Evaluation of School District Poverty Estimates: Predictive Models using IRS Income Tax Data[1]

Jerry J. Maples (jerry.j.maples@census.gov) and William R. Bell
Bureau of the Census, Washington DC 20233

Keywords: small area estimation, share models, poverty estimates, IRS income tax data

1. Introduction

The U.S. Census Bureau, with support from other Federal agencies, created the Small Area Income and Poverty Estimates (SAIPE) program to provide more current estimates of selected income and poverty statistics than are available from the most recent decennial census. Estimates are created for states, counties, and school districts. The main objective of this program is to provide updated estimates of income and poverty statistics for the administration of federal programs and the allocation of federal funds to local jurisdictions. In addition to these federal programs, there are hundreds of state and local programs that depend on income and poverty estimates for distributing funds and managing programs (U.S. Census Bureau, 2005). In this paper we will be focusing on estimating the number of poor school-age children (between the ages of 5 and 17) for every school district.

The No Child Left Behind Act of 2001 directs the Department of Education to distribute Title I basic and concentration grants directly to school districts on the basis of the most recent Census Bureau estimates of school-age children in poverty in each school district in the U.S. When constructing these estimates we first split up school districts that cross over county boundaries into *school district pieces* corresponding to the parts of the district that overlap each county. We form estimates for the school district pieces and then aggregate these results across pieces within each school district. (Many school districts, however, are equal to or are contained within a single county, and so have only one piece.) Breaking the school districts into these pieces facilitates controlling the school district estimates to agree with SAIPE county level estimates for the number of poor school-age children. For this research, we are using school districts as defined by the 1999-2000 school district boundaries. There are 14,334 school districts which are split up into 20,177 school district pieces.

SAIPE estimates for states and counties make use of data from the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC). The CPS ASEC is designed, however, for estimates at the national level, and its sample sizes for individual states and counties are not large enough to produce sufficiently reliable direct estimates except for a few of the largest states and counties. Hence, small area models like that of Fay and Herriott (1979) are used to improve the estimates. The models relate true poverty to other variables obtained from administrative records sources including IRS income tax data and food stamp program participation data, as well as corresponding poverty estimates from the previous decennial census. Additionally, all estimates are raked to be consistent at aggregate levels, e.g. the state estimates are raked to be consistent with the direct national poverty estimates given by the CPS ASEC, and the SAIPE county estimates are raked to agree with the state estimates.

Estimation for school districts presents more severe data problems. First, CPS ASEC data is much too sparse to use at the school district level, the problem being more severe than that for counties due to the much larger number of school districts (14,334 school districts versus 3,141 counties). The large number of school districts with no or very small CPS samples would make it difficult even to use CPS ASEC data in school district level models.

[1]Disclaimer: This presentation is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical and methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.

School district estimates from the American Community Survey (ACS) are some years away until the ACS can accumulate up to five years of data for the small districts. Data from the administrative records sources used in the state and county models have previously not been tabulated down to the level of school districts. Administrative data that have been available for school districts (free and reduced price lunch data and school enrollment data) face data quality problems and other issues (e.g., school enrollment data cover only children in public school and not children who are in private or parochial schools, or are home-schooled.)

Given these data limitations, SAIPE school district estimates for post-censal years have used a crude updating scheme that assumes that the ratio of poor school-age children in a district piece to the county number of poor school-age children remains constant over time, with this ratio estimated from results of the previous census. These ratios are then carried forward for the post-censal years and multiplied by updated county estimates of the number of poor school-age children obtained from the SAIPE county model. (Section 3 gives a mathematical formulation of this procedure.) While this procedure uses updated information about poverty at the county level (via the county model), it does not account for any changes in poverty within counties that differentially affect school districts. (This is apart from school district boundary changes, which are accounted for by retabulating the previous census results each year.)

Recently, however, IRS income tax data has been tabulated for school districts for possible use in formulating population and poverty estimates at the school district level. Maples and Bell (2004) investigated this possibility by fitting models relating census school district poverty estimates to school district tabulations of income tax data. The present paper discusses an extension of this work to an evaluation of school district poverty estimators that make use of IRS data. Different school district poverty estimates are obtained for income year (IY) 1999 and compared against corresponding estimates for the 2000 census to assess the accuracy of the various estimates in comparison to the official SAIPE school district estimates (that involve crude updating of results from the 1990 census). The various estimation procedures are then reversed in time to "predict" school district poverty in IY 1989 and compared against corresponding estimates from the 1990 census to provide a second evaluation of the alternative estimators.

