# Use of Labor Market Indicators in Small Area Poverty Models*

Wesley Basel and Nora Albert
U.S. Census Bureau, Washington, D.C. 20233

**Abstract**
Current production models of poverty and income utilized by the Census Bureau for the Small Area Income and Poverty Estimates (SAIPE) program do not utilize labor market indicators. The mismatch between the geographic definition of a labor market versus residential-based poverty is the primary impediment to their use. This paper will use data from the Local Employment Dynamics program at the Census Bureau to construct a residence-based indicator of employment. The indicator will then be included in county-level SAIPE poverty models to test its usefulness.

**Key Words:** small area, small domain, employment, census, SAIPE, poverty

## 1. Background

The SAIPE program produces model-based estimates of poverty that combine direct estimates from the American Community Survey (ACS) with regression predictions based on administrative records, postcensal population estimates and decennial census data. For both the survey data and the explanatory data, individual units are aggregated for the specified geographic area and year, producing inputs and estimates that are interpreted as single-year or annual data. The modeling techniques allow the SAIPE program to produce single-year estimates of child poverty for all school districts and all counties, regardless of population size.

The current model used for published SAIPE estimates relies on indicators drawn from aggregated federal income tax returns, participation in the Supplemental Nutrition Assistance Program (SNAP), and population estimates to construct regression predictions. Employment indicators have been considered in the past, but have not been found useful due to the high explanatory power of the concepts included in the model, and the difference in the geographic definition of most employment indicators. The

* This report is released 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.

second issue will be solved for the analysis of this paper by re-tabulating the employment data by place of residence, rather than the usual place-of-work identification. The first issue with the use of employment data will always be a limiter for its use in published SAIPE program estimates, but many other needs for poverty models occur where the use of federal personal income tax data is not approved. So the joint effect of including both, neither, or one of the employment and/or tax data will be tested in this analysis.

The layout of the paper is as follows. First, we introduce the structure of the model in Section 2, focusing on the list of indicators included in the regression model. In Section 3, we describe the compilation of the employment data used in the analysis. In Section 4, we present empirical results for four different alternative sets of variables. We give concluding remarks in Section 5.

## 2. Structure of the Model

In general, the SAIPE program's county poverty model follows a Fay-Herriot, or shrinkage, approach, by specifying both a sampling model and a regression model for the true value of log (poverty) (Fay & Herriot, 1979; Bell, 1999). The empirical-Bayes, best linear unbiased predictor (EBLUP), under the assumptions of the model, is then a weighted average of the direct estimate from the ACS sample and the predictions from the regression model.

For this paper, we follow the same approach, but applied to a structure on the poverty rate scale, rather than the log-level scale. In the published SAIPE model, most of the concepts in the regression model are closely related to poverty, such as tax-based poverty, participation in SNAP, and so on. There is a nearly direct link between ACS poverty and these explanatory variables, so a log-level model has performed well. For the addition of employment indicators, we are using data without such a direct linkage, and the scale of the concepts could be considerably different. For this reason, we are examining a rate model in this paper.

For a single year, the specification of this model is given in (1) below.

For $i$ = 1,2, …, $m$ areas,

$$\begin{aligned} y_i &= Y_i + e_i \qquad && \text{Sampling Model} \\ Y_i &= \mathbf{x}_i'\beta + u_i \qquad && \text{Linking Model} \end{aligned} \tag{1}$$

where $y_i$ represents the direct survey estimate of the poverty rate for a specified age group from a single-year sample of the ACS, $Y_i$ is the unobservable true value of the poverty rate, and $\mathbf{x}_i$ is a k x 1 vector of explanatory variables, also in a rate form. The model errors, $u_i$, are assumed i.i.d. with variance $\sigma_u^2$, and the sampling errors, $e_i$, are assumed independent with sampling variance, $v_i$, specific to each county and year. The ACS sampling variance, $v_i$, for a given county is estimated directly from the sample using a successive difference replication method described in the ACS documentation (U.S. Census Bureau, 2009), and then are assumed known. Both the model and sampling error terms are assumed normally distributed in the rate scale, and both the regression parameters and the model error variance are estimated by maximum likelihood.

