Small Area Income and Poverty Estimates Program[1]

Donald M. Luery, U.S. Census Bureau[2]

Summary

The U.S. Census Bureau's Small Area Income and Poverty Estimates program (SAIPE) uses an empirical Bayes estimation method to produce annual estimates of the poverty rates and counts of poor school age children for states and counties. The dependent variable is from the American Community Survey (ACS) and the predictor variables are from administrative data sources from other Federal agencies along with estimates from the decennial census and the population estimates program. The No Child Left Behind Act of 2001 (NCLB) directs the Department of Education to distribute Title I grants directly to school districts on the basis of the most recent estimates of children in poverty available from the Census Bureau. The SAIPE program produces these estimates for the Department of Education using a synthetic estimation method controlled to the county SAIPE estimates. This paper discusses the data sources and difficulties encountered in their compilation; the estimation methods for states, counties, and school districts; and a summary of our plans for future research.

1. Introduction

Prior to the creation of the SAIPE program, the decennial census long-form was the only source of income distribution and poverty statistics for households, families, and individuals if one needed data for "small" geographic areas, i.e., counties, cities, and other sub-state areas. The ten-year span between censuses left a large gap in information concerning the economic situations of local areas. In the 1990s, federal agencies and the U.S. Congress asked the Census Bureau to develop intercensal estimates of income and poverty. In September 1994, Congress passed the Improving America's Schools Act (PL 103-382) of which Title I specified that the distribution of federal funds be made to school districts based largely on "the number of children ages 5 - 17, inclusive, from families below the poverty level on the basis of the most recent satisfactory data ... available from the Department of Commerce."

This law further required the Secretary of Education to use updated Census Bureau data on school-aged children in poverty for counties starting with allocations for the 1997-98 school year, and for school districts beginning in the 1999-2000 school year, unless the Secretaries of Education and Commerce determined that the use of the updated estimates would be "inappropriate or unreliable." The No Child Left Behind Act of 2001 (NCLB) directs the Department of Education to distribute Title I basic and concentration grants directly to school districts on the basis of the most recent estimates of children in poverty available from the Census Bureau. The U.S. Department of Education allocated Title I funds of about $14.5 billion dollars in 2009.

The U.S. Census Bureau developed the Small Area Income and Poverty Estimates (SAIPE) program to produce: 1) annual reliable median household income estimates and poverty by age group estimates for U.S. states and counties during intercensal periods and 2) timely and reliable estimates of income and poverty for the administration of federal programs and the allocation of federal funds to local jurisdictions and state and local programs that depend on income and poverty estimates. Annually SAIPE produces for states and counties estimates of the number of persons in poverty, children under 18 in poverty, related children ages 5-17 in poverty, children under age 5 in poverty (for states only), and median household income. For school districts, SAIPE produces estimates of the total population, children ages 5-17, and related children ages 5-17 in poverty. The SAIPE population and poverty estimates of school-aged (ages

[1] Presented at the 27th SCORUS (Standing Committee for Regional and Urban Statistics) Conference in Riga, Latvia

[2] This paper is released to inform interested parties and any views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

5 - 17) children for school districts are key inputs for the Department of Education's Title I allocation process.

The first release for the county and state estimates was in 1997 for calendar year 1993 and the first release for school district estimates was in 1999 for calendar year 1995. The state and county estimates have been produced using Fay-Herriot (1979) models fitted to direct income and poverty estimates initially from the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS). Beginning with the 2005 release, the SAIPE estimates have used the American Community Survey (ACS). The CPS ASEC is designed for estimates at the national level and its sample size (about 100,000 addresses) is not large enough to publish reliable direct survey estimates except for the largest states and counties. As the ACS has a much larger sample size (about 3 million addresses) than the CPS ASEC and much wider county coverage, the ACS data supports direct survey estimates for much smaller areas than will the CPS ASEC. The ACS supports single-year direct estimates for counties and other places with populations of 65,000 or more. In 2006, the Census Bureau switched the source of the official direct survey state and substate estimates of income and poverty from the CPS ASEC to the ACS. The CPS ASEC remains the source of the official direct national estimates of income and poverty. Following the Census Bureau's switch of new methods investigated by SAIPE and an external review that took place in 2007, we adopted ACS as the survey source for the SAIPE state and county estimates.

Due to a lack of data for modeling at the school district level, the school district estimates have been produced using a simple updating scheme that takes the estimated shares of the number of school-aged children in poverty among school districts within each county and multiplies these by the current year's county level SAIPE estimates of the number of 5 -17 year old children in poverty. These shares are derived from the previous decennial census in conjunction with the latest Internal Revenue Service (IRS) data on child exemptions below the poverty thresholds. The shares approach used for the school district estimates ensures that they aggregate to the county estimates, which are first raked so they add to their corresponding state estimates. The state estimates are raked so they add to the ACS direct national estimate.

Section 2 discusses the data sources used for the SAIPE estimates. Section 3 discusses the poverty models for counties and states. Section 4 discusses the estimation for school districts. Section 5 discusses research plans.

2. Poverty Measure and Data Sources

This section discusses the poverty measure and the major data sources used for the estimation of the state and county level estimates and for the school districts.

2.1. Poverty Measure

The Census Bureau uses a set of money income thresholds that vary by family size and composition to determine who is in poverty. If a family's total income is less than the family's threshold, that family and every individual in it is considered in poverty. The official poverty thresholds do not vary geographically but they are updated for inflation using the Consumer Price Index (CPI-U). The official poverty definition uses money income before taxes and does not include capital gains or noncash benefits (such as public housing, Medicaid, and food stamps).

Poverty status cannot be determined for all individuals. Unrelated children under age 15 (such as foster children) cannot be assigned a poverty status because income questions are asked of people age 15 and older, and since they are not living with a family member, there is no income to use to determine their poverty status. These individuals are excluded from the "poverty universe." For SAIPE, the estimates of children in poverty are only those for "related" children. Other individuals also excluded from the poverty universe are people in institutional group quarters (such as prisons or nursing homes), college dormitories, military barracks, and living situations without conventional housing. More information on the poverty measure can be found at: http://www.census.gov/hhes/www/poverty/about/overview/measure.html.

