UAA Institute of Social and Economic Research
University of Alaska Anchorage

Using Administrative Data to Improve ACS Small Area Estimates

Experimental Synthetic Data for Rural Alaska Communities

Matthew Berman[1]
Lance Howe[2]

2019 ACS Data Users Conference
American University Washington College of Law
May 14-15, Washington, DC

[1] Institute of Social and Economic Research, University of Alaska Anchorage; 907 786 5426, matthew.berman@alaska.edu

[2] Department of Economics, University of Alaska Anchorage; 907 786 5409, elhowe@alaska.edu

Challenge with ACS small-area estimates: Balancing data privacy vs. usability

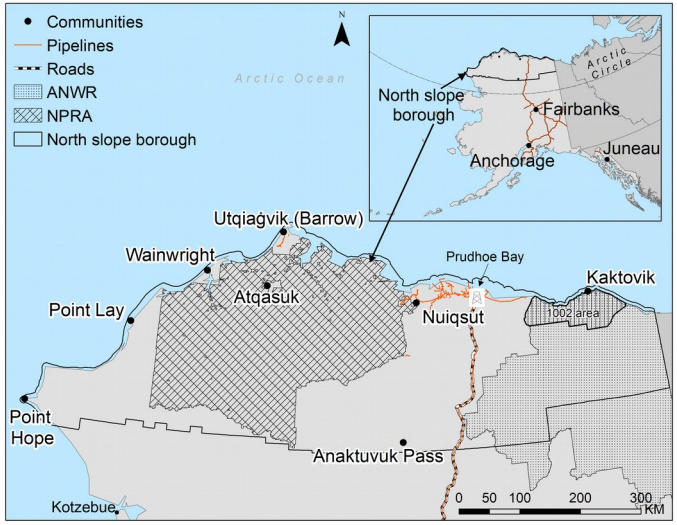
Census Bureau historical approach

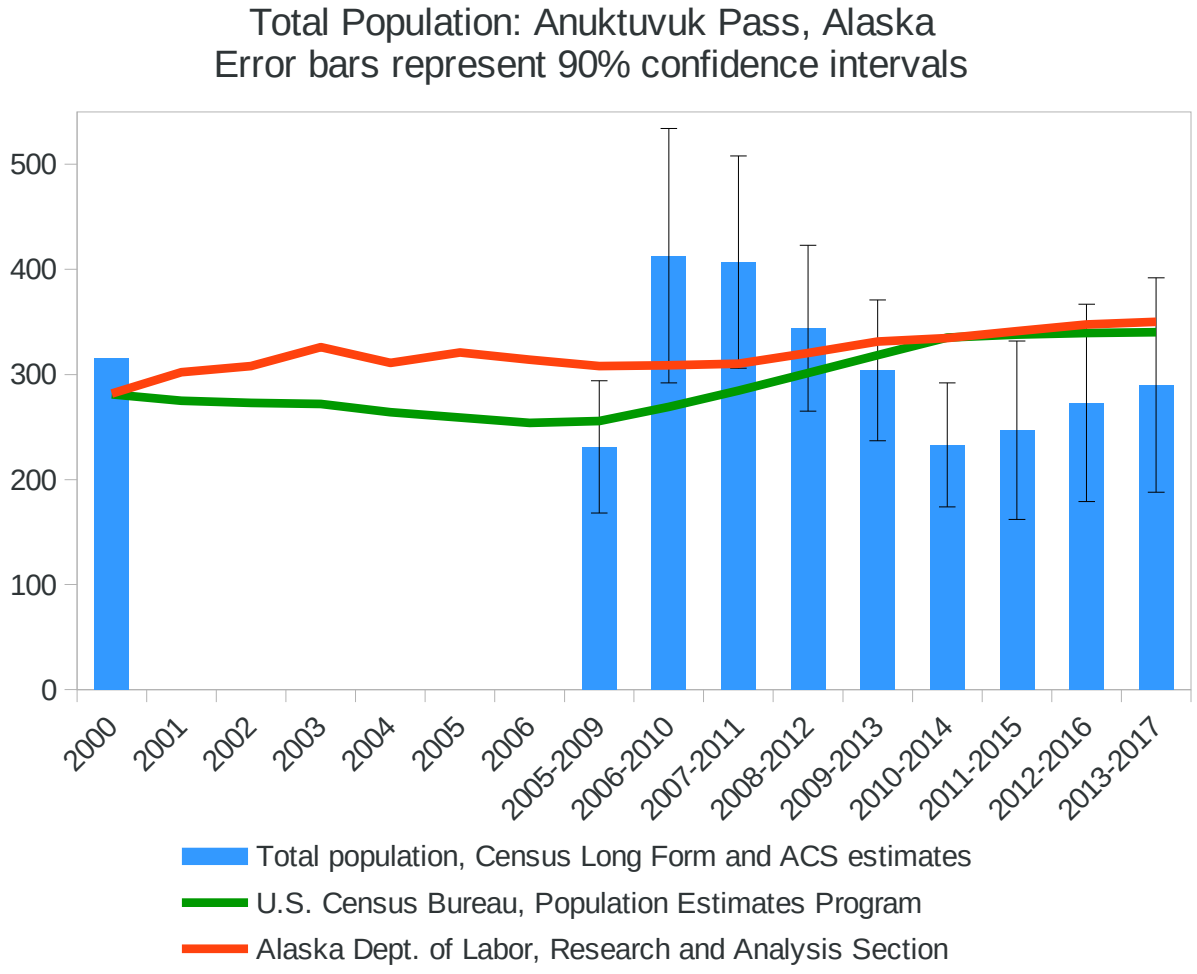
- Randomly select a small number of housing units for sampling each year
- Replace survey non-response items with data from randomly selected “similar” individuals (imputation)
- Apply additional obfuscation techniques: top and bottom coding, swapping, etc.
- Add edited survey responses to generate total counts by community
- Report 5-year moving averages