Section 2 of this paper discusses the IRS data used in the models and the issues that arise in tabulating these data for school district pieces. Section 3 discusses the alternative school district estimation methods considered including the official SAIPE method. Section 4 then discusses results from the evaluations of these alternative estimators, and Section 5 summarizes the conclusions.

2. IRS Income Tax Data at the School District Level

Recently, IRS income tax data have been tabulated at the level of school district pieces to investigate the possible benefits of using such data for constructing poverty estimates at the school district level. Each income tax return contributes a number of exemptions some of which are identified as dependent child exemptions. If a return reports an adjusted gross income below the official poverty threshold for a family of the size implied by the number of exemptions (total and child) on the return, then all of the exemptions on that return are considered to be "in poverty," which we label as "poor exemptions." Thus, for our purposes here, we record four main quantities from each return: total exemptions, total poor exemptions, child exemptions, and poor child exemptions.

To tabulate any of these four quantities up to the level of school district pieces we must first assign income tax returns to individual school district pieces. This process first goes through an intermediate step where we attempt to *geocode* the address of the tax return to a census block. Then we tabulate results over all census blocks within each given school district piece. As we cannot successfully geocode all tax returns to census blocks we are left with a pool of non-geocoded exemptions to deal with. We discuss this issue in the next section. Work is currently ongoing to determine if many of the currently non-geocoded exemptions can be geocoded directly to the larger school district pieces instead of to census blocks in order to

reduce the number of non-geocoded exemptions.

Another problem is that some school districts have overlapping boundaries, as when an area is serviced by separate elementary and secondary school districts. The process for assigning tax exemptions to school districts does not take this into account; therefore, the same tax exemptions may be assigned to multiple school districts. In the next section, we also address this issue.

2.1 Adjusted Income Tax Exemptions

As just noted, before tabulations of tax exemptions can be used in any model, two issues need to be addressed. One issue concerns the age range of children that are serviced by the school district. Although no age is recorded for child exemptions on tax returns, we will assume that child exemptions refer to the population of children ages 17 years old and under. We obtain the grade range for each school district from the NCES (National Center for Education Statistics) Common Core of Data. The most typical grade ranges are unified (K-12), elementary (K-8), and secondary (9-12). In some areas of 17 states there are separate elementary and secondary school districts, each exclusively responsible for providing education in a subset of the grades in their shared territory. In these areas, the census estimates of poor children (refered to as "relevant" children) reflect the grade range of the school districts. Therefore, we will modify the tabulated numbers of child tax exemptions and poor child tax exemptions to reflect the grade range of the school district.

We are attempting to make estimates relative to school-age child population. For a unified district this includes grades K through 12, which we assume corresponds to children aged 5-17. A simple assumption we could make would be that the distribution over single years of age among children 0 to 17 years old is uniform. However, demographic estimates for number of children by each single year of age are available at the county level. These estimates refer to the population as of July 1st for a given year. Using this demographic county level data we can account for differences in age distributions between counties. Most counties (2,874 out of 3,141) only contain school districts that are unified and for districts in these counties the adjustments for age corresponding to grade range cancel out, that is, all districts within the county are adjusted by the same proportional amount. Only for 267 counties with school districts that do not contain the full grade range (K though 12) do the age-grade adjustments differentially affect school district pieces within the counties. To construct the age-grade range adjustment we multiply the number of geocoded child tax exemptions in school district piece j within county i, denoted $T_{g,ij}$, by the proportion from the demographic population estimates of the 0-17 year-old children in the county who are in that school district age-grade range. We denote this adjustment factor by π_{ij} and thus have an age adjusted geocoded child exemption count ($T^*_{g,ij}$) of

$$T^*_{g,ij} = \pi_{ij} \times T_{g,ij}$$

for school district piece j in county i.

The second issue concerns the pool of non-geocoded child tax exemptions in each county, the number of which we denote by $T_{ng,i}$. The non-geocoded exemptions are first adjusted as just discussed to reflect the target population of age 5 to 17 year old children, i.e., we form $T^*_{ng,i} = \pi^c_i \times T_{ng,i}$, where π^c_i is the county i ratio of 5-17 year-old children to 0-17 year-old children from the demographic population estimates. We then assume that the non-geocoded exemptions are distributed among the districts within a county proportional to the school-age population that resides in the school district pieces. For IY 1999 we can use the Census 2000 short form population count of school-age children for this purpose, but for a post-censal year we must instead use estimates of school district school age population. Let p_{ij} denote this estimated proportion of the county school-age population that resides in school district piece j of county i. We assign $p_{ij} \times T^*_{ng,i}$ of the non-geocoded exemptions to school district piece j. Combining the grade range adjustment and non-geocoded exemption adjustment together, we get the final adjusted number of poor child tax exemptions for a school district piece, denoted T_{ij}, and given by

$$\begin{aligned} T_{ij} &= T^*_{g,ij} + (p_{ij} \times T^*_{ng,i}) \\ &= (\pi_{ij} \times T_{g,ij}) + (p_{ij} \times \pi^c_i \times T_{ng,i}) \end{aligned}$$