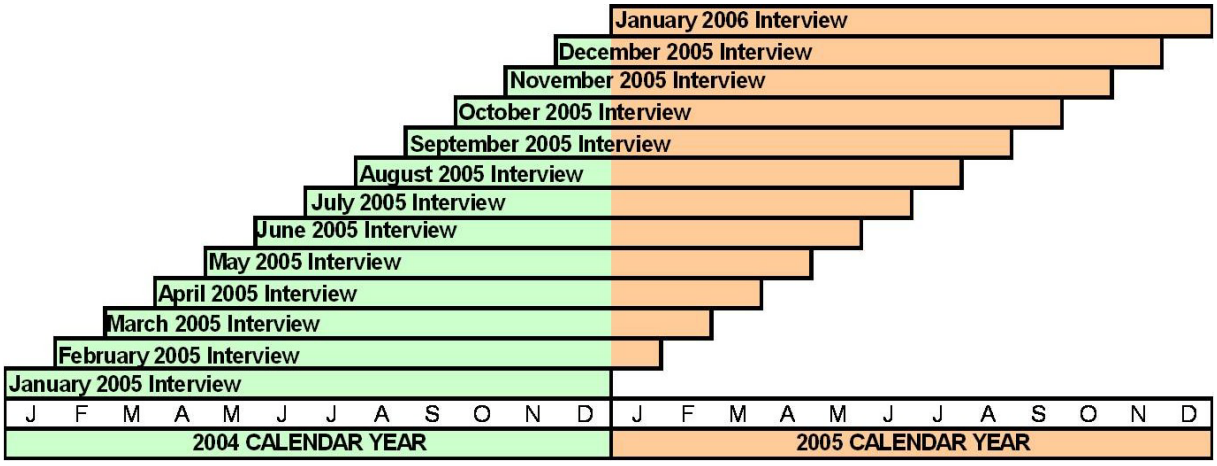
2.2. American Community Survey

The American Community Survey (ACS) is a nationwide survey designed to provide communities with reliable demographic, social, economic, and housing data estimates every year. Prior to the advent of the ACS, the sole source of these data for communities was from the decennial census long form, which could not provide timely community level data for the years between the censuses. Some surveys conducted by the Census Bureau such as the Current Population Survey are able to produce estimates for states and major metropolitan areas but no other survey can provide county and community level estimates. With the 2010 decennial census, only the basic demographic information and household composition is being collected. Information on income is not being collected so estimates of poverty will not be available from this census. The United States now relies solely on the ACS for these community level data and the primary source for state and county data on poverty and other characteristics.

The ACS has a sample size of roughly 3 million addresses, and the sample is selected from all counties and county-equivalents in the United States, and from all municipios in Puerto Rico. Single-year survey estimates are published for counties and other places with population size 65,000 or larger and three-year estimates are published for counties and other places with population size 20,000 or larger. ACS has published estimates for two different three-year spans, 2005-2007 and 2006-2008, and will publish for 2007-2009 this year. The three-year estimates use data collected in three different annual surveys so that data collected in a single year will be used in three different three-year estimates. For example, data from the 2007 ACS is used in all of the above three spans. Starting this year, the ACS will provide five-year direct survey estimates for the period from 2005 through 2009 for all counties and school districts, as well as for other small geographic areas. The ACS data used to produce SAIPE income and poverty estimates are state-level and county-level, single-year direct estimates, regardless of population size.

The ACS collects data on a monthly basis from one twelfth of the annual sample. The data for a single year are collected in January through December of that year using three models of data collection: mail. telephone, and personal interviewing. At each interview, income responses are based on the previous twelve months, not for the previous calendar year as is done for the Annual Social and Economic Supplement of the Current Population Survey. This results in a rolling reference period of 23 months for annual estimates of income where, for example, income reported in 2005 is for income earned between January 2004 and November 2005, and centered in December 15, 2004. The following illustration is from Webster (2007).

Graph 1 Illustration of overlapping reference period of income by month of ACS interview: 2005



2.3. Decennial Census Income and Poverty Data

The 2000 decennial census sample data provide income, poverty, and population estimates. The income and poverty estimates are used as predictors in the income and poverty models, and are used in

apportioning of the SAIPE county estimates to school districts. Estimates from the decennial censuses also have been used to investigate the relationships between administrative data and income and poverty, and evaluating estimates from alternative models. By constructing models that predict income and poverty of the census year, we can assess the models' performance. We also use the census to estimate approximate variances for the school district estimates.

2.4. Federal Individual Income Tax Returns

Access to federal individual income tax return data is vital to the U.S. Census Bureau's SAIPE income and poverty estimates. Each year, the Census Bureau receives selected items of information from all tax returns. The SAIPE program uses the street address on each tax return to assign a block and, subsequently, process the resulting "geocoded" data file to provide state level, county level, and school district level variables for use as predictors in the models. For returns that cannot be geocoded to a specific census block, no school district identification occurs but state and county identification is imputed using an approximate crosswalk between zip codes and state-county location. These data are stripped of any personally identifiable information and are kept under the strictest control and in complete confidentiality. Disclosure of IRS data may lead to cumulative penalties of up to $600,000 and 10 years in jail.

The total number of exemptions attributed to a return includes 1) the filer, 2) the spouse of the filer, 3) the number of child exemptions for the household, and 4) the number of exemptions for people 65 and older. Even though the adjusted gross income on a return is not the same as family income, if it is below the official poverty threshold[3] for a family of the size implied by the number of exemptions, all the exemptions on the return are treated as poor. We do not use the related child delineation in the official poverty matrix. Instead, we compare AGI to the published weighted average threshold appropriate to the size of the family unit. Where we require age groups, we assume that child exemptions correspond to the population under 18 years even though it can include persons over 18, and we obtain exemptions for the 65 and over age group directly. We do not use returns that claim no exemptions as they are filed by persons reported as dependents on some other return.

SAIPE uses different variables as predictors for the different models. For median household income, SAIPE uses median AGI and the ratio of exemptions to population. For the number of poor in an age group, the county-level model uses the number of poor exemptions (people), the number of child exemptions, and total number of exemptions. The state-level poverty models use somewhat similar variables except they are in terms of the ratio of number of poor exemptions to the population size, or the complement of these ratios.