Multiple Problems with the Census Bureau Approach to Small Area 5-year Moving Average Estimates

- The total population is estimated from the survey responses, generating high margins of error for smaller communities.
- ACS sample sizes are smaller than Census Long Form sample sizes: the small number of households randomly selected for interviews, even across a 5-year period, may not represent the community as a whole.
- Social and economic conditions change from year to year, adding to the margin of error in moving average estimates.
- Limited and problematic choices for imputing values for respondents who don't answer certain questions

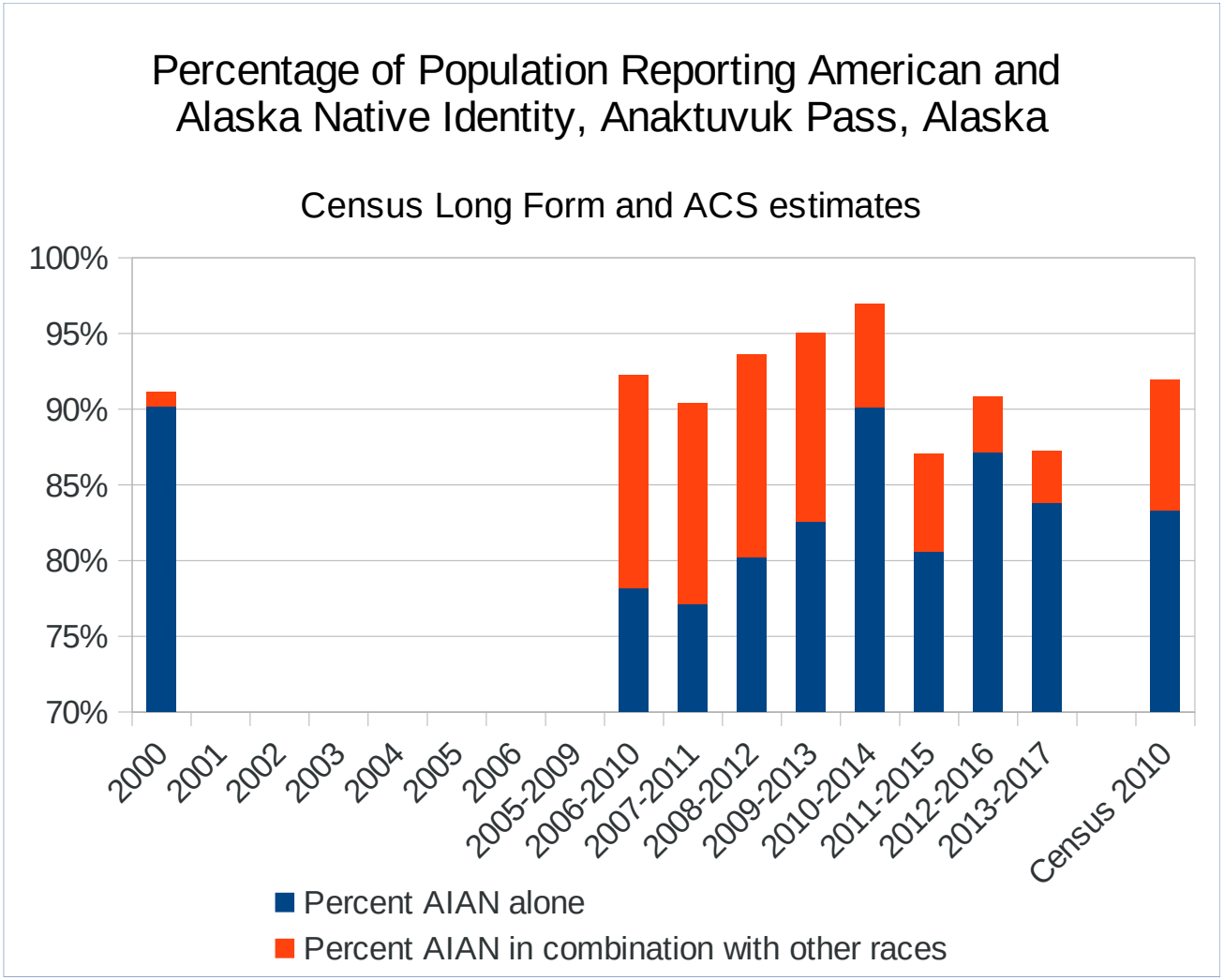
Strike 1: Population is unknown and estimated from the survey response rate





