Synthetic Area Weighting for Measuring Public Opinion in Small Areas

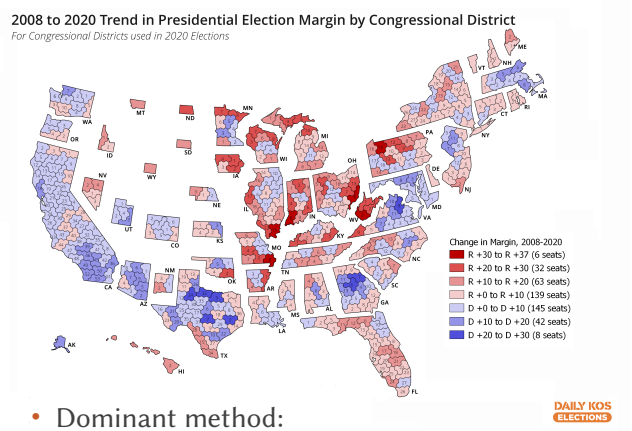
Shiro Kuriwaki and **Soichiro Yamauchi**

PolMeth 2021

July 2021

Small Area Estimation

- Enables local geography / subgroup estimates



- Dominant method:
  Multilevel Regression + Poststratification (MRP)

But at what cost?

> "I think MRP is good. I think it's overrated also ... [I understand MRP as asking] what the polls would look like in a particular state or district without having any polls of that state or district... it's a good approximation, but it loses a lot of local variation."

Nate Silver (2018)

What this paper does

1. Derive identification conditions for unbiasedness when borrowing information across small areas
2. A weighting approach clarifies conditions
   - compared to an outcome-based approach like MRP
3. We leverage existing data not used in MRP
   - initial (non-small area) survey weights
   - covariates only measured in the survey

Our method in a nutshell: Estimating FL-27 (Miami)

- Index people by i, areas of by j
- Survey inclusion for person i: $S_i \in \{0, 1\}$
- Person i is in Area j: $A_{ij} \in \{0, 1\}$
- Estimand: $\mathbb{E}\left[Y_i | A_{ij} = 1\right]$

Combine

1. Direct estimator

   (Only use FL-27 respondents)

   - Weight by inverse of $\Pr\left(S_i = 1 \mid \text{Covariates}_i, A_{ij} = 1\right)$

2. Indirect estimator

   (Use non-FL-27 respondents, reweight to look like FL-27)

   - Weight by $\Pr\left(A_{ij} = 1 \mid \text{Covariates}_i, S_i = 1\right)$
   - Weight them *again* to look like population

Partial pooling needs two identification conditions

Assumption 1 (Sampling Ignorability)

$$Y_i \perp\!\!\!\perp S_i \mid \mathrm{X}_i^P, A_{ij} = 1$$

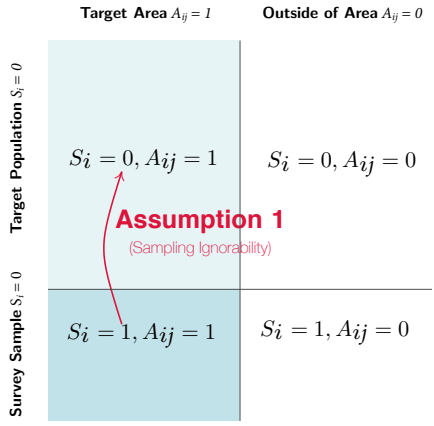
- X_i^P: Poststratification variables (with Population target)

  e.g. age, sex
- Required for almost any survey estimate

→ If satisfied, **direct estimator** using weights

$$\frac{1}{\Pr\left(S_i = 1 | \mathrm{X}_i^P, A_{ij} = 1\right)}$$

is unbiased.