The IRS data is also used in conjunction with the decennial census data to apportion SAIPE's county estimates of related children ages 5-17 in poverty to school districts. Using the zip-code cross-reference and block-level geocoding, about 99% of tax returns are identified to the county level. However, to determine the school district that covers the address of the return, the block where the address resides needs to be determined. Tax returns coded to blocks can be geocoded to school districts using the boundaries established in the School District Review Program. Those that cannot be geocoded are referred to as "non-geocoded." Of the 3,140 counties with IRS data, about 15% of them have block geocoding rates less than 50% and about 23% have block geocoding rates greater than 90%. Table 1 shows geocoding of tax data by county size and bloc geocoding rate. Non-geocoded returns are concentrated in rural and fast-growing areas. Most large counties have geocoding rates over 90%.

[3] Information about how the Census Bureau measures poverty can be found at http://www.census.gov/hhes/www/poverty/methods/index.html.

Table 1 Sub-county geocoding of tax data for tax year 2007

County Size by all-age Pop	% of counties with > 90% block ID	% of counties with 50-90% block ID	% of counties with < 50% block ID
Under 10k	3.5	57.7	38.7
10-20k	7.4	72.5	20.1
20-65k	17.7	75.0	7.3
65-250k	48.8	50.1	1.1
Over 250k	86.0	14.0	0.0
All sizes	23.3	61.5	15.2

2.5. Supplemental Nutrition Assistance Program Benefits Recipients

The Supplemental Nutrition Assistance Program (SNAP), formerly the Food Stamp Program, is the one low-income assistance program that is uniform in its eligibility requirements and benefit levels across states (except for Alaska and Hawaii that have more lenient eligibility requirements). While the definitions of income, household composition, and the resource income cutoffs are different from those used in the official measure of poverty, a household's eligibility for the program is determined by a standard that is tied to the poverty level. Participation in the SNAP is variable across states ranging from 49% to 95% with an average of 65% for 2005.

2.5.1. SNAP Benefits Recipients for Counties

We obtain counts of the number of people participating in SNAP from the United States Department of Agriculture, Food and Nutrition Service (USDA/FNS). In most states, we use counts of participants for the month of July in the estimation process. In a few cases, however, the states are able to provide data only for other reference periods. We control the county-level SNAP benefits variable values to the state-level SNAP benefits variable.

2.5.2. SNAP Benefits Recipients for States

We calculate the number of recipients by state as a 12-month average. The average has a delay of six months relative to the model's income year. For example, we used the average between July 2007 and June 2008 for the 2008 model.

A 12-month average produces a better measure of the typical number of poor that receive SNAP benefits than a one-month figure. A 12-month average removes seasonality, smoothes out irregularities probably not related to income, and is more comparable with poverty rates based on annual income figures.

We adjust isolated extreme values so that they are compatible with long-term trends. Outliers often result from special SNAP benefits issuance in response to natural disasters. We remove the additional recipients resulting from more lenient eligibility requirements in Alaska and Hawaii based on estimates of the number of recipients in these states who would not be eligible in the continental United States from the "Characteristics of Food Stamp Households" survey of the USDA/FNS.

2.6. Intercensal Estimates of the Population

The Population Estimates Program (PEP) publishes total resident population estimates and demographic components of change (births, deaths, and migration) each year. The Census Bureau also publishes the estimates by demographic characteristics (age, sex, race, and Hispanic origin) for the nation, states, and counties. The reference date for estimates is July 1. To estimate the household population, PEP uses a technique known as the cohort-component method. A cohort refers to a group of individuals born in the

same time period. The cohort-component method applies the components of population change to groups of individuals based on when they were born.

SAIPE requires estimates of the total resident population and the following age groups: under age 5, ages 5 to 17 years, under age 65, and age 65 and over.

SAIPE uses the intercensal population estimates along with tax return data as predictor variables in the models. The population estimates cover all residents, while the tax data cover people with filing requirements. The tax data omit two groups, the non-compliant and those without filing requirements. The non-compliant are not easily described by age or income class, but those without filing requirements are. People with low income and the elderly are less likely to have income that exceeds tax-filing thresholds. In conception, the difference between a population estimate and the corresponding number of exemptions is an indicator of the tax non-filers.

In the state-level models, the dependent variable, the variable predicted for each state, is the ratio of numbers of people in poverty to population as measured in the American Community Survey (ACS). To transform these ratios into estimated numbers of people in poverty, we multiply each estimated ratio by a demographic estimate of the population as covered by the ACS.

Finally, we use estimates of the poverty universe at both the state and county levels to compute the percentages of people in poverty. We form poverty universe estimates from the household population estimates by adjusting them to exclude other population subgroups (e.g., foster children under age 15) and to limit the estimates of the number of children to related children.

2.7. Supplemental Security Income (SSI) Recipients

The federal SSI program provides monthly cash assistance to low-income elderly persons and low-income disabled persons. The Social Security Administration tabulates various measures of SSI participation for states. After the number of recipients is tabulated, we use the estimates as predictor variables in a model to estimate the number of people 65 and over in poverty in states. These estimates are not released separately but are included in the estimates of total persons in poverty.

2.8. Treatment of Unusable Data

Beyond potential measurement error in the auxiliary data, there are cases of unusable auxiliary data that must be imputed/allocated. These cases occur either due to counties lacking data for that variable or for counties that have extreme, implausible data for that variable. The outliers are detected primarily by comparison with various other data available for that county. If the variable is extreme for a given county relative to the other data then the variable for that county is flagged as an outlier. Another way of detection of extreme data is to look at changes across many years. Although finding wild change over time is not necessarily indicative of a problem in the variable (it may occur for a real economic reason on the ground), it suggests need for scrutiny in order to learn why the variable is changing so much.

For the SNAP in particular, four simple regression models are run using random state effects in which the logarithm of the variable under study is the dependent variable and logarithms of some of the remaining auxiliary variables are predictors. We flag counties for possible replacement when the average of the four standardized residuals exceeds 2.5 in absolute value. When replacement is recommended, we replace the extreme value with the exponentiated prediction (of the four) that is closest to the raw extreme value.