The results reported in this paper will be the regression parameters, the model error variance, and summary statistics for the standard errors of the Fay-Herriot shrinkage estimate. The shrinkage estimate is obtained when the direct and indirect estimates are combined using an efficient weighting described in Bell (1999), producing the EBLUP, or empirical Bayes best linear unbiased predictor. The equations for this shrinkage estimate and its standard error is shown in (2) and (3) below. See Bell (1999) for more details.

$$\begin{aligned} \hat{Y}_i &= (1-w_i)y_i + w_i\mathbf{x}_i'\hat{\beta} \\ w_i &= v_i / (\hat{\sigma}_u^2 + v_i) \end{aligned} \qquad (2)$$

$$\mathrm{Var}(Y_i - \hat{Y}_i) = w_i\hat{\sigma}_u^2 + w_i^2\mathbf{x}_i'\mathrm{Var}(\hat{\beta})\mathbf{x}_i \qquad (3)$$

The regressor variables used in the regression model for the four alternative specifications of this model are rate forms of the variables used in the published SAIPE model, plus the jobs ratio, whose construction is discussed below. Results for two age groups, independently estimated and analyzed, are given. The two age groups are school-age children, defined as ages 5 to 17 in families, and the all-age group. The variables for the school-age children estimates are listed in Table 1. For the all-age regression, we substitute all-age for 5-17 in families, and all exemptions for child exemptions.

For the employment ratio concept used in this analysis, the desired concept for the numerator would be the total employed persons resident within the county. The Annual Supplement for Economic Characteristics to the Current Population Survey has an estimate of this concept, but this estimate is not available for every county, and has considerable sampling error. An administrative records source, that is, a tabulation of actual workers, would better fit our needs. Most published employment data is tabulated by place of work, rather than place of residence, which for many counties within large metropolitan areas is only indirectly related to the residence. Fortunately, the Local Employment Dynamics (LED) program at the Census Bureau has such a source of employment data available, which is identified by place of residence. The LED program combines several different data sources to arrive at this concept. See http://lehd.did.census.gov/led/led/led.html for more details.

The underlying data is number of jobs, not number of employees, so there will be some double counting. However, the database is coded by both the county of residence for the job-holder, as well as the usual county of the worksite. This double tabulation makes it possible to aggregate the jobs by the residence of the job-holder. The jobs ratio then used as a regressor is the aggregate jobs by place of residence divided by the Bureau's population estimate for the county for ages 18 to 64.

**Table 1: Variable definitions for the poverty rate model, Ages 5-17 in families**

| Short Name | Description |
|---|---|
| Dependent Variable | |
| ACS poverty rate for 5-17 | Estimated county poverty rate from the 2010 ACS sample, ages 5-17 in families. (Both published and unpublished ACS county-level estimates are used.) |
| Regressors | |
| IRS child tax-poverty ratio | Number of county tax-poverty 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, divided by the total child exemptions. |
| SNAP participation ratio | Number of county SNAP participants reported in July (data from the USDA Food and Nutrition Service), raked to a control total obtained from state SNAP participant data, divided by population ages 0-64. |
| Dependency ratio | County population, ages 0 - 17, as of July 1, 2010, from the Census Bureau's Population Estimates Program (PEP) of post-censal demographic estimates, divided by the all-ages population. |
| IRS Filing coverage ratio | Total IRS child exemptions divided by the age 0 to 17 population estimate. |
| Jobs Ratio | Total jobs held by residents of the county (see below) divided by the population estimate ages 18-64. |
| Further information about these input data is available on the SAIPE program's webpage, http://www.census.gov/hhes/www/saipe/techdoc/inputs/datintro.html. | |

## 3. Employment Data

As mentioned in the previous section, LED data and population estimates data is used to create the jobs ratio. The specific source of LED data used is LED's OnTheMap program (http://lehd.did.census.gov/led/datatools/onthemap.html). OnTheMap estimates are based on data from each state's Unemployment Insurance (UI) system, which excludes certain types of employment depending on the state. Employment not typically included in UI data include: most federal employees, members of the Armed Services, some state and local government employees, most farmers and agricultural laborers, domestic workers, self-employed workers, and some employees of nonprofit or religious organizations.

OnTheMap data is available for every state except Massachusetts, New Hampshire, and the District of Columbia. To fill in the gaps, we use the Quarterly Census of Employment and Wages (QCEW) from the Bureau of Labor Statistics (http://www.bls.gov/cew/). QCEW covers most of the same industries as the OnTheMap data, but to maintain as much consistency as possible we exclude federal employment from the QCEW employment count. To obtain a proxy for OnTheMap data, we weight the employment (from QCEW) for missing counties by the estimated proportion of people working in the county who reside in each source county. These estimated proportions were obtained from the ACS five-year sample, 2006-2010.

## 4. Results

The purpose of this paper is to examine the usefulness of the jobs ratio both in the published SAIPE model, and its potential in situations where the use of aggregate IRS data is not allowed. For this reason, we consider four different alternative specifications, differing only by the list of included regressors. All specifications include the SNAP participation ratio and the dependency ratio, but two omit the regressors dependent on the tax data, and two omit the jobs ratio. Table 2 lists the regressors for the four models.

