Towards Developing Best Practices for Using SAE for Equity Research

Carolina Franco

NORC at the University of Chicago

JSM 2024

With co-authors: Angelo Cozzubo, Taylor Wing, Carissa Villanueva

Small Area Estimation (SAE)

- Objective: To estimate quantities of interest for many geographic and/or demographic domains or other subsets of the population
- Problem: Often, surveys cannot estimate all quantities of interest through "direct" methods with acceptable precision
  Direct estimator: *Based only on sample data for the domain*
  Small area: *Domain where the sample size is too small for reliable direct estimation*
- SAE: Through modeling, incorporates auxiliary information from other sources; assumes common relationships across domains to "borrow strength" (censuses, administrative records, other surveys, previous vintages, satellite data, etc.)
- Result: Obtain more accurate estimates at granular level (with lower measures of uncertainty)

Illustration: Applying SAE to home ownership by race/ethnicity and state using GSS data and ACS covariates

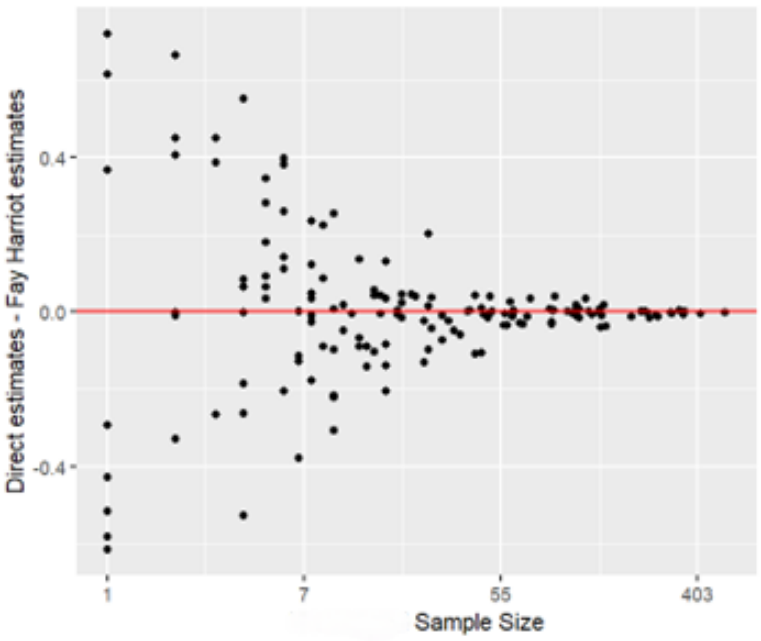

Figure 1: Prediction differences

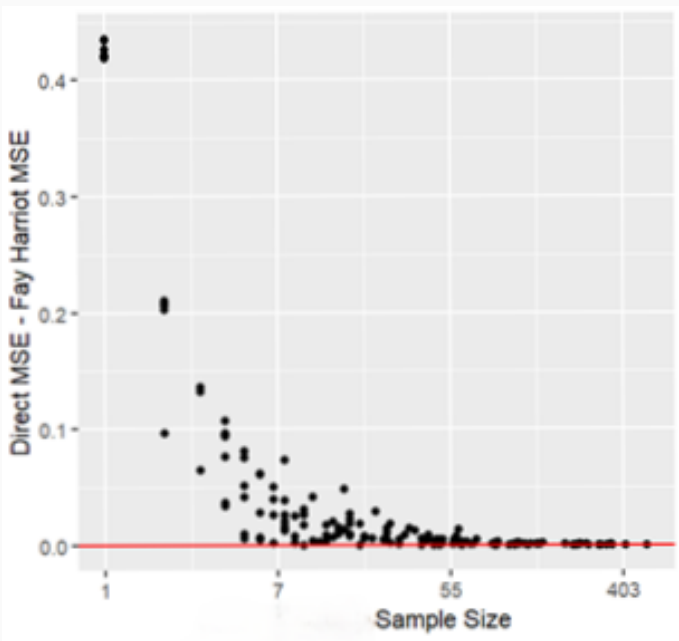

Figure 2: MSE differences

General Social Survey (GSS): full-probability, personal-interview survey designed to monitor changes in both social characteristics and attitudes

SAE 101: The Fay-Herriot Model (FH, 1979)