3. SAIPE Poverty Models

3.1. Introduction

This section discusses small area models applied to single-year survey estimates from the ACS for county and state level poverty estimates of school age (5 - 17) children. SAIPE uses a basic area level model

used for small area estimation (Rao 2003) that reflects the general form suggested by Fay and Herriot (1979). The SAIPE county model has the data in log-level form. This model is referred to as the log-level model since it involves modeling estimates of the log of the number in poverty. This requires us to drop counties with estimates of zero from the ACS since the log of zero is undefined. This amounts to about 130 counties out of roughly 3,140 counties in the United States. At the end of the modeling, we translate back from the log scale to the level scale by exponentiating the shrinkage estimate (3) and applying a log-bias adjustment. We then control the county totals up to the state level. The SAIPE state model uses the data in linear rate form. We use specially adjusted demographic population estimates (from the Population Estimates Program) where needed. These specially adjusted population estimates conform to the survey concept of the population used by the ACS. We also use these specially adjusted population estimates to provide suitable denominators to compute rates for the estimates (esp. for poverty).

The log-level model used for counties is presented in section 3.2. Section 3.3 discusses the definitions of the variables used in the county model. Section 3.4 compares the SAIPE estimates with ACS estimates. Section 3.5 discusses the state poverty model, which is a model of poverty ratios and section 3.6 discusses adjusting the estimates derived from each of the models to ensure consistency with the national poverty estimates from the CPS ASEC and consistency of the county estimates with the state estimates.

3.2. County Log-level Model

The SAIPE county model has the data in a log-level form. The dependent variable in the SAIPE county model is the logarithm of the direct survey estimate of the number of related children (5 -17) in poverty. The log-level model is

$$\log\left(y_i\right) = \log\left(Y_i\right) + e_i \quad \text{where} \quad e_i \sim ind.\ N\left(0, \mathrm{v}_i\right) \tag{1}$$

$$\log\left(Y_i\right) = x_i'\beta + u_i \quad \text{where} \quad u_i \sim i.i.d.\ N\left(0, \sigma_u^2\right) \tag{2}$$

where, for county i :

y_i = ACS survey estimate of the number 5 -17 in poverty,
Y_i = true population value of 5 -17 poverty that y_i is intended to estimate,
$e_i = \log(y_i) - \log(Y_i)$ = sampling error in $\log(y_i)$ as an estimate of $\log(Y_i)$,
x_i = vector of regression variables,
β = vector of regression parameters, and
u_i = area-specific random effect (county random effect).

We assume that the random errors, e_i and u_i, both have independent normal distributions, and the sampling errors, e_i, and area-specific random effects, u_i, are mutually independent from each other. The sampling error variances, v_i, of the $\log(y_i)$, are treated as known. In reality, they are estimated by replication methods and subject to estimation error. These estimation errors "can affect the accuracy of the small area [estimates] and lead to bias in stated mean square error" (Bell 2008). Generalized variance functions (GVFs) for smoothing these estimates are being investigated.

The model defined by equations (1) and (2) is estimated by maximum likelihood, the unknown parameters being β and σ_u^2. The shrinkage estimates (empirical best predictions) in the log scale, i.e., predictions of $\log(Y_i)$, and the corresponding prediction error variances are computed by plugging the parameter estimates into the following standard formulas (Bell 1999):

$$\widehat{\log\left(Y_i\right)} = \left(1 - w_i\right)\log\left(y_i\right) + w_i\left(x_i'\hat{\beta}\right) \tag{3}$$

$$\text{where } w_i = v_i \Big/ \left(\hat{\sigma}_u^2 + v_i\right)$$

$$\mathrm{Var}\left[\log\left(y_i\right) - \widehat{\log\left(Y_i\right)}\right] = w_i\hat{\sigma}_u^2 + w_i^2\left(x_i'\,\mathrm{Var}\left(\hat{\beta}\right)x_i\right) \tag{4}$$

The estimator in (3) is called a shrinkage estimate in that it shrinks the effect of the survey estimates on the SAIPE small area estimate. When the sampling variance, v_i, is small the shrinkage weight, w_i, is nearly zero so that there is not much shrinkage and the SAIPE estimate is little different from the survey estimate. When the sample variance is large, the shrinkage weight is near one and the SAIPE estimate is shrunk nearly to the prediction estimate, $x_i'\hat{\beta}$.

Predictions of Y_i on the original scale (not logged) and associated variances are defined from the above results using an inverse transformation appropriate for the assumed normal distribution on the random errors.

$$\hat{Y}_i = \exp\left(\widehat{\log(Y_i)}\right)\exp\left[\left(\hat{\sigma}_u^2 - Var\left(\log(Y_i) - \widehat{\log(Y_i)}\right)\right)\Big/2\right] \quad (5)$$

$$\mathrm{Var}\left(\hat{Y}_i\right) = \hat{Y}_i^2\left[\exp\left(\mathrm{Var}\left[\log(Y_i) - \widehat{\log(Y_i)}\right]\right) - 1\right] \quad (6)$$

3.3. Variable Definitions

Table 2 lists the variable definitions for the SAIPE county estimates for 2006. The dependent variable is the log of the direct 2006 ACS county estimates of the number of related children ages 5 - 17 in poverty. **T**he 2007 ACS represents a population distribution (controls) from the U.S. Census Bureau's Population Estimates Program (PEP) for July 1, 2007, while the poverty estimates use income reports covering 12-month spans that start as early as January 2006 and end as late as November 2007. The regression variables are those from the model used to produce the official 2006 SAIPE county poverty estimates. The tax data are for income year 2006, the SNAP participants are for July 2006, and the Census 2000 estimates refer to income year 1999. Population estimates for July 1, 2007 are used to correspond with the ACS survey controls and to correspond approximately with the population distribution of personal income tax filings, which are mostly made between January and May of the year following the income year.

