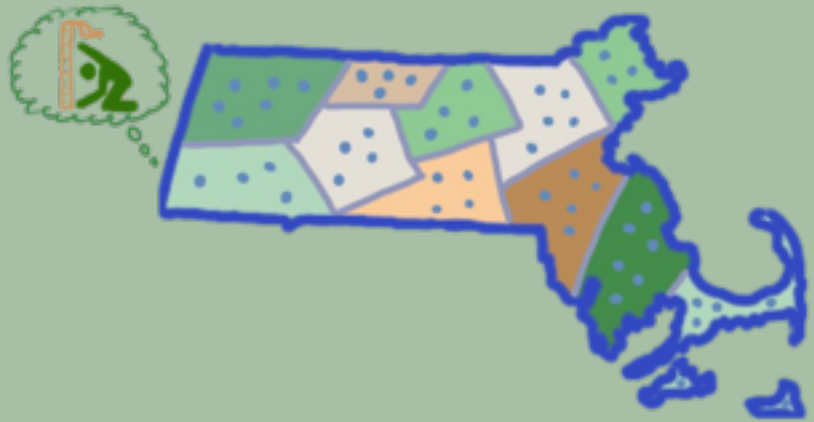
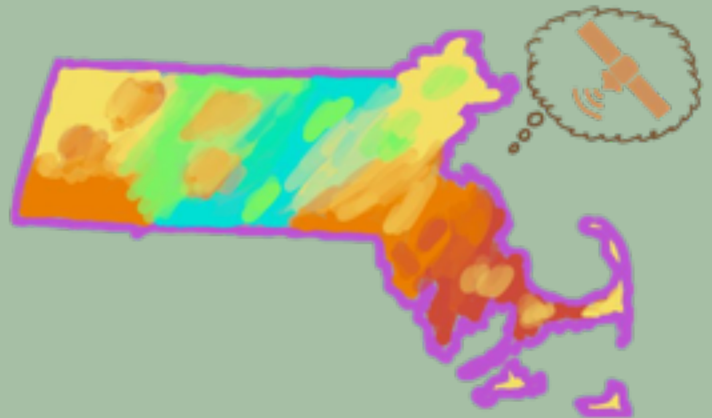
# Comparison Study of Survey Sampling Estimators


Professor Kelly McConville (Harvard)

Asteria Chilambo (Harvard Math ‘23),
Jing Shang (Fudan Economics ‘23,
Visiting Student at Harvard Statistics ‘22)

# How does typical forestry data look like?





**Ground Plot:**
- Directly measures variables of interest
- Precise, but expensive and sparse!

**Remote sensor census:**
- Indirectly provides related information
- Cost-effective, but not exactly what we want

# How can we use auxiliary variables  to assist estimation for variables of interest? 

Generalized Multivariable Difference Estimator (GMDE)

&

Generalized Regression Estimator (GREG)

# GMDE

Step 1. Make use of Horvitz-Thompson $\hat{\mathbf{t}}_{\mathbf{y}\pi}$ and auxiliary residual $\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x},\pi}$


Step 2. Among all linear combination of $\hat{\mathbf{t}}_{\mathbf{y}\pi}$ and $\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}\pi}$, choose one with minimal variance

# GREG

Step 1. Start with linear regression

Step 2. Optimize regression coefficient

Step 3. Use **population** information of x to predict y **through regression**, plus **sample** information of residual

# How do GMDE and GREG perform under different sampling scenarios?

# What is dominating the difference?

- **GMDE**

$$\hat{\mathbf{t}}_{\mathbf{y},gmde} = \hat{\mathbf{t}}_{\mathbf{y}\pi} + \hat{\mathbf{V}}_{\mathbf{yx},\pi} \cdot \hat{\mathbf{V}}^{-1}_{\mathbf{x},\pi} \cdot (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x},\pi})$$

Pos. Corr.