- For m small areas:

$$y_i = Y_i + e_i \qquad i = 1, \ldots, m$$
$$Y_i = \mathbf{x}_i'\beta + u_i$$

- Y_i is the population characteristic of interest for area i
- y_i is the direct survey estimate of Y_i
- e_i is the sampling error in y_i, usually assumed as $N(0, v_i)$, independent with v_i known
- u_i is the random area effect for area i, usually assumed as *i.i.d.* $N(0, \sigma_u^2)$ and independent of the e_i

More on the FH Model

- Best predictor of Y_i (β and σ_u^2 known):

$$\hat{Y}_i = (1 - \gamma_i) y_i + \gamma_i \mathbf{x}_i' \beta$$

  where

$$\gamma_i = \frac{v_i}{v_i + \sigma_u^2}$$

- Linear combination of the direct estimator (y_i) and the "synthetic" estimator ($\mathbf{x}_i'\beta$)
- Smaller sampling variances imply more weight to y_i
- Fitting via hierarchical Bayes (HB) or empirical Bayes (EB)
- If no covariates are available, you can replace "$\mathbf{x}_i'\beta$" with μ, "shrinkage to the mean" (e.g, Carter and Rolph, 1974)

Motivation

There is an **increased need to produce granular statistics not just by detailed geographic levels, but by other characteristics** of the population...

For example, the United Nations SAE toolkit, states about Sustainable Development Goals Indicators:

"SDG indicators should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics...

As sound statistical methods are vital to overcome this challenge, **Small Area Estimation (SAE) constitutes an important topic in the way forward**"

`https://unstats.un.org/wiki/display/SAE4SDG`

Motivation

National Academies, "Towards a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources," 2023

"*In the panel's view, much more research is needed to better understand the effects of data-combination methods on equity...*"

This work is a small step in this direction

Some recent examples

- **Hearing loss in the United States:** by county by groups defined by age, race and ethnicity, gender, under a grant for the CDC.
  See Rein, Franco, et al. (2024)
  `https://soundcheckmap.org/`
- **EPOP Small Area Estimates:** Entrepreneurship activities by state or MSA, by race/ethnicity and gender, for the EPOP project sponsored by the Kauffman Foundation
  `https://epop.norc.org/us/en/epop.html`

The current scope of the work

- SAE models trained mostly by data from majority. Can we accurately capture estimates for minorities, differences?
- What challenges arise and what are potential strategies to address them?
- Approach:
  - Critical literature review
  - Design-based simulation study inspired by a case study
  - Empirical examples
  - Case study: GSS home ownership rates by state and race/ethnicity (Hispanic, Non-Hispanic Black, Non-Hispanic White, or "Other")
  - Discussion with other experts – room for a lot more!
- Work in progress

About the design-based simulation

- Simulation's artificial population: ACS PUMS data
- ACS variable of interest: Home ownership
- Covariates also from ACS PUMS. Examples: Income to poverty ratio, % foreign born, % employed, % at least high school, median age
- 4 different sampling designs, 500 replications
- Tested various modeling choices using the FH model
- Computed bias and MSE from replications

Sampling designs of simulations

Strata/domains are defined by race/ethnicity and state

1. **Equal Sample Size 1%**: equal sample size for all domains, sampling fraction 1% ($n = 30,269$).
2. **Equal Sample Size GSS Size**: total sample comparable to 2021 GSS ($n = 4,080$).
3. **PPS 1%**: sampling fraction of 1% ($n = 30,943$).
4. **PPS with GSS Size**: probability proportional to race/ethnicity pop size with a total sample size comparable to 2021 GSS ($n = 4,080$)

Some of the questions analyzed with the simulation

- Effectiveness of using SAE to improve domain estimates/ estimates of differences
  - Generally, substantial MSE reductions applying SAE in this setting, compared with the direct estimators
  - In some cases with small sample sizes, SAE decreased bias (direct estimators design consistent but not design unbiased)
- Impact of various modeling choices (examples: one model for all domains vs. models for each race/ethnicity group; impact of race/ethnicity fixed effect)
- Show that using SAE in this context can be very effective when "done right"

Some results of the simulation

Table 1: Median % MSE improvements, FH vs direct estimates by race (with simple covariates and race/ethnicity fixed effect)

Races / Designs	**eq_1**	**eq_gss**	**pps_1**	**pps_gss**
Non-Hispanic White	47	71	30	66
Non-Hispanic Black	52	83	83	92
Hispanic	44	80	81	92
Other race or multirace	53	79	74	91
Overall	48	78	73	89