Table 2 Variable definitions

Short Name	Description
Dependent Variable	
Log (ACS poor, ages 5-17 related)	Log estimated county number of 5 - 17 related children in poverty from the 2007 ACS.
Regressors	
Intercept	
Log (IRS child tax-poor exemptions)	Log number of county tax-poor child exemptions from IRS administrative records, where tax-poor is defined as Adjusted Gross Income (AGI) below the poverty level for a household size defined by the total number of exemptions on the return.
Log (SNAP participants)	Log number of county Supplemental Nutrition Assistance Program participants reported in July (data from the USDA Food and Nutrition Service), raked to a control total obtained from state SNAP participant data.
Log (PEP population, ages 0-17)	Log county population, ages 0 - 17, as of July 1, 2007, from the Census Bureau's Population Estimates Program (PEP) of intercensal demographic estimates.
Log (IRS child tax exemptions)	Log total number of county child exemptions from IRS administrative records.
Log (Census 2000 poor, ages 5-17 related)	Log estimated county number of 5 - 17 related children in poverty from Census 2000.

Bell, Basel, et al. (2007, pages 24-25) report for the model based on 2005 data that

> The coefficient estimates on all variables bear the expected sign, with the coefficient estimates positive except for the coefficient on the IRS child tax exemptions variable, which is negative as expected. The results can be summarized as follows: (*i*) the coefficient estimates are each individually highly significant, (*ii*) the log (IRS child tax-poor exemptions) variable is the most important (highest t-statistic), (*iii*) the R-squared is large [(0.936)] indicating that a lot of the variation in the dependent variable (mostly due to county population size) is being explained, and (*iv*) the sum of the coefficients on the log (PEP population, ages 0-17) and log (IRS child tax exemptions) variables [1.050+(-1.037)] is not statistically different from zero. This reinforces the interpretation of the net effect of these variables as relating to the effect of a log (tax nonfiler rate) variable.

They show a table with the regression coefficients, their standard errors, and significance level and they show a table with the correlation matrix of these coefficients estimates. They report that

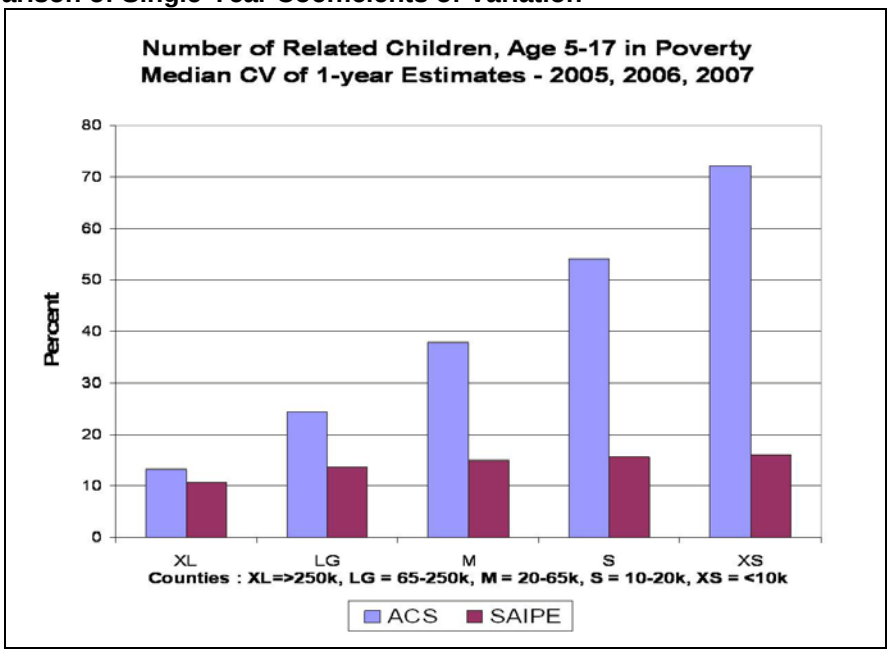
> The correlation for the coefficients on the IRS child tax exemptions and PEP population, ages 0-17 variables is nearly negative one. Since the estimate values are equal and have opposite signs as well, this is further evidence that the two terms could be combined without appreciably changing the model predictions or standard errors.

For the dependent variable and the shrinkage estimates, SAIPE uses 2007 ACS estimates for every county even though ACS releases estimates only for counties with a population greater than or equal to 65,000. For some counties with small samples, the estimate of the number of related children ages 5 – 17 in poverty is zero by random chance. Since logs cannot be taken of these estimates, such counties are excluded from the regression prediction.

3.4. Improvement in Accuracy of SAIPE Compared to ACS

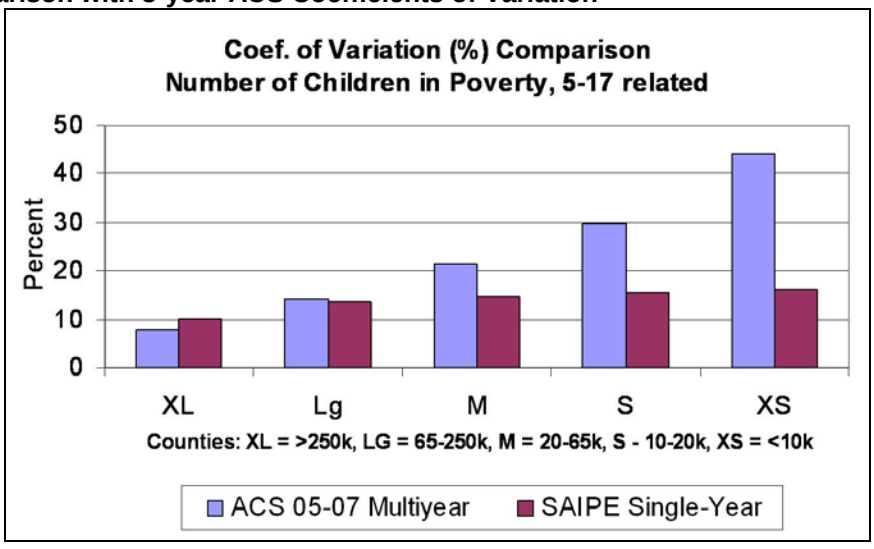
The SAIPE county estimates can dramatically improve the accuracy of related children ages 5-17 in poverty compared to the single-year estimates from the ACS. In addition, the SAIPE are also more accurate for small counties compared to three-year ACS estimates. Graph 2 shows a comparison of the coefficients of variation (CVs) between of the 2006 SAIPE estimates and the average of the single-year 2005 through 2007 ACS estimates. A coefficient of variation is the standard error of the estimate divided by the estimate. The graph groups counties by size: extra large (XL) have population greater than or equal to 250,000, large (LG) have population between 65,000 and 250,000, medium (M) have population between 20,000 and 65,000, small (S) have population between 10,000 and 20,000, and extra small (XS) have population less than 10,000. The height of the bars is the median CV for the group and the left bar in each pair is for the ACS and the right bar for SAIPE. For the largest counties, the CVs are about the same and as the sizes of the counties decrease, the CVs for the ACS grow to a median over 70 percent for the extra small counties while the CVs for the SAIPE estimates increase moderately to only about a 16 percent CV.