We do not round T_{ij} to be an integer. We use this procedure for making age-grade range adjustments and dealing with non-geocoded exemptions with both total child exemptions and poor child exemptions. To distinguish these two we denote the allocated total and poor child exemptions by T_{ij}^{tot} and T_{ij}^{poor}, respectively.

3. Models

Our goal is to estimate the number of poor children in each school district for post-censal years. The unit of analysis is the school district piece. We have a collection of school district pieces ($j = 1, ..., J_i$) in each county i ($i = 1, ..., I$). Data from the 2000 and 1990 Census long form (estimates of related children aged 5-17 in families in poverty) and data from IRS income tax returns for income years 1989 and 1999 will be used to fit, evaluate, and validate the models. Preliminary work by Maples and Bell (2004) using the 2000 Census and 1999 IY IRS income tax data have shown that models for school district piece to county poverty shares based on tax data work better than modeling the number of poor children in school district pieces directly.

The current methodology to estimate the number of poor children in school district pieces for a post-censal year is based on a synthetic approach. The most recent census data are used to estimate school district piece to county shares of poverty, and these estimated shares are then multiplied by the SAIPE model-based county estimate for the current year:

$$\text{Poor } \widehat{\text{Children}}_{ij} = \frac{\text{Census}_{ij}}{\sum_j \text{Census}_{ij}} \times CNTY_i \quad (1)$$

where Census_{ij} is the previous census long form estimate of related 5-17 children in families in poverty for school district piece j of county i, and $CNTY_i$ is the SAIPE model-based estimate for county i. The underlying assumption with this approach is that the distribution of poverty among school district pieces within a county does not change between censuses. We want to explore estimation methods that use the current-year IRS data to reflect changes over time in the distribution of poverty within the county.

3.1 Share Models

Since any estimates for school district pieces will be controlled to the official county estimate of poor children age 5-17 to maintain consistency, it is sufficient to estimate the school district piece to county poverty share. Share models thus attempt to describe the distribution of poor 5-17 children among the school district pieces within a county. We can also view the shares as the probability that each poor 5-17 child in the county should assigned to a particular school district piece. Note that within a county the estimated shares must add up to 100 percent.

We will present four competing models for estimating the school district piece to county poverty shares of 5-17 related children:

1. census share: use shares from the most recent census (official method)
2. tax poor share: use shares of the adjusted poor child tax exemptions
3. tax poverty rate share (a): use the IRS data to estimate a "child poverty rate" for each school district piece, multiply these poverty rates by the official estimate of the child population for the school district piece, and compute poverty shares for the school district pieces from these results
4. tax poverty rate share (b): same as approach 3 except replace the official child population estimates for the school district pieces by a new estimate of child population that makes use of the IRS child tax exemption data for the school district pieces.

The new estimate of the child population of school district pieces will be discussed shortly.

We now formalize these four estimators using notation specific to estimation for IY 1999 and that distinguishes between estimates from the 1990 and 2000 censuses. Let $C_{2K,ij}$ be the Census 2000 estimate of related 5-17 children in families in poverty for school district piece j of county i, and let $T_{99,ij}^{\text{poor}}$ and $T_{99,ij}^{\text{tot}}$ be the number of adjusted poor child exemptions and number of adjusted total child exemptions for IY 1999 for school district piece j

in county i constructed as discussed in the previous section. We similarly define $C_{90,ij}$, $T^{\text{poor}}_{89,ij}$, and $T^{\text{tot}}_{89,ij}$ for Census 1990 estimates and IRS poor child and total child exemptions for IY 1989 respectively. Instead of using the SAIPE model-based county estimate of the number of poor 5-17 children ($CNTY_{99,i}$) to scale up our share estimates for IY 1999, we shall instead use the Census 2000 county estimate, $C_{2K,i\cdot}$. Since we will use the Census 2000 school district estimates for evaluation, we are thus pretending that we have a "perfect" county estimator. We do this for the purposes of this paper because we only want to identify models that best describe the distribution of poor school aged children within a given county.