Example of simulation results

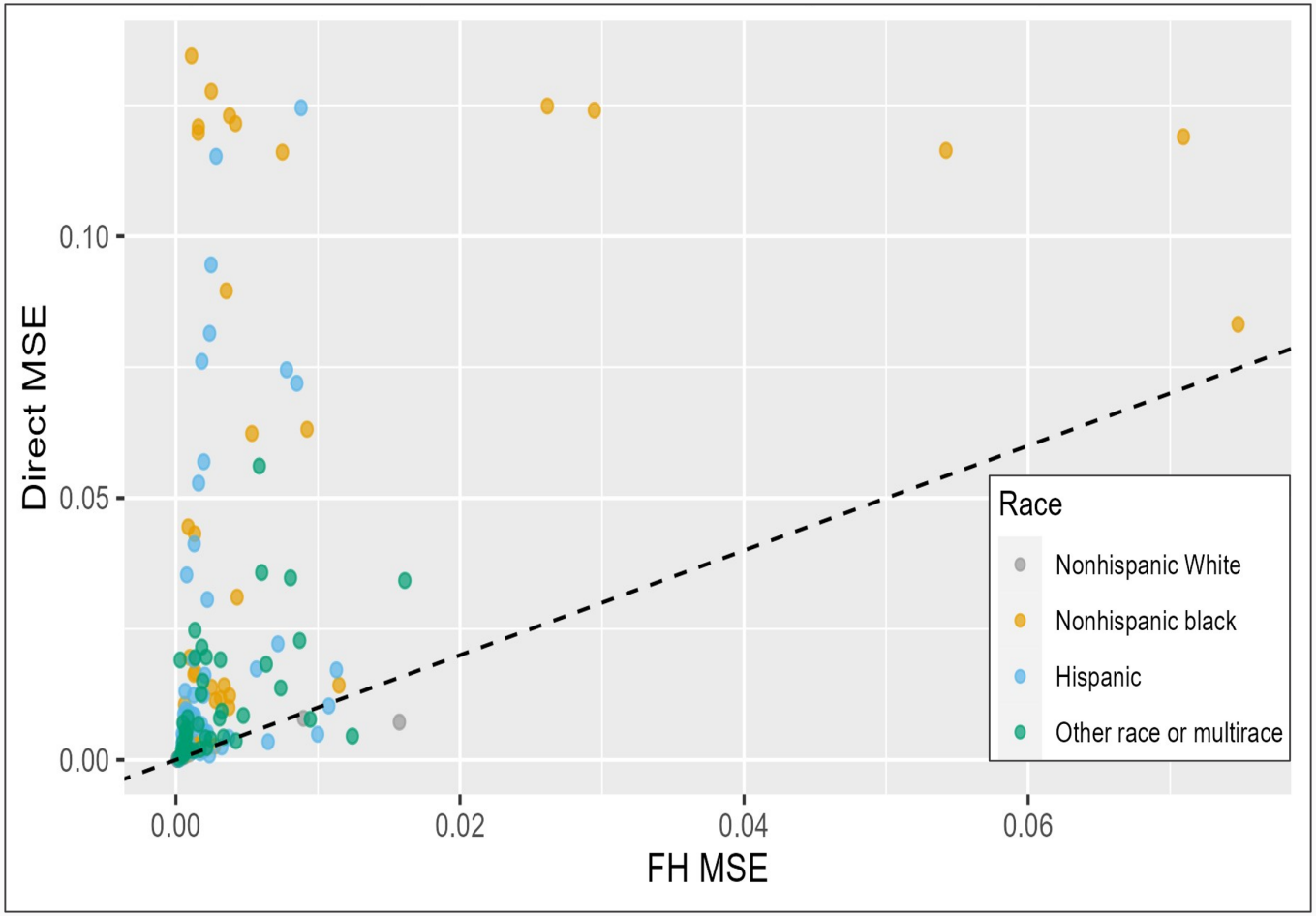

Figure 3: MSEs for Direct vs FH estimate; Model: stepwise covariates with fixed effects; Design: PPS 1% sampling fraction

“Random” sample of recommendations

R: Have a diverse multi-disciplinary team and make your research community engaged

Ideally projects should include, in all stages of the research process, a combination of:

- Subject matter experts
- SAE experts
- Community stakeholders
- A diverse, multi-disciplinary team

R: Consider the SAE strategy and the equity research goals at the survey design stage

