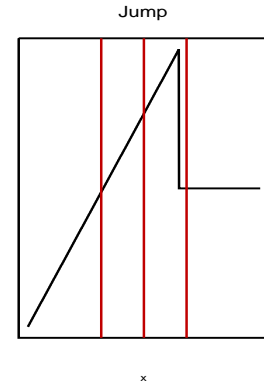
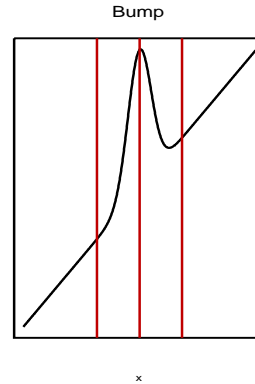
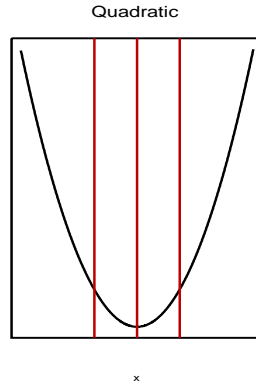
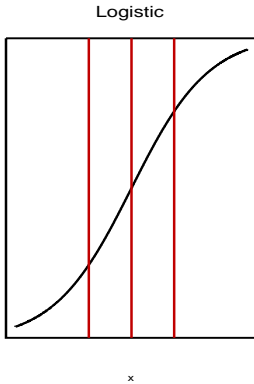
## Simulation experiments

---

- Comparison of three estimators:
  - Horvitz-Thompson (HTE),  $PSE(\boldsymbol{\lambda})$ , and  $EPSE(\hat{\boldsymbol{\lambda}})$
  - obtain weights for each
  - apply weights to all eight study variables
- Linear and logistic classifiers
  - high and low noise
  - $n = 100, 200, 500$
  - $H = 2$  or 4 strata
- Compare efficiency of estimators and percent relative bias of variance estimators

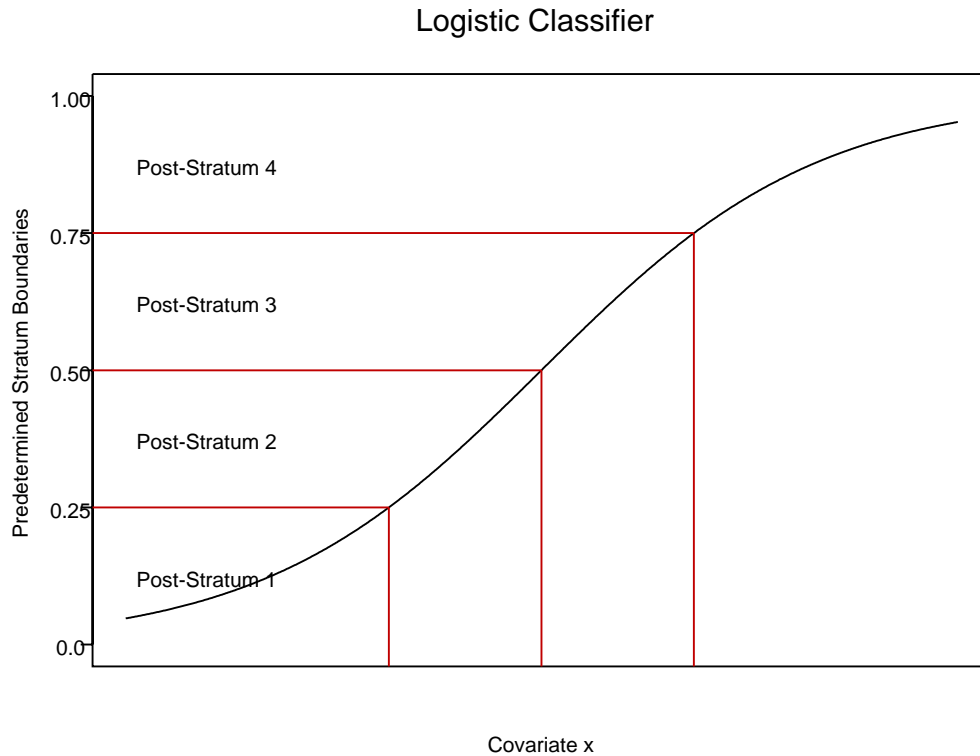
# Apply each set of weights to eight populations

---



# Focus on results for logistic model

---



## Logistic EPSE efficiency: $MSE(^*)/MSE(EPSE)$

---

- 1000 simple random samples of size  $n = 200$
- Greater than one favors EPSE

Estim.	Number of Strata	Mean $m$	Quadr.	Bump	Jump	Expon.	Cycle 1	Cycle 2	Flat
HTE	2	1.35	0.99	1.71	1.05	1.05	2.33	1.00	0.99
	4	1.47	1.06	2.39	1.34	1.12	2.44	1.02	0.97
PSE	2	0.99	1.00	0.85	0.97	1.00	0.90	1.00	0.99
	4	1.02	1.01	0.92	0.98	0.99	0.92	1.00	0.99

## Percent relative bias of variance estimator

---

- 1000 simple random samples of size  $n = 200$

Number of Strata	Estim.	Mean $m$	Quadr.	Bump	Jump	Expon.	Cycle 1	Cycle 2	Flat
2	PSE	2.81	2.27	7.58	6.72	-5.64	3.94	3.45	2.77
	EPSE	2.98	2.62	6.95	7.07	-5.13	-1.57	3.87	2.50
4	PSE	-0.88	0.76	7.92	8.58	-3.71	3.31	3.27	2.89
	EPSE	2.25	2.01	2.59	10.34	-3.95	-3.51	4.45	1.84

## Conclusions

---

- EPSE may be useful in forest inventory
  - allows training of classifier with survey data
- Asymptotically equivalent to traditional PSE under generalized linear model
  - effect of estimating  $\lambda$  can be ignored
- Simulation study supports theoretical results
  - efficiency of EPSE comparable to PSE
  - EPSE's variance estimator no worse than PSE's

## Further research

---

- Alternative asymptotic formulation for GLIM case?
- Simulations for non- or semi-parametric case
  - CART, MARS, neural nets, and other algorithmic classifiers
  - large, growing number of “parameters”
- Theory for non-, semi-parametric case?