The Census Bureau has a good program to estimate annual small-area populations; so does the Alaska Dept. of Labor, using independent sources. This information is not used for ACS estimates.

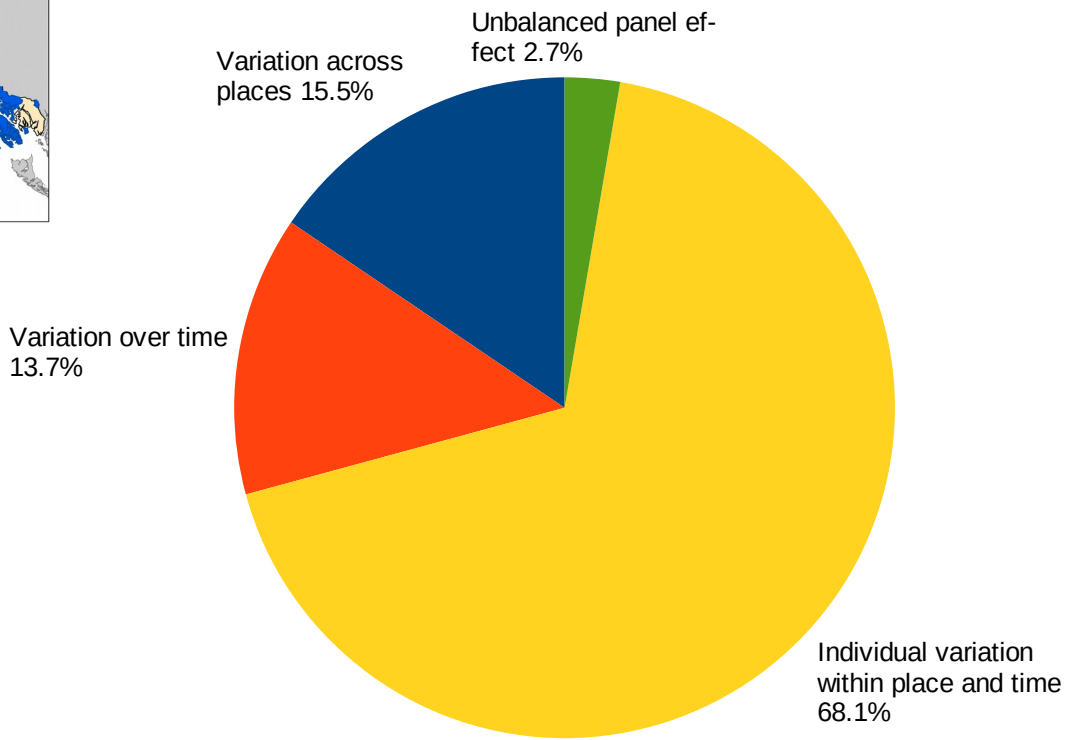
Strike 2: People interviewed each year differ in lots of ways



- Population characteristics likely change gradually over time
- Information on repeated sampling of the population not used