- SAE often done as an afterthought, but whenever possible estimation goals should be considered during design (e.g., Singh et al 1994, Marker 2001)
- Particularly important when objective is equity research
- Consider including demographic groups as planned domains and oversampling even if planning to do SAE
- See the literature on sampling "hard to reach populations" (e.g., Tourangeau et al. 2014, Kalton 2009, Kalsbeek 2003, Dutwin and Lopez 2014, Burgard et al. 2016)
- Literature about optimal sampling design for SAE goals is available and could be extended for this use (e.g., Longford 2006, Choudhry et al. 2012, Keto and Pahkinen, 2017, Molefe and Clark 2015)

R: Consider smoothing the sampling variances

- Especially important with small, unplanned domains which often occurs when cross-classifying by demographic characteristics
- e.g: Rein, Franco, et al. (2024) point out an example where ACS yields an estimated effective sample size that is magnitudes larger than the world's total population!
- More research needed on smoothing the variance for SAE in general, but an interesting recent paper is You and Hidiroglou (2023)
- One can also jointly model the sampling variance estimator and direct estimator (e.g, You and Chapman 2006, Dass et al. 2012; Maiti et al. 2014; Sugasawa et al. 2017)

R: Evaluate and address measurement error in auxiliary information

ME in the covariates in SAE models should be evaluated, e.g.,

- Tendencies to underdiagnose certain conditions for minorities means administrative records can have a bias (e.g., diagnosis of dementia, Gianattasio, 2019)
- Other quality issues with admin records (e.g., differences in accuracy among races in death certificates, Arias, Heron, Hakes, 2016).
- Differential errors in data linkages among races (e.g., Lariscy 2017; Black et al. 2017; Bohensky et al. 2010)
- Differential errors in imputation (Brown et al. 2022)

Measurement error continued

What to do about ME in SAE models in this context?

- Erciulescu, Franco, and Lahiri (2021) has a good discussion of what qualities to look for in covariate for used in models
- There is ample literature on ME for SAE models (e.g. Ybarra and Lohr, 2008, Bell, Datta, Franco, 2019, Arima, Bell, Datta, Franco, Liseo, 2017) , but most assume that ME has known variance
- Certain model assumptions could potentially alleviate ME problems (i.e., more flexible models that allow for different intercepts, etc.)
- See also NAS 2023 report, Section 3.5, "Assess and Reduce Measurement Error"

Recommendations on modeling

- R: Beware of "shrinkage to the mean" when the mean comes primarily from data from the majority group (applies also when you have very weak covariates)
  - For example, in our simulations, the MSE of the estimate of the difference in home ownership between NHW and NHB tended to be higher than the MSE of the corresponding direct estimator when using a pure shrinkage model
  - Having strong covariates or race fixed effects can help
- R: Give careful thought to model assumptions and whether they are realistic for all groups being studied
  - i.e., measurement error, is the model flexible enough, etc.
- R: The best models for prediction are not always best for interpretation of parameters or casual inference. Be aware of the main goal of the study and design your modeling strategy accordingly (see for instance Shmueli 2010)

Impact of some modeling choices in home ownership estimation according to simulation study

Model/design	eq_1	eq_gss	pps_1	pps_gss
Separate models	0.0009	0.0040	0.0023	0.0075
Joint model with FE	0.0008	0.0024	0.0013	0.0032
Joint model no FE	0.0009	0.0029	0.0014	0.0035

Table 2: Median MSE in simulation study, estimates of home ownership by state for **Hispanics**, models with hypothesized covariates; comparison of modeling home ownership by race/ethnicity separately, jointly, and with and without fixed effects

The importance of model validation

- R: Allow time for model selection, diagnostics, and validation. Using social or demographic groups to cross-classify brings additional challenges
  - Paper contains discussion of several techniques with discussion on how they apply to this setting, but more research is needed
  - Some assumptions hard to check
  - Apart from more formal techniques, external validation whenever possible
  - Importance of common-sense checks, involving subject matter experts

Comments

- Sample of content of paper, more recommendations, discussion about each recommendation, examples, are in the paper
- More research is needed!
- Intention is to foster more discussion and raise awareness of challenges
- In SAE, no "one-size-fits-all" solutions, but it's important to be aware about benefits and pitfalls
- What are others?
- Please reach out with questions, comments, or related works that should be covered in our paper

Thanks for your attention!

Questions?

Franco-Carolina@norc.og

NORC at the University of Chicago