Graph 2 Comparison of Single-Year Coefficients of Variation



Graph 3 shows a comparison of the coefficients of variation of SAIPE estimates based on a single-year of ACS with the CVs for the three-year 05-07 ACS estimates. We see a less dramatic difference between the SAIPE and ACS estimates. The ACS estimates are somewhat better for the largest counties and about the same for the large counties but the SAIPE shows considerable increase in accuracy for the small and smallest counties. Note also that the multiyear ACS estimates are not measuring a single-year poverty concept but rather a three-year average.

Graph 3 Comparison with 3-year ACS Coefficients of Variation



Each of these graphs shows a strong positive correlation between the ACS estimate of log number of children in poverty and each of the independent variables. Including IRS child tax exemptions and PEP population, ages 0 – 17 in the model has in the past led to coefficient estimates with magnitudes that were not statistically significantly different, but with a negative coefficient for the log (IRS child tax exemptions) variable and a positive coefficient for the log (PEP population, ages 0-17) variable. This has been

interpreted as providing a proxy for a log tax "filers" rate [= log (IRS child tax exemptions) - log (PEP population, ages 0-17)], for which a negative relation to poverty is expected.

3.5. State Poverty Models

The SAIPE state poverty ratio models are in the form of equations (1) and (2) except the models are applied to untransformed poverty ratios instead of logarithms of the poverty estimates. The regression variables in x_i, in addition to the intercept term, are ratios related to poverty. The regression variables in the state 5 - 17 poverty ratio model include the following:

- IRS tax-poor child exemption rate = (number of child exemptions on tax returns with incomes below the poverty threshold in the state) / (total child exemptions in the state),
- IRS nonfiler rate = 1 – (number of exemptions in the state) / (state population),
- state rate of participation in the SNAP = (number of food stamp participants) / (state population), and
- "Census residuals" obtained by regression of the previous census state 5 - 17 poverty ratio estimates on the above regression variables defined in the census income year (1999 for Census 2000).

The model is given a Bayesian treatment with flat priors on σ_u^2 and on the regression parameters β (Bell 1999). The results obtained are posterior means and variances of the true poverty rates Y_i, which are obtained from the Bayesian analogues of equations (3) and (4).

The results from modeling the 2005 ACS data show little effect on the point estimates or their model error variances for most states – that is, for most states the shrinkage estimates and their variances from equations (3) and (4) are about the same as the direct ACS results. This is due to the small sampling error variances of the ACS survey estimates for most states. However, for the ten or so states with the largest sampling error variances, the modeling and shrinkage estimation can have an appreciable effect on the results.

3.6. Raking of Estimates

To date, SAIPE state and county model-based estimates of number of children in poverty have been raked (proportionally adjusted) to maintain consistency from the county to the state to the national level. That is, prior to the 2005 SAIPE estimates, state estimates of number of children in poverty from the state CPS ASEC model have been raked to direct CPS ASEC national estimates of number of children in poverty, and estimates from the county CPS ASEC model have been raked to these raked state model estimates. The same raking procedure is currently used, with the difference that all models are based on ACS as the direct survey source. So the SAIPE estimates for any level, 2005-2008, will sum to the single-year national ACS number in poverty.

4. SAIPE School District Estimates

Three estimates are provided for each school district:

- total population
- children ages 5-17
- related children ages 5-17 in families in poverty

The number of related school-age children in families in poverty in each school district is provided as a component of the determination of Title I grants. The estimate of the total population of each district is provided for use in the small district (fewer than 20,000 people) provision. The number of school-age children is provided so that the proportion of children in poverty can be determined. This proportion is required for determining eligibility for grants.

School districts are not always nested within counties. In some states, districts and counties are geographically identical. In other states, a county will contain many school districts and school districts may cross county boundaries. School districts often are “unified” in the sense that they are responsible for educating all elementary and secondary grades. Some states have areas with separate "elementary" and "secondary" school districts, each exclusively responsible for providing education in some grades in their shared territory. There are also some states that have school districts with different school-age ranges in different parts of their territory. 83.1 percent of school districts are unified, 13.1 percent are elementary, and 3.4 percent are secondary. School districts are very variable in size. The Table 3 shows the distribution of school districts by child population. Most school districts are small with 95 percent having fewer than 10,000 children and 42 percent having fewer than 750 children.

Table 3 Distribution of 13,754 school districts by child population

Child Population	Percentage
<100	8.0%
100-250	10.8%
250-750	23.2%
750-1,500	18.1%
1,500-3,500	20.1%
3,500-10,000	13.9%
10,000+	5.5%

This school district geography poses a problem for estimation for school districts. SAIPE must estimate concepts for a domain defined along two dimensions—geography and age. It is reasonably straightforward to define such a domain for survey data such as the ACS, where we already have detailed geographic identification and single-year of age for each response. For tax data, however, no age identification is currently available. So in the cases where a school district does not span the complete school-age range, a synthetic adjustment is applied based on the proportions of the specific ages obtained from post-censal population estimates.