Strike 3: Conditions change systematically annually

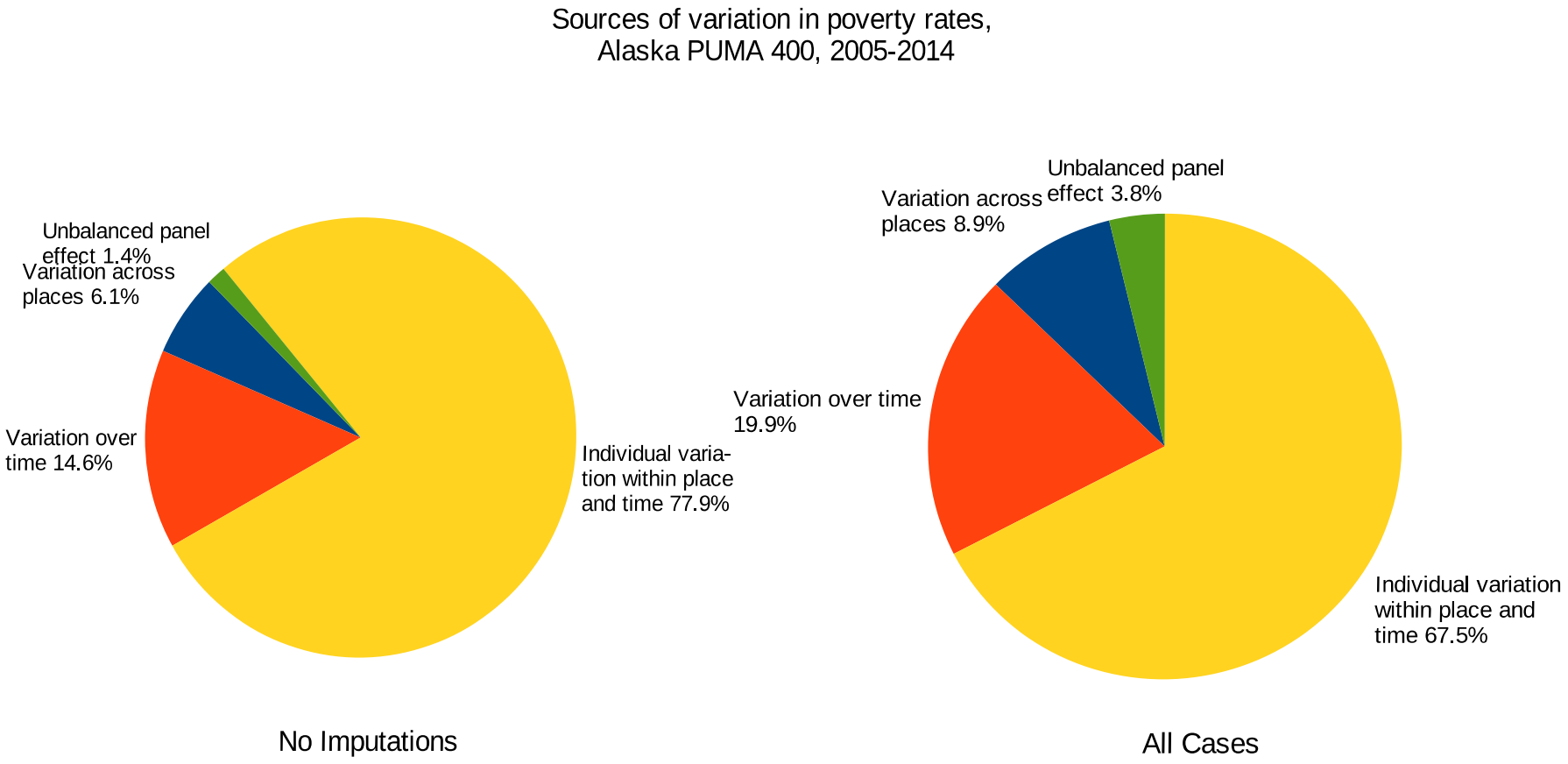
Sources of variation in household income,
Alaska PUMA 400, 2005-2014



- National and regional economic fluctuations: information ignored
- changing state budgets affecting social outcomes: information ignored

Strike 4: Imputation adds error

Higher time variation relative to place variation in imputed values

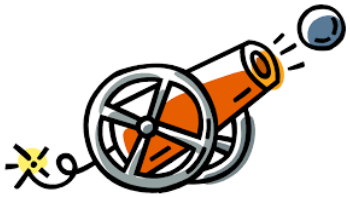


- Assumptions about whose answer to use are poorly documented
- Small sample means that many imputed values have to be drawn from households in other communities
- Information about community differences relevant to imputation not known

The data privacy challenge in the era of Big Data

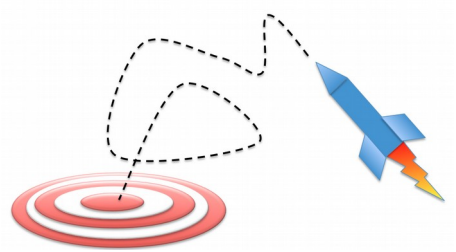
- The Census Bureau has figured out that a sophisticated statistician with a fast computer and lots of time can possibly identify individual responses from the 5-year moving average estimates.

- CB approach to data privacy: add random noise to further obfuscate the data



- Noise infusion is a crude weapon against an adversary that exists only in theory at this point
  - Data usability is already severely compromised, and will degrade further
  - Individual responses will still be identifiable in the published results; however, now they will be incorrect.