Partial pooling needs two identification conditions

Assumption 2 (Area Ignorability)

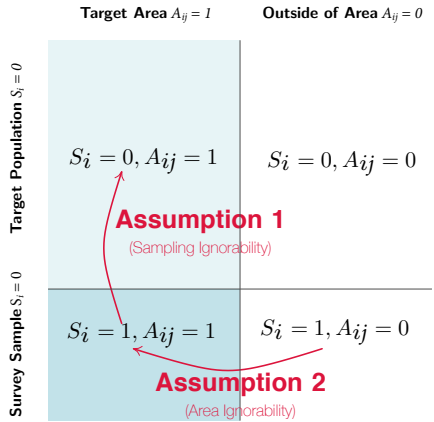
$$Y_i \perp\!\!\!\perp A_i \mid \mathrm{X}_i^P, \mathrm{X}_i^S, S_i = 1$$

- X_i^S: Covariates only in the Survey Sample
  e.g., party ID, news interest
- Unique to small area estimation

→ If satisfied, **indirect estimator** using weights

$$\frac{1}{\Pr\left(S_i = 1 | \mathrm{X}_i^P, A_{ij} = 1\right)} \cdot \frac{\Pr(A_{ij} = 1 | \mathrm{X}_i^P, \mathrm{X}_i^S, S_i = 1)}{1 - \Pr(A_{ij} = 1 | \mathrm{X}_i^P, \mathrm{X}_i^S, S_i = 1)}$$

is unbiased.



Our synthetic area estimator

Simplest case: if post-stratification covariates (X_i^P) were sufficient, then compute weights

$$\widehat{w}_{ij}^{\mathsf{SA}} \propto \underbrace{\Pr(A_i = j \mid \mathrm{X}_i^P)}_{\text{Area adjustment}} \times \underbrace{\frac{1}{\widehat{\Pr}(S_i = 1 \mid \mathrm{X}_i^P)}}_{\text{Selection adjustment}}$$

and take the weighted average across all respondents.

Result: With Assumptions 1 and 2, $\sum_{i=1}^n \widehat{w}_{ij}^{\mathsf{SA}} Y_i$ is unbiased for $\mathbb{E}\left[Y_i | A_{ij} = 1\right]$.

More generally,

$$\widehat{w}_{ij}^{\mathsf{SA}} \propto \overbrace{\frac{\widehat{\Pr}(A_i = j \mid \mathrm{X}_i^P, \mathrm{X}_i^S, S_i = 1)}{\widehat{\Pr}(A_i = j \mid \mathrm{X}_i^P, S_i = 1)}}^{\text{Adjustment from new } X_i^S} \times \Pr(A_i = j \mid \mathrm{X}_i^P) \times \frac{1}{\widehat{\Pr}(S_i = 1 \mid \mathrm{X}_i^P)}$$

Difference of our approach vs. other methods

1. *vs.* Traditional MRP
   - ⇝ Fay and Herriot (1979) style estimator partially pools by random effects (global + local shrinkage estimator)

2. *vs.* Machine learning (MR)P
   - ⇝ Ghitza and Gelman (2013); Bisbee (2019); Ornstein (2020); Goplerud (2020) all extract as many interactions as possible from X_i^P, not partially pooling

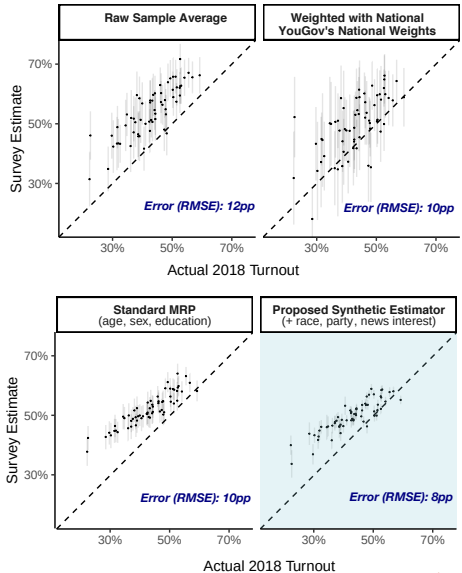
3. *vs.* "Multilevel regression with synthetic *poststratification*"
   - ⇝ Leemann and Wasserfallen (2017) expand X_i^P through missing data estimation.
     (We estimate the *area*, not *population*, synthetically)