To produce poverty estimates at this level, SAIPE uses a synthetic approach that combines within county population and poverty shares from two sources: the latest decennial census survey and the aggregated tax data discussed earlier. Prior to 2006, the school district estimation used the Census share method. A share is the proportion in a county of the number of related children ages 5-17 in poverty that are in a school district piece. (A school district is contained within a state but may be parts of more than one county. A school district piece is the part of a school district contained in a single county.) The Census share method assumes that this share does not change between censuses so that the distribution of poverty does not change in the intercensal years. The estimate for a school district piece is the product of the share from the latest census and the SAIPE county estimate of related children ages 5-17 in poverty. The estimate for the school district is the sum over the school district pieces.

In 2006, the “minimum change” method using both IRS income tax data and census data (Maples and Bell, 2007) replaced the Census share method. This method relaxes the assumption that the within distribution of related children ages 5-17 in poverty does not change between censuses by incorporating IRS data on children poor exemptions. We adopted this new method because income data from tax returns is informative about poverty status, it contains geographic detail to link income returns, thus exemptions, to individual school districts, it covers all school districts, and it is useful in predicting poverty for states and counties. It has also been shown to be informative about child poverty from the Census 200 and by extension would be informative about the distribution of child poverty with counties for non-census years (Maples and Bell 2004). There are some limitations to the IRS tax data in using it for estimation for school district pieces: age information is not recorded for child exemptions, school districts may overlap

geographically by covering different grade ranges, and not all tax returns can be geocoded to the school district pieces.

Section 4.1 discusses the key steps prior to making the school district estimates: the boundary update and geocoding federal tax returns, and assigning children to school district piece. Section 4.2 discusses the minimum change method for the school district estimation.

4.1. Key Steps in School District Estimation

4.1.1. Spatial Boundaries and Geocoding Tax Returns

SAIPE provides poverty estimates for all school districts with boundaries updated by the School District Review Program. It consists of two phases: the Annotation Phase and the Verification Phase. In the Annotation Phase, the Census Bureau provides state officials with materials containing the most current school district boundaries and information for their state. These data are reviewed and any changes in the school district boundaries or attributes are reported to the Census Bureau. In the Verification Phase, when state officials review the results of any changes submitted, after the Census Bureau has incorporated those changes into the Master Address Files (MAF)/Topologically Integrated Geographic Encoding and Referencing (TIGER) System (MAF/TIGER) database.

Initially, tax returns are geocoded to blocks based on addresses provided on the forms. Tax returns coded to blocks can be geocoded to school districts using the boundaries established in the School District Review Program. Not all tax returns can be geocoded to blocks. These tax returns typically have rural addresses or are from areas with new streets in fast growing areas. These “non-geocoded” blocks can only be tabulated at the county level. To make estimates for the school districts, the non-geocoded exemptions from these tax returns must be apportioned among the school districts.

4.1.2. Allocation to School Districts

For each school district, SAIPE estimates pertain to all resident school age children ages 5-17, inclusive, whether enrolled in public or private school, or not enrolled. Where two districts divide the children of an area between them by grade, the estimates do so as well. Grade ranges for each district are collected during the boundary update and supplemented with phone calls to districts. Census assigns a single grade range to each district that, in the case of spatially overlapping districts, leaves no grade unclaimed and no grade claimed by more than one district.

To tabulate the data for each district, each child is assigned a grade. In the Census 2000 sample, where responses to the "long-form" questions are available, 97 percent of children are assigned a grade based on their edited reports of the grade in which they were enrolled. For Census 2000 short-form data, where school enrollment and educational attainment are not available, children were assigned the modal grade for their age on October 1, 1999. With the Census 2000 record for each child assigned to a single school district to which that child is said to be "relevant," we tabulate for each district: total population, children ages 5-17, and related children ages 5-17 in families in poverty.

The data on children tax exemptions for the IRS only provides the number of exemptions and not ages of the children. The data on children tax exemptions for the IRS only provides the number of exemptions and not ages of the children. The proportions of the relevant ages for each district, calculated from the county-level post-censal population estimates by single-year of age, is multiplied to total geocoded child exemptions and poor child exemptions at the school district piece-level to obtain an estimate of relevant child exemptions and poor child exemptions.

4.2. School District Estimation

The first part in making school district poverty estimates is to compute the school district piece (SDP) tax-based child poverty rate using federal tax information obtained from the IRS. To do this we would need to adjust the county related children ages 5-17 poverty rate by estimates from the tax data of the ratio of the SDP child poverty rate over the county child poverty rate. The formulae for these two rates are:

$$\text{SDP TaxChildPovRate} = \frac{\text{SDP TaxChildPoor}}{\text{SDP TaxChild}}$$

$$\text{County TaxChildPovRate} = \frac{\text{County TaxChildPoor}}{\text{County TaxChild}}$$

The ratio of these can be rewritten as the ratio of the school district piece to county share of "tax-poor exemptions" over the share of total "child tax exemptions." Thus the formula for the tax-based poverty rate for a school district piece is:

$$\frac{\text{SDP TaxChildShare}}{\text{SDP TaxPoorShare}} \times \left(\text{SAIPE County ChildPovRate}\right)$$

As discussed above, not all tax returns can be geocoded down to a specific school district piece. However, the total number of exemptions in a county is known. The tabulated child tax exemptions and child tax-poor exemption counts are adjusted to reflect the appropriate grade range of the school district piece as discussed above. The next step in calculating the tax-based shares is to estimate the school district piece to county share of relevant children age 5-17 and relevant children age 5-17 in poverty from Census 2000. The non-geocoded exemptions are also adjusted to reflect the target 5-17 year old population and then allocated to the school district pieces to minimize the difference between the tax-based shares and the corresponding census-based shares using the Minimum Change method (Maples and Bell, 2007). After allocating the non-geocoded exemptions, the tax-based poverty rate for a school district piece can be computed.

Table 4 illustrates three examples of the Minimum Change method. There are three school district pieces in a county. The shares from the census are in the second column. In each example, the first column is the tax shares – the proportion geocoded to each school district piece and the proportion not geocoded. The second column is the resulting share allocation using the Minimum Change method. In the first example, all tax shares geocoded to school district pieces are less than the census shares. The Minimum Changes shares are then the census shares. In examples 2 and 3, the tax share for the first school district piece is greater than the census share. For this school district piece, the Minimum Change share is the tax share and the non-geocoded share must be allocated to the other two school district pieces minimizing the change from the census shares. The method attempts to allocate the remaining tax shares (32 = 100-68) proportional to the remaining census shares (30 and 10). The trial shares are then 24 and 8. In example 2, none of the tax shares is greater than the trial shares so the trial shares are the Minimum Change shares. In example 3, the tax share (9) for school district piece 3 is greater than the trial share (8) for it so its Minimum Change share is the tax share. The remaining share (23) is assigned to the second school district piece.