Synthetic data: A better approach for producing ACS small-area estimates



- Utilize the wealth of information available from multiple administrative sources to supplement and interpret ACS survey results
- Regression-based estimates of community conditions
- Potential outcomes:
  - More information about contemporary conditions in small areas
  - Better data privacy and improved data usability

Synthetic data approach

- **Use administrative data to estimate community populations**, not ACS sample data
  - Use Census Bureau population estimates program to estimate annual small-area population
  - Use information from decadal census counts to provide benchmark housing, occupancy, and household demographic information
- **Estimate rates, not counts**, from ACS surveys
  - Use co-varying characteristics of ACS sample households to estimate rates independent of population size
  - Multiply estimated rates by estimated population in step 1 if counts are desired
  - Assume rates change systematically, not randomly over time, in response to demographic, economic, and social processes revealed by the sample data
- **Use local area administrative data** to separate systematic changes (to be reported) from random variation (sampling error to be ignored)
  - State and local employment data, BEA LAPI, tax returns, provide economic data
  - School enrollments, social program enrollments, provide social data
- **Apply statistical methods to impute non-response** rather than random substitution

Synthetic data approach: technical details

- Model: indicator y_{ijt} for individual, household, or family *i* in community *j* in year *t* is assumed to be a function of individual and household demographic characteristics (measurable in the decennial census), *x*, community indicators, *z,* a community-specific error term, *u,* and random sampling error, *ε*:

$$y_{ijt} = f\left(x_{ijt}, z_{jt}, t\right) + u_{jt} + \varepsilon_{ijt} \qquad u_{jt} \sim N\left(0, \sigma_u^2\right) \qquad \varepsilon_{ijt} \sim N\left(0, \sigma_\varepsilon^2\right)$$

- Using the ACS survey observations for *x* and *y*, and publicly available administrative data for z, estimate the relationship *f* (estimated function *g*) and estimated community-specific random effect *v*, and use the predicted values to generate synthetic annual community-level indicators *s* with a variance estimate σ_ε^2 , which can be considered as a Bayesian prior for y_{jt}:
-

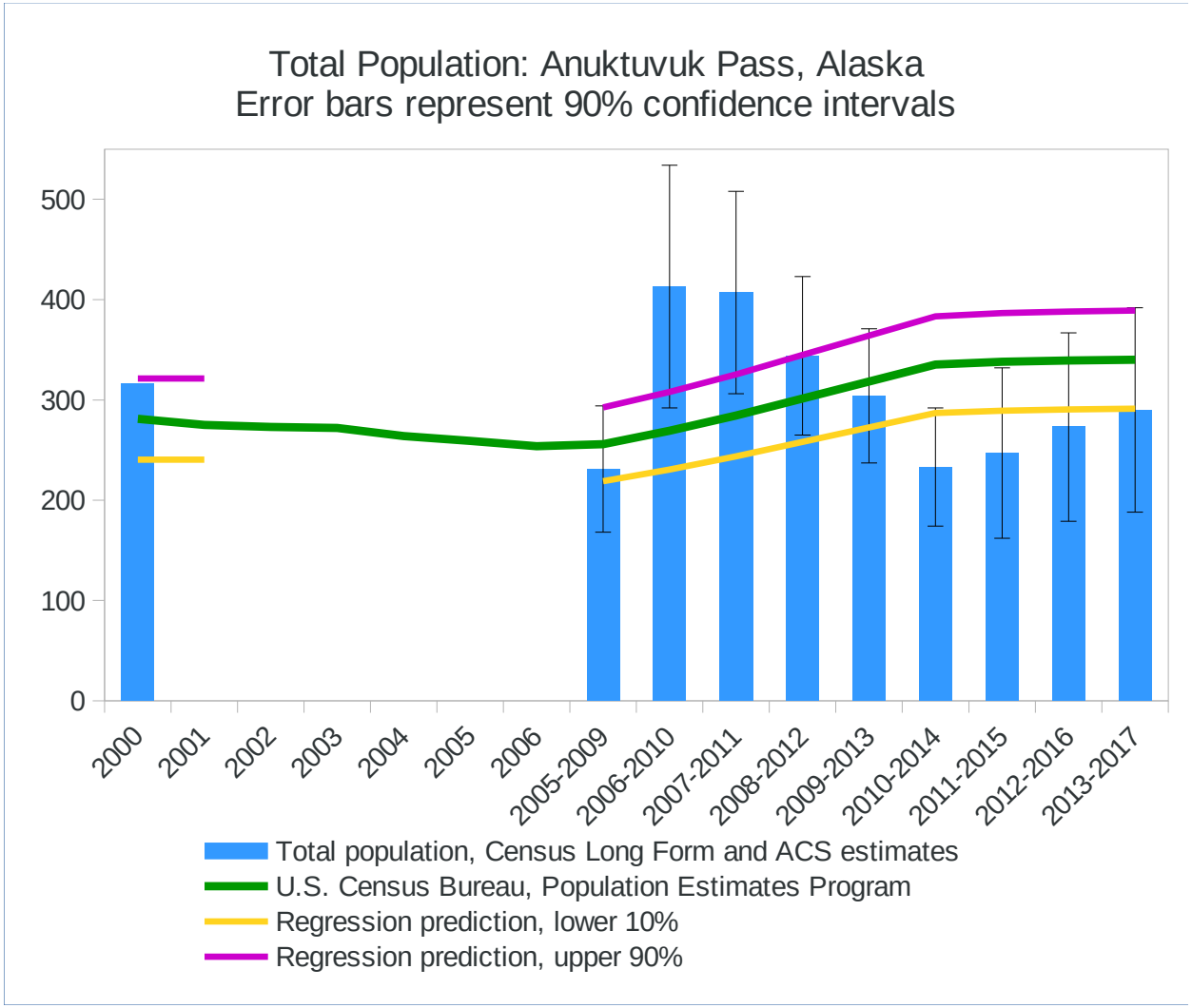
$$s_{jt} = \sum_{i=1}^{n_t} g\left(x_{ijt}, z_{jt}, t\right)/n_t + v_{jt}$$