4. *vs.* "Subgroup Balancing Propensity Score"
   - ⇝ Ben-Michael, Feller, and Rothstein estimate the propensity score by partial pooling, but not the outcome

Combined estimator reduces variance and bias

- CCES 2018
  - 63 Congressional Districts in Texas + Florida, $n \approx 140$ each

- MRP vs. Proposed estimator (`synthArea`)
  - X_i^P: Age group + sex + education
  - X_i^S: Race × Party ID × News interest



Insight 1: The combined estimator is a rescaling of national weights

$$\widehat{w}_{ij}^{\mathsf{SA}} \propto \frac{\widehat{\Pr}(A_i = j \mid \mathrm{X}_i^P, \mathrm{X}_i^S, S_i = 1)}{\widehat{\Pr}(A_i = j \mid \mathrm{X}_i^P, S_i = 1)} \times \Pr(A_i = j \mid \mathrm{X}_i^P) \times \underbrace{\frac{1}{\widehat{\Pr}(S_i = 1 \mid \mathrm{X}_i^P)}}_{\text{Plug-in } w_i^{\texttt{national}}}$$

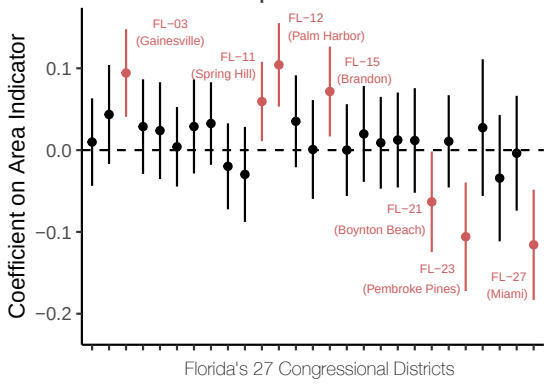
The survey weights $w^{\texttt{national}}$...

- already given in the dataset
- adjustment for covariates beyond researcher's X_i^P
- not used in MRP

Insight 2: Violations of area ignorability can be empirically tested

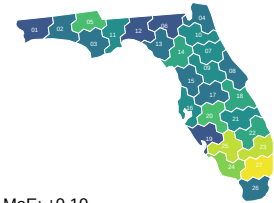
- For each area j, regress $Y_i \sim A_{ij} + X_i^P + X_i^S$
- Area ignorability is valid for area j when the coefficient on A_{ij} is 0.

Florida turnout example:

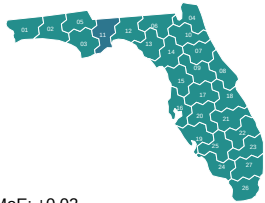


Insight 3: Lack of covariates can cause partial pooling to "oversmooth"

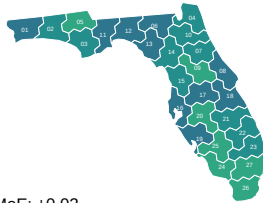
Direct (No Pooling)	Age + Sex + Education adjustment (baseline)	[baseline] + Race adjustment	[baseline] + Race*Party ID + Hispanic Origin + Religion adjustment



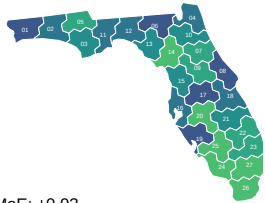
MoE: ±0.10



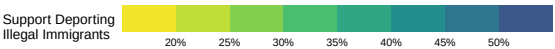
MoE: ±0.02



MoE: ±0.02



MoE: ±0.03



Takeaways

1. Small Area Estimation is not assumption-free
2. "Area Exchangeability" is key assumption in Small Area Estimation
3. Synthetic weighting (R package `synthArea`) can incorporate existing pollster's survey weights and survey-only variables