SAIPE: Poverty Mapping in the United States

Carolina Franco

U.S. Census Bureau
Center for Statistical Research and Methodology

ECLAC-UNSD Webinar on Poverty Mapping Using Small Area Estimation Techniques, July 1, 2021

About Small Area Estimation (SAE)

- Cross-classification (i.e., geographic, demographic) often leads to small sample sizes even in very large surveys.
- Surveys often cannot estimate all quantities of interest through "direct" methods with acceptable accuracy.

  Direct Estimator: *based on sample data for domain of interest alone.*

  Small Area: *domain where sample size is too small for reliable direct estimation*.
- SAE: Through modeling, incorporate information from other domains and auxiliary data sources to "borrow strength."

Examples of sources to borrow information from in Small Area Estimation

- **Administrative Records**– See Erciulescu, Franco, and Lahiri (2021). Use of administrative records in small area estimation
- **Censuses**
- **Same survey, different year**
- **Other surveys** e.g., Franco and Bell (2021)
- **Commercial data, satellite data, cell phone data, etc.**

Poverty estimation at the Census Bureau

- The U.S. Census Bureau's SAIPE (Small Area Income and Poverty Estimates) program estimates poverty for various age groups by levels of geography.
- In the US, a **family**, and all individuals from the family, are in poverty if their **total money income** (pre-tax) is less than the poverty threshold for the family size and age composition.

Poverty thresholds 2020

https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html

Size of family unit	Related children under 18 years								
	None	One	Two	Three	Four	Five	Six	Seven	Eight or more
One person (unrelated individual):									
Under age 65	13,465								
Aged 65 and older	12,413								
Two people:									
Householder under age 65	17,331	17,839							
Householder aged 65 and older	15,644	17,771							
Three people	20,244	20,832	20,852						
Four people	26,695	27,131	26,246	26,338					
Five people	32,193	32,661	31,661	30,887	30,414				
Six people	37,027	37,174	36,408	35,674	34,582	33,935			
Seven people	42,605	42,871	41,954	41,314	40,124	38,734	37,210		
Eight people	47,650	48,071	47,205	46,447	45,371	44,006	42,585	42,224	
Nine people or more	57,319	57,597	56,831	56,188	55,132	53,679	52,366	52,040	50,035

Source: U.S. Census Bureau.

Poverty statistics produced by SAIPE

- All people in poverty (state and county)
- Children under age 18 in poverty (state and county)
- **Related children aged 5-17 in poverty** (state, **county***, and school district)
- Children under age 5 in poverty (states only)
- Supports the Elementary and Secondary Education act of 1965 (reauthorized by the Every Student Succeeds Act of 2015)

United States Census Bureau

County school-aged (5-17) children poverty model

- SAIPE uses **area-level model** (Fay Herriot)–models direct estimates at the domain level
- Alternative: **unit-level models** model unit level data, and typically require having covariates for all units in the population
- Main data source is American Community Survey, also uses administrative records and 2000 Census long form

United States Census Bureau

About the American Community Survey (ACS)

- Approximately 3.5 million addresses per year.
- Questions about demographics, income, health insurance, education, disabilities, etc.
- Complex survey (stratification, clustering of people within households, sub-sampling of initial non-respondents).
- Survey-weighted estimates.
- 1-year and 5-year estimates produced annually.
- Supplanted the census long form (sent to about 1/6 of the population during decennial census).

The Fay-Herriot Model (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, generally assumed to be $N(0, v_i)$, independent with v_i known.
- u_i is the area i random effect, usually assumed to be *i.i.d.* $N(0, \sigma_u^2)$ and independent of the e_i.

More on the Fay-Herriot Model

- Best linear 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 is placed on y_i.
- Hierarchical Bayes or empirical Bayes fitting.

The SAIPE 5-17 county production poverty model

A univariate Fay-Herriot Model:

- y_i = log of the ACS estimate of the number of children age 5-17 in poverty for county i.
- Y_i = log of the corresponding true quantity.
- β and σ_u^2 are estimated by ML.
- $\mathbf{x}_i$ is the regressor variable vector on the log scale.
- Prediction results are translated back from the log scale using properties of the lognormal distribution.