- Using the sample mean, $\bar{y}_{jt} = \sum_{i=1}^{n_t} y_{ijt}/n_t$ with its sampling variance estimate σ_{jt}^2 adjust s_{jt} to generate the Bayesian posterior distribution,

$$\bar{s}_{jt} = g\left(\bar{x}_{jt}, z_{jt}, t\right) + v_{jt} + \mu_{jt} \quad \text{where} \quad \mu_{jt} = \frac{\sigma_{jt}^2 s_{jt} + \sigma_\varepsilon^2 \bar{y}_{jt}}{\sigma_{jt}^2 + \sigma_\varepsilon^2}$$

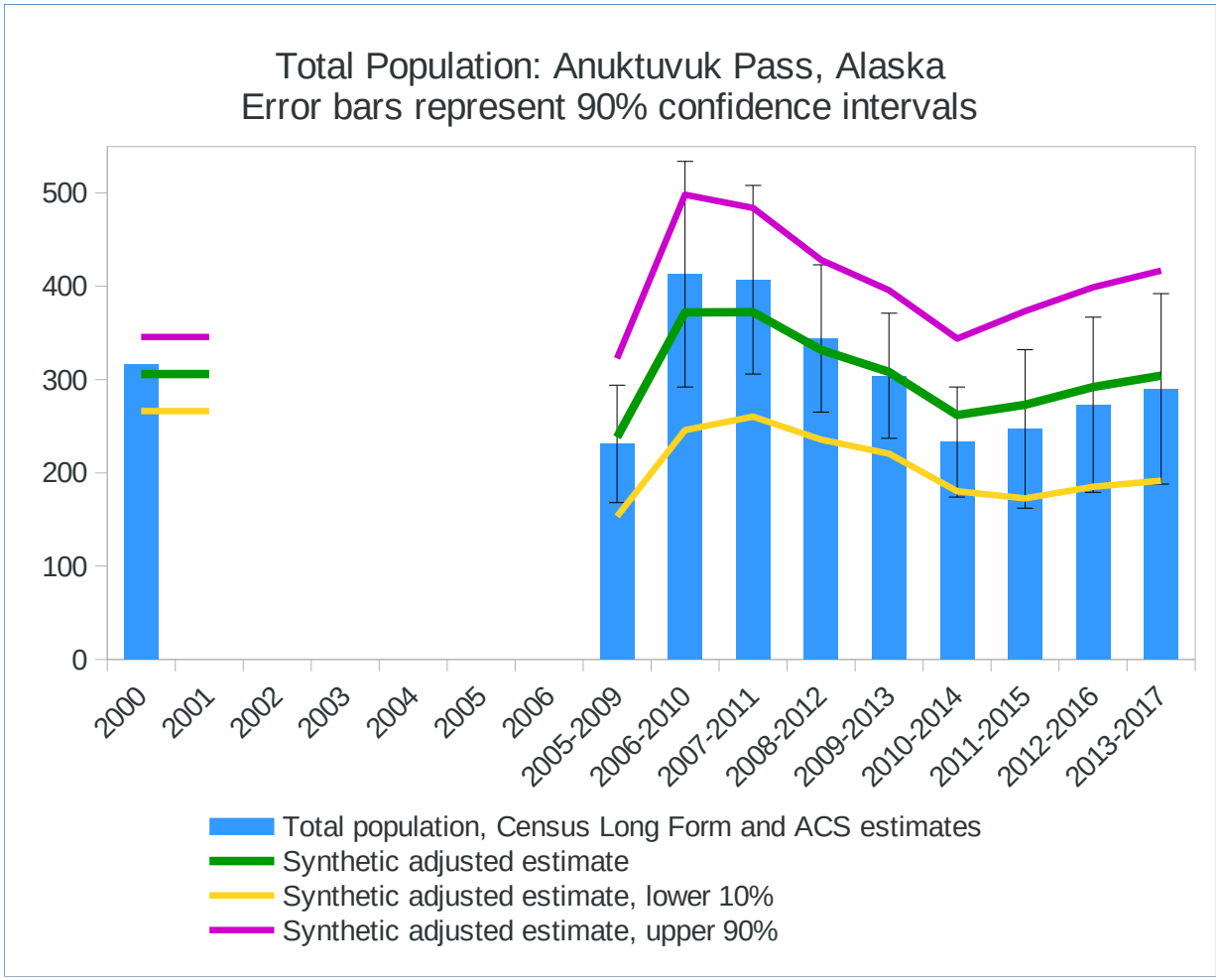
Illustrate method with a simple example, population of Anaktuvuk Pass