Table 4 sub-county geocoding of tax data

		Example 1		Example 2		Example 3	
SD Piece	Census Share	Tax Shares	MC Shares	Tax Shares	MC Shares	Tax Shares	MC Shares
1	60	47	60	68	68	68	68
2	30	23	30	14	24	15	23
3	10	1	10	7	8	9	9
Non-geocoded		29		11		8	

The second part in creating the school district poverty estimates is to multiply the school district piece poverty rate by the official estimate of relevant child population for the school district piece. These estimates are then raked (ratio adjusted) to agree with the county estimates for number of children age 5-17 in poverty. Finally, the raked school district piece estimates are adjusted using "controlled rounding."

The final step is to reassemble the school district pieces into the school districts, simply by adding their controlled-rounded numbers of children in poverty together.

5. Current and Future Research

SAIPE has an extensive research effort aimed primarily to improve the precision and timeliness of poverty estimates for school-age children at the school district level, and secondarily, to expand the methods to domains and geographies not currently covered. In addition, there has been a parallel effort to develop and publish health insurance coverage estimates for states for domains defined by age, sex, race, Hispanic origin, and several income groups defined by multiples of the poverty thresholds, and for counties for domains defined by age, sex, and income groups similarly defined. We have made three releases using CPS ASEC data and may release estimates for 2008 in late 2010 using data from the ACS for the first time.

The major recent innovations for SAIPE have been the replacement of the CPS ASEC with ACS for the survey data, the use of tax data for school district estimates using the minimum change method, and an accelerated production schedule. Other recent advances have been in estimating margins of errors for school district poverty estimates and providing margins of errors for intertemporal and cross-sectional comparisons. The latter work has also enabled the construction and publication of other aggregate domains, such as multi-year averages, metro and micropolitan areas, and regional transportation blocs, each provided with associated margins of error.

Major ongoing and future projects involve improvements to the county model, improved estimation for sub-county areas, and ways to use multiple or multi-year ACS estimates.

- The current county model uses the survey estimates of the sampling variance that are subject to estimation error and can lead to bias in stated mean square error of the SAIPE estimates (Bell 2008), and it excludes counties with zero poverty estimates because of the logarithmic trans-formation. We are researching methods for smoothing these sampling variances and alternative models that would allow the inclusion of the zero poverty estimates.
- To improve the sub-county estimates we need to improve the identification of the sub-county geo-coding of tax returns. To do this, we need to improve sub-county identification of addresses in our master address file and improve geographic location of tax returns for households served by rural and PO boxes in the tax data.
- We will also research direct modeling of sub-counties making use of multiple or multi-year estimates from the ACS. The ongoing improvements to the county model involving modeled sampling variances and inclusion of zero-poverty areas, which will become much more frequent, will be crucial for sub-county estimation.
- The 2000 decennial long form data is becoming increasingly dated and new data on poverty from the 2010 will not be available. The most precise survey data will be from multi-year ACS but with a time lag because of the averaging over three or five years. Research will look at the best way to incorporate these multi-year ACS estimates or several years of single ACS estimates into the SAIPE estimation.
- To expand SAIPE to include more age groups than currently possible would likely require age-identification of tax data, which would generally require additional IRS authorization than is currently approved.
- The larger ACS sample size would allow the estimation of more domains defined by multiples of the poverty threshold. These domains would be estimated consistently and would be useful to inform policy and other public programs.

References

1. Bell, W.: Accounting for Uncertainty about Variances in Small Area Estimation, 1999, U.S. Census Bureau. http://www.census.gov/did/www/saipe/publications/files/Bell99.pdf

2. Bell, W.: Examining Sensitivity of Small Area Inferences to Uncertainty About Sampling Error Variances, 2008, In *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* 327-334*.*
http://www.census.gov/did/www/saipe/publications/files/Bell2008asa.pdf

3. Bell, W, Basel, W., Craig, C., Dalzell, L., Maples, J., O'Hara, B., and Powers, D.: Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties, 2007, U.S. Census Bureau. http://www.census.gov/did/www/saipe/publications/files/report.pdf

4. Fay, R. and Herriot, *R.: Estimates of Income for Small Places:* An Application of James-Stein Procedures to Census Data, 1979, *Journal of the American Statistical Society*, **74**, 269-277.

5. Maples, J. and Bell, W.: Investigating the Use of IRS Tax Data in the SAIPE School District Estimates, 2004, In *Proceedings of the Section of Survey Research Methods of the American Statistical Association*, 1656-1663.

6. Maples, J. and Bell, W.: Small Area Estimation of School District Child Population and Poverty: Studying Use of IRS Income Tax Data, 2007, *Research Report Series, Statistics #2001-11*, U.S. Census Bureau. http://www.census.gov/srd/papers/pdf/rrs2007-11.pdf

7. Maples, J. and Bell, W.: Small Area Estimation of School District Child Population and Poverty: Studying Use of IRS Income Tax Data, 2007, *Research Report Series, Statistics #2001-11*, U.S. Census Bureau. http://www.census.gov/srd/papers/pdf/rrs2007-11.pdf

8. Powers, D., Basel, W. and O'Hara, B.: SAIPE County Poverty Models Using Data from the American Community Survey, 2008, In *Proceedings of the Section on Government Statistics of the American Statistical Association*, 2378-2383.

9. Rao, J.N.K.: *Small Area Estimation*, 2003, Wiley.

10. *Webster* Jr., B.: Evaluation of Median Income and Earnings Estimates: A Comparison of the American Community Survey and the Current Population Survey, 2007, U.S. Census Bureau. http://www.census.gov/acs/www/Downloads/Evaluation_of_Income_Estimates31207.doc