The SAIPE 5-17 county production poverty model–regression variables

On the log scale, for each county, intercept plus:

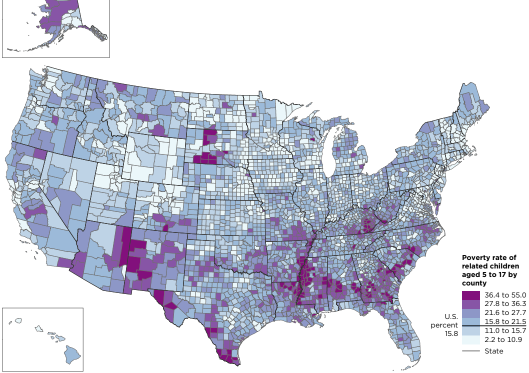
- Number of "poor child exemptions" (child exemptions on tax returns with incomes below the poverty line).
- Number of Supplemental Nutritional Assistance Program (SNAP) benefits recipients.
- Estimated population age 0-17 from Population Estimates Program.
- Number of child tax exemptions.
- Census 2000 estimate of the number of school-aged (ages 5 to 17) children in poverty.

The state and school district methodologies

- State model: FH model directly on the poverty rate, with no log transformation
- Bayesian implementation to avoid model error variance estimates of 0.
- Lower geographic levels ratio-adjusted to sum up to higher geographic levels (school district estimates sum to county estimates, county estimates sum to state estimates, state estimate sum to ACS national estimate)
- The school district estimation methodology does not make use of a formal model. Will not be discussed here due to lack of time.

County poverty map

Figure 5.
Poverty Rate of the School-Age Population by County: 2019

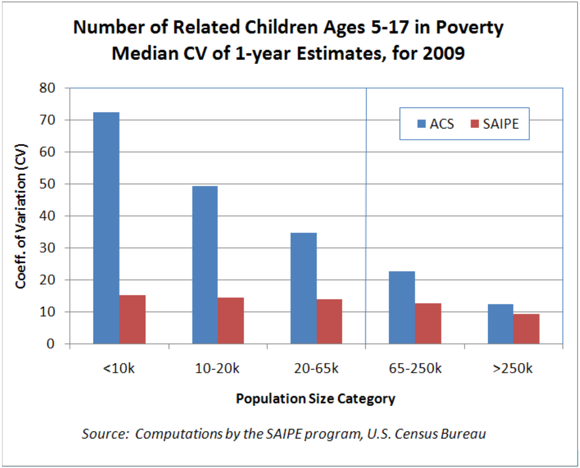


Note: The data provided are indirect estimates produced by statistical model-based methods using sample survey, decennial census, and administrative data sources. The estimates contain error stemming from model error, sampling error, and nonsampling error.
Source: U.S. Census Bureau, Small Area Income and Poverty Estimates (SAIPE) Program, December 2020.

The SAIPE interactive poverty map

In the SAIPE website, you can find the SAIPE interactive tool: `https://www.census.gov/data-tools/demo/saipe`
There, you can see poverty maps filtering by age groups and geography, as well as graphs about the time series of poverty rates

County reductions in coefficients of variation



Challenges and recent/current research

- The next few slides will discuss some challenges and recent research
- The methods that follow are **not** currently part of official production

Challenges/selected research, related to SAIPE 5-17 county model

- Need to drop counties with zero estimates ($\approx 5\%$) due to log transformation; lack of good estimates of sampling variances.

  Potential Solution: Use a Generalized Variance Function (GVF) to produce estimates of sampling variances. Model rates directly rather than log counts. See Maples (2011), Franco and Bell (2013), and Franco (2020).

- Data are inherently discrete/possible improvements to normality assumptions (?).

  Potential Solution: Consider other GLMMs, such as Binomial/Logit Normal (BLN) model. See Franco and Bell (2013, 2015), Franco (2020)

Challenges/research continued

- Long form discontinued in 2000, covariate becoming outdated *Potential Solution: Consider borrowing information from past ACS estimates instead.*
- **Note**: Using survey estimates directly as covariates, without accounting for sampling error, can result in suboptimal predictions, incorrect estimation of MSEs. (See Bell, Chung, Datta, and Franco, 2019)
- Sampling error can be captured through bivariate or measurement error models (e.g., Huang and Bell, 2012, Franco and Bell, 2013, 2015, or Arima, Datta, Bell, Franco Liseo, 2019).
- Can also borrow strength from past ACS via temporal models (e.g., Franco and Bell, 2015, Taciak and Basel, 2012)