Population, $y_t = f(x,z,t) = z_t$, using Census population estimate for z_t:



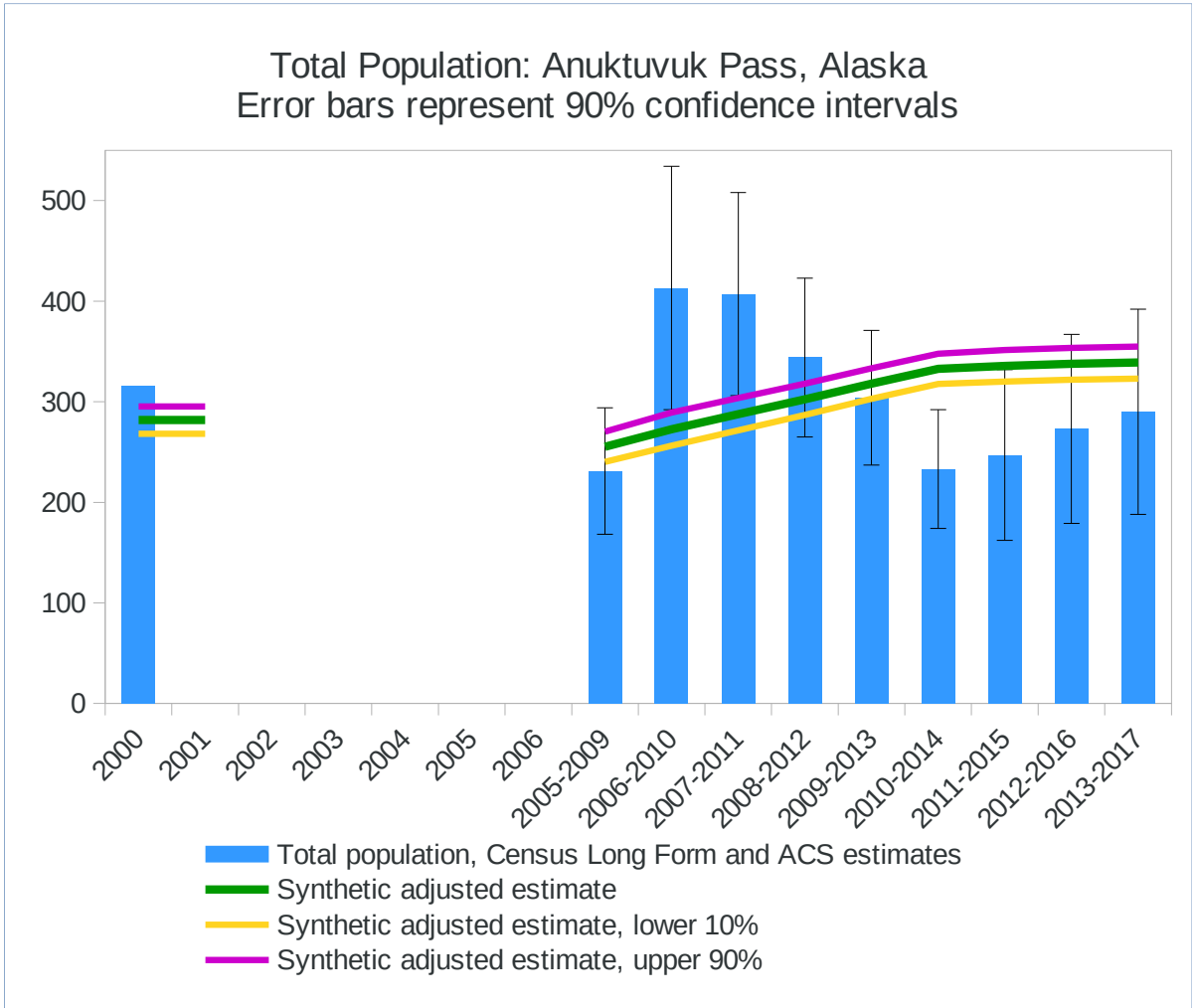
Illustrate method with a simple example, population of Anaktuvuk Pass