**Table 2: List of included regressors for four alternative specifications.**

| Model 1<br>No Tax/No Jobs | Model 2<br>No Tax/With Jobs | Model 3<br>With Tax/No Jobs | Model 4<br>With Tax/With Jobs |
|---|---|---|---|
| SNAP Participation Ratio | SNAP Participation Ratio | SNAP Participation Ratio | SNAP Participation Ratio |
| Dependency Ratio | Dependency Ratio | Dependency Ratio | Dependency Ratio |
| | | IRS Tax-Poor Ratio | IRS Tax-Poor Ratio |
| | | IRS Filing Coverage Ratio | IRS Filing Coverage Ratio |
| | Jobs Ratio | | Jobs Ratio |

Note that the two specifications including the jobs ratio differ from the specification to its left by only that variable. Thus, a simple overall test of the two models can be obtained by examining the T-statistic against the 90% critical value of 1.645. Expected signs are positive for the SNAP participation ratio and the IRS tax-poor ratio, as they are both clearly positively related to poverty. For the dependency ratio, we generally expect a higher poverty rate in areas with higher proportions of children relative to the overall population, mostly because this is indicative of a younger workforce. But the relation might be weak. For the IRS filing coverage ratio, higher poverty means that more households will have incomes lower than required for filing a tax return, so we expect a negative coefficient here. For the jobs ratio, increased employment should generally be associated with lower poverty, on average. Industry-mix effects could mask this effect when looking at individual counties, however.

Tables 3 and 4 list the coefficient estimates and some summary statistics for the alternative specifications for the two different age groups. All SAIPE-style Fay-Herriot models on the log scale have high R-squared values, so for evaluation purposes, we have included the estimate of the model error variance, and the mean standard error of the shrinkage estimate as measures of overall fit. Using these measures, for either age group, the addition of the jobs ratio adds little to the overall fit. However, the jobs ratio coefficient is significant and the expected sign for all models except for the all-ages specification with the full variable list (wTax/wJobs).

**Table 3: Regression Results for Four Alternative Specifications, Age 5-17 in Families**
(T-statistic in parentheses)

| Predictor | | Specification | | | |
|---|---|---|---|---|---|
| | | noTax/noJobs | noTax/wJobs | wTax/noJobs | wTax/wJobs |
| Intercept | $\beta$ | 0.07 (5.52) | 0.14 (8.49) | -0.16 (-3.78) | -0.16 (-3.70) |
| SNAP Participation Ratio | $\beta_1$ | 0.59 (38.17) | 0.54 (31.58) | 0.31 (11.71) | 0.31 (11.83) |
| Dependency Ratio | $\beta_2$ | -0.08 (-1.41) | 0.003 (0.05) | 0.07 (1.29) | 0.12 (2.05) |
| IRS Child Tax-Poverty Ratio | $\beta_3$ | | | 0.55 (13.85) | 0.51 (12.18) |
| IRS Child Filing Coverage Ratio | $\beta_4$ | | | 0.12 (3.76) | 0.16 (4.68) |
| Jobs Ratio | $\beta_5$ | | -0.12 (-6.68) | | -0.07 (-3.57) |
| sqrt(model error variance) | | 0.074 | 0.074 | 0.072 | 0.071 |
| Mean standard error of shrinkage estimate | | 0.048 | 0.047 | 0.047 | 0.047 |
| Mean ACS percent in poverty, ages 5-17 | | 0.220 | 0.220 | 0.220 | 0.220 |

For the model of ages 5 to 17 in families, the only age-specific regressors are the ones derived from aggregate tax data, the tax poverty and filing coverage ratios. So the specifications in the left-most two columns have no age-specific regressors, and yet perform comparatively well, at least by the broad statistics given in Table 3. A more detailed look in Table 4 shows more advantages to using the tax data.

**Table 4: Regression Results for Four Alternative Specifications, All-ages**
(T-statistic in parentheses)

| Predictor | | Specification | | | |
|---|---|---|---|---|---|
| | | noTax/noJobs | noTax/wJobs | wTax/noJobs | wTax/wJobs |
| Intercept | $\beta_0$ | 0.10 (14.82) | 0.16 (19.15) | 0.26 (18.85) | 0.26 (18.90) |
| SNAP Participation Ratio | $\beta_1$ | 0.41 (52.01) | 0.37 (42.99) | 0.23 (19.14) | 0.23 (18.85) |
| Dependency Ratio | $\beta_2$ | -0.16 (-5.78) | -0.088 (-3.19) | 0.11 (3.87) | 0.11 (4.02) |
| IRS all-age Tax-Poverty Ratio | $\beta_3$ | | | 0.22 (9.01) | 0.23 (9.25) |
| IRS all-age Filing Coverage Ratio | $\beta_4$ | | | -0.28 (-17.73) | -0.30 (-16.14) |
| Jobs Ratio | $\beta_5$ | | -0.102 (-11.68) | | 0.022 (2.11) |
| sqrt(Model error variance) | | 0.037 | 0.035 | 0.032 | 0.032 |
| Mean standard error of shrinkage estimate | | 0.023 | 0.022 | 0.021 | 0.021 |
| Mean ACS percent in poverty, all ages | | 0.166 | 0.166 | 0.166 | 0.166 |

The interpretation of the small difference between the overall fit and standard errors despite the significance of the jobs ratio in the regression could be that the explanatory power of the other covariates, particularly the SNAP participation and tax ratios, are high enough that additional information provides little benefit.