United States Census Bureau

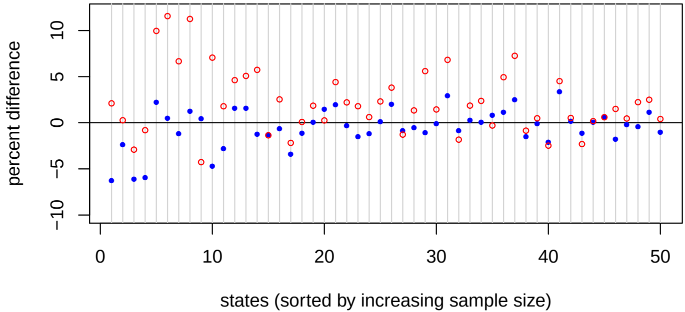
Binomial/Logit Normal (BLN) model

- y_i sample count, n_i sample size, p_i true proportion

$$y_i | p_i, n_i \sim \text{Bin}(n_i, p_i) \qquad i = 1, \ldots, m \tag{1}$$
$$\text{logit}(p_i) = \mathbf{x}_i' \beta + u_i \tag{2}$$

- $\text{logit}(p_i) = \log[p_i/(1 - p_i)]$, $u_i \overset{i.i.d}{\sim} N(0, \sigma_u^2)$.
- May be more appropriate for discrete data
- Naturally handles zero estimates, skewness.
- Complex sampling can be addressed by using **effective sample size**
- Can be readily extended to bivariate, temporal (e.g. AR(1))

Comparison of bivariate BLN and production (unraked) county estimates summed to state



BLUE=Bivariate BLN (borrows strength from previous non-overlapping 5-year ACS estimates and AR covariates)
RED=SAIPE (unraked)

Other ways of borrowing strength from past

- We compared temporal (AR(1), AR(5)), and bivariate BLN/FH models as alternative ways to borrow strength
- The time series models jointly model several one-year ACS estimates
- We found the performance of the two were similar, perhaps slightly better for bivariate model
- However, time series models facilitate estimating year to year changes
- For more on these comparisons, see Franco and Bell, 2015

Some final remarks

- SAIPE is a great example of a successful small area estimation program
- By leveraging other data sources (e.g. tax records), SAIPE is able to provide improved estimates that "borrow strength."

Where to read more about SAIPE

- The SAIPE website:
  `www.census.gov/programs-surveys/saipe.html`
- Bell, W. R. , Basel, W. W, and Maples, J. J. An overview of the US Census Bureau's Small Area Income Poverty Estimates Program. In "Analysis of Poverty by Small Area Estimation" (2016), edited by Monica Pratesi. New York, Wiley

Other references from this talk

- Arima, S., Bell, W. R., Datta, G. S., Franco, C., and Liseo, B. (2017). Multivariate Fay-Herriot hierarchical Bayesian estimation of small area means under functional measurement error. Journal of the Royal Statistical Society–Series A. 180 (4), 1191-1209
- Bell, W. R., Chung, H. C., Datta, G. S., and Franco, C. (2019). Measurement error in small area estimation: Functional versus structural versus naive models. Survey Methodology, 45, 61-80.
- Erciulescu, A., Franco, C., and Lahiri, P. (2021). Use of administrative records in small area estimation. Chun, A. Y. and Larsen, M. (Eds.) Administrative records for survey methodology. New York: Wiley

References continued

- Franco, C. (2020). Comparison of small area models for estimation of U.S. county poverty rates of school-aged children using an artificial population and a design-based simulation. Census Bureau Research Report Series RRS2020/04. Available online athttps://www.census.gov/library/working-papers/2020/adrm/RRS2020-04.html
- Franco, C. and Bell, W. R. (2021). Using American Community Survey data to improve estimates from smaller U.S. surveys through bivariate small area estimation models. To appear in Journal of Survey Statistics and Methodology
- Franco, C. and Bell, W. R. (2015). Borrowing information over time in binomial/logit normal models for small area estimation. Joint issue of Statistics in Transition and Survey Methodology. 16, 4, 563-584.

United States Census Bureau

References continued

- Franco, C. and Bell, W. R. (2013). Applying bivariate/logit normal models to small area estimation. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 690-702.
- Huang and Bell (2012) An empirical study on using previous American Community Survey data versus Census 2000 data in SAIPE models for poverty estimates. Research Report RRS2012/04. Center for Statistical Research and Methodology, U.S. Census Bureau

References continued

- Maples (2011) Using small area modeling to improve design-based estimates of variance for county level poverty rates in the American Community Survey. Research Report RRS2011/02. Center for Statistical Research and Methodology, U.S. Census Bureau
- Taciak, J. and Basel, W. (2012). Time series cross sectional approach for small area poverty models. *Proceedings of the American Statistical Association, Section on Government Statistics*

United States Census Bureau

Questions? Preguntas?

Carolina.franco@census.gov
(English or Spanish)

Disclaimer

This presentation is to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.