If ppt. detected from sensor is bigger than est. from sample

- **GREG**

$$\hat{t}_{y,greg} = \hat{t}_{y\pi} + \left(\sum_{i \in S} \frac{x_i y_i}{\pi_i}\right)^{\intercal} \left(\sum_{j \in S} \frac{x_j x_j^{\intercal}}{\pi_j}\right)^{-1} \cdot (t_x - \hat{t}_{x\pi})$$

# Simple Random Sampling


## Single Study Variable

- **GMDE**

$$\hat{\mathbf{V}}_{\mathbf{yx},\pi}\hat{\mathbf{V}}^{-1}_{\mathbf{x},\pi} = \frac{\frac{1}{n}\sum_{i\in S} x_i y_i - \left(\frac{1}{n}\sum_{i\in S} y_i\right)\left(\frac{1}{n}\sum_{i\in S} x_i\right)}{\frac{1}{n}\sum_{i\in S} x_i^2 - \left(\frac{1}{n}\sum_{i\in S} x_i\right)^2}$$

$$\frac{\text{Cov. of } y \text{ and } x}{\text{Cov. of } x}$$

- **GREG**

$$\left(\sum_{j\in S}\frac{x_j x_j^\top}{\pi_j}\right)^{-1}\left(\sum_{i\in S}\frac{x_i y_i}{\pi_i}\right) = \frac{\frac{1}{n}\sum_{i\in S} x_i y_i - \left(\frac{1}{n}\sum_{i\in S} y_i\right)\left(\frac{1}{n}\sum_{i\in S} x_i\right)}{\frac{1}{n}\sum_{i\in S} x_i^2 - \left(\frac{1}{n}\sum_{i\in S} x_i\right)^2}$$

**GMDE and GREG are the same under SRS!**

# Simple Random Sampling


## Multiple Study Variables

- **GMDE**

$$(\hat{\mathbf{t}}_{\mathbf{y},gmde})_m = (\hat{\mathbf{t}}_{\mathbf{y}\pi})_{\mathbf{m}} + \underbrace{(\hat{\mathbf{V}}_{\mathbf{yx},\pi})_{\mathbf{m}}(\hat{\mathbf{V}}^{-1}_{\mathbf{x},\pi})_{\mathbf{m}}}_{\frac{\text{Cov. of } m^{th} y \text{ and all } x}{\text{Cov. of all } x}}\,(-\hat{\mathbf{t}}_{\mathbf{x}\pi} + \mathbf{t}_{\mathbf{x}})$$

- **GREG**

$$\hat{t}_{y,greg} = \hat{t}_{y\pi} + \overbrace{\left(\sum_{i\in S}\frac{x_i y_i}{\pi_i}\right)^{\intercal}\left(\sum_{j\in S}\frac{x_j x_j^{\intercal}}{\pi_j}\right)^{-1}} \cdot (t_x - \hat{t}_{x\pi})$$

Using GREG to estimate the $m^{th} y$

**GMDE and GREG are the same under SRS!**

# Stratified Simple Random Sampling


Intuition:

- When sample mean varies significantly across strata, then GMDE is more precise.
- Because $\hat{V}_{yx,\pi} \cdot \hat{V}^{-1}_{x,\pi}$ is measured within each stratum, while reg coefficient is computed across strata. The den shrinks faster.

## No simple form for two ratios, but have numerical similarities:

| GMDE_DR | GREG_DR |
|---|---|
| 0.601 | 0.591 |
| 0.599 | 0.584 |
| 0.595 | 0.580 |
| 0.600 | 0.579 |
| 0.608 | 0.595 |

Comparison of dominant ratio (also dominates var.)

| GMDE | GREG | GMDE_Var | GREG_Var |
|---|---|---|---|
| 1811.995 | 1814.122 | 0.895 | 0.923 |
| 1616.635 | 1616.157 | 0.721 | 0.815 |
| 1484.934 | 1483.678 | 1.151 | 1.267 |
| 1765.941 | 1765.211 | 1.146 | 1.178 |
| 1769.369 | 1766.972 | 1.188 | 1.233 |

Comparison of estimator and variance

# Main takeaways

A. Both GMDE and GREG make use of **auxiliary variables** and can **improve performance** compared with Horvitz-Thompson estimator
B. Under **simple random sampling**, GMDE and GREG performs the same
C. Under **stratified** simple random sampling, GMDE performs better

# Future work

**GMDE deserves more attention!**

- Take different estimators of variance into consideration (eg. Hajek-Berger, Hansen-Hurwitz)
- Extend sampling scenarios (eg. two phase sampling)

# Special Thanks to:

## Ray Czaplewski

**Emeritus Scientist**

**Forest Inventory & Analysis**

## George Gaines

**Research Mathematical Statistician**

**Forest Inventory & Analysis**

## Kelly McConville

**Harvard University**

**Department of Statistics**