Another potential explanation is that the standard error and model error variance estimates are derived from the assumptions of the model outlined in Equation (1), and in particular, the assumption of constant model error variance across all counties. Since in the maximum likelihood estimation larger counties, with lower sampling variance, will be weighted more strongly, the estimate of the model error variance, and standard errors, will be more greatly influenced by residuals for the larger counties.

Since the purpose of small area estimation programs, like SAIPE, are to supply good estimates of smaller counties and domains, some evaluation that doesn't rely on the constant model error variance is warranted. To supply such an evaluation, Tables 5 and 6 report an alternative measure of fit for partitions of the counties by population size. We do not use the calculated standard error in this case because the derived standard error is primarily dependent on the model error variance (Equation 3), which is assumed constant. Rather tables 5 and 6 report the adjusted root mean-squared difference

(ARMSD) between the ACS estimate and the fitted value, adjusted for the sampling variance. This measure is defined by Equation 4.

$$\mathrm{ARMSD}_k = \left(\frac{1}{K}\sum_{i=1}^{K}\left(y_i - x_i'\hat{\beta}\right)^2 - \frac{1}{K}\sum_{i=1}^{K} v_i\right)^{1/2} \qquad (4)$$

Where $y_i$ is the ACS poverty rate estimate as before, and $k$ represents a group of counties, $i = 1,2,\ldots K$, characterized in this case by a range of all-age population. Note that despite the adjustment for average sampling variance, in general this measure is expected to give an estimate of dispersion somewhat larger than the standard error of the estimator because it does not account for uncertainty in the ACS due to non-sampling errors. However, when comparing the performance of two different regression estimators as in this case, for the same set of counties and thus direct ACS estimates, it could provide a useful comparison regarding different specifications.

**Table 5: ARMSD for Population Partitions, Age 5-17 in Families**

| | Specification | | | |
|---|---|---|---|---|
| Population Size | noTax/noJobs | noTax/wJobs | wTax/noJobs | wTax/wJobs |
| 0-10,000 | 0.090 | 0.083 | 0.070 | 0.069 |
| 10,000-20,000 | 0.061 | 0.061 | 0.057 | 0.058 |
| 20,000-65,000 | 0.059 | 0.058 | 0.051 | 0.052 |
| 65,000+ | 0.045 | 0.043 | 0.035 | 0.035 |

**Table 6: ARMSD for Population Partitions, All Ages**

| | Specification | | | |
|---|---|---|---|---|
| Population Size | noTax/noJobs | noTax/wJobs | wTax/noJobs | wTax/wJobs |
| 0-10,000 | 0.049 | 0.045 | 0.043 | 0.044 |
| 10,000-20,000 | 0.034 | 0.033 | 0.030 | 0.030 |
| 20,000-65,000 | 0.037 | 0.036 | 0.030 | 0.030 |
| 65,000+ | 0.036 | 0.033 | 0.026 | 0.026 |

No distribution has been derived for this statistic, so no statistical comparisons can be performed. For predicting the poverty of children ages 5 to 17 in families, the addition of the jobs ratio does little to decrease error for any population size when the tax-based variables are included. In models where the tax variables are omitted, however, there is a difference. For smaller population sizes, the error for the model including the jobs ratio is approximately 8% lower than the error associated with the model that does not include the jobs ratio. This may or may not be statistically significant. The same trend is present for predicting the poverty of people of all ages, with the largest differences for the smallest population sizes.

## 4. Conclusion

Overall, we found the inclusion of an employment-derived regressor resulted in a small improvement in the dispersion of the predictors for less populous counties. This effect was seen only when tax-derived regressors were not included in the model. Further research will include relaxing the constant model error variance assumption, as the adjusted RMSD evaluation yields some evidence of variable predictor dispersion across population size. Also, variations of the employment-derived regressor could be constructed. Particularly, industry exposure indices, such as construction or retail industry dependence might be explanatory for poverty during some periods of the business cycle.

## References

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

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

Fay, R. E., III, and R.A. Herriot. June, 1979. Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association.* Vol. 74, No. 366, 269-277.

U.S. Census Bureau. 2009. Chapter 12: Variance Estimation. *ACS Design and Methodology*. <http://www.census.gov/acs/www/methodology/methodology_main>