Population estimate = $s_t + \mu_t$. Note that this is not very accurate: only 10 observations for the synthetic estimates yields a high margin of error.



Illustrate method with a simple example, population of Anaktuvuk Pass

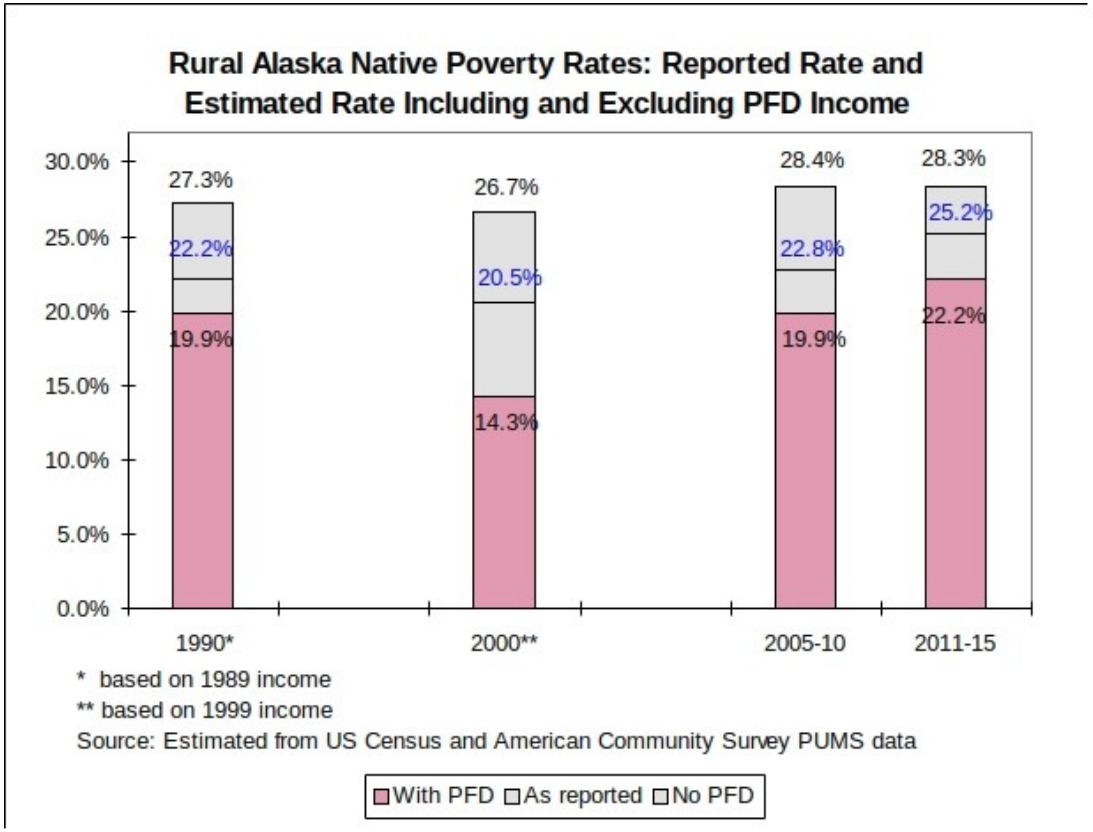
Population estimate = $s_t + \mu_t$. If we had 100 communities (which we do in PUMA 400, MOE for synthetic estimates would decline by 90%, which would yield:



Final note: we have only discussed sampling error. Administrative data can reduce non-sampling error, too.

Two examples using Alaska Permanent Fund Dividend applications:

- ACS estimates of out-migration from Alaska overestimated by 100 percent – helped Census Bureau find and fix a coding error
- Income of children not counted, resulting in large overestimate of poverty in all Alaska, especially among children



UAA Institute of Social and Economic Research
University of Alaska Anchorage

Thank you!

The authors gratefully acknowledge financial support for this research from the National Science Foundation, award #1216399, and logistical support provided by the U.S. Census Center for Economic Studies, and the California Census Research Data Center at the University of Southern California.

USC University of Southern California

United States Census Bureau

FSRDC

NSF