The census share estimator (Method 1) for IY 1999 corresponds to (1), but modified to replace the SAIPE model-based county estimate by the Census 2000 county estimate:

$$EST_{1,ij} = \frac{C_{90,ij}}{\sum_j C_{90,ij}} C_{2K,i\cdot}$$

A generalization of the tax poor share approach (method 2) has the form:

$$\begin{aligned} EST_{2,ij} &= \frac{(T^{\text{poor}}_{99,ij})^{\beta}}{\sum_j (T^{\text{poor}}_{99,ij})^{\beta}} C_{2K,i\cdot} \\ &= \frac{\exp(\beta \log T^{\text{poor}}_{99,ij})}{\sum_j \exp(\beta \log T^{\text{poor}}_{99,ij})} C_{2K,i\cdot} \quad (2) \end{aligned}$$

This model form of exponentially weighted shares can be derived from a model of the log number of poor children where the intercept is allowed to vary by county. A special case of this model when $\beta = 1$ is just the simple tax poor share approach discussed earlier. This model lacks an intercept term in the $\exp(\cdot)$ part because it would factor out and cancel with the same term in the denominator, and so would be unidentifiable. Estimates of the parameters in (2) fitted to both the 1990 and 2000 Census data (modifying $T^{\text{poor}}_{99,ij}$ to $T^{\text{poor}}_{89,ij}$ when fitting to the 1990 census data) are given in Table 1. In both of the census years, we have β parameters whose 95% confidence interval contains $\beta = 1$. Thus, we cannot outright reject using a simple share method in favor of this more general method.

Table 1 - Parameter estimates from the modeled tax share method

Year	β	Std. Err.
1990	1.022	.048
2000	1.024	.048

For the tax poverty rate share model we first create a pseudo-estimate of the poor 5-17 children in a school district piece by multiplying the IRS income tax poverty rate, based on adjusted poor child and total child exemptions, by the estimated number of school age children.

$$Y_{ij} = \frac{T^{\text{poor}}_{99,ij}}{T^{\text{tot}}_{99,ij}} \widehat{\text{Child Pop}}_{ij} \quad (3)$$

$$EST_{3,ij} = \frac{Y_{ij}}{\sum_j Y_{ij}} C_{2K,i\cdot} \quad (4)$$

A full discussion of the estimation of the child population in school district pieces is beyond the scope of this paper (report forthcoming). The current production approach is analogous to (1) but uses census counts of total children rather than census estimates of poor children, and replaces the county model-based child poverty estimate by a demographic county child population estimate. An alternative child population estimate for 2000 that uses current IRS income tax data (for IY 1999) and the most recent census data (Census 1990) is

$$\begin{aligned} \widehat{\text{Child Pop}}_{ij} &= \frac{1}{2}\left(\frac{T^{\text{tot}}_{99,ij}}{\sum_j T^{\text{tot}}_{99,ij}} + \frac{C^{\text{tot}}_{90,ij}}{\sum_j C^{\text{tot}}_{90,ij}} \right) \\ &\quad \times [\text{County Child Pop 5-17}]_i \quad (5) \end{aligned}$$

where $C^{\text{tot}}_{90,ij}$ is the 100% count of the number of relevant children in school district piece j of county i from the 1990 census. Our alternative estimate averages the shares from the IRS income tax data and the most recent census and then multiplies that share by the estimated number of children age 5-17 in the county. Preliminary results show that this hybrid population estimator performs better than either of the two individual population estimators. Additional work on improving school district child population estimators is currently being done in

an independent project. We use these child population estimators (either the official estimator or the alternative given by (5)) rather than the actual Census 2000 population counts because use of the latter would add extra information to the poverty rate share estimators, which could lead to overestimating the accuracy of the child poverty estimates for the tax poverty rate share methods compared to the other two methods. Also, in practice when we are estimating for a post-census year, we will have to produce an estimate of school district piece child population.

4. Evaluation of Estimators

Our goal is to develop an estimator based on the IRS income tax data that performs better than the current production methodology, most recent census share. Although in practice, we would not use the 1990 Census shares to make estimates for income year 1999 since we have the Census 2000 available, we will use the 1990 Census data to form the census shares to replicate the official methodology for non-census years pretending that we do not have the Census 2000 data, except for evaluation purposes.

In this section we compare the accuracy of the various estimators at predicting the Census 2000 estimates. Although we make estimates of school district pieces, our goal is to make the most precise estimate for whole school districts. Thus, our unit of analysis for evaluation will be whole school districts. Our metric to compare estimators is the mean squared difference of the log number of poor 5-17 children estimated by our models compared to the corresponding Census 2000 long form estimate:

$$\text{MSDiff} = 1/N \sum_{sd} [log(C_{sd}+1) - log(EST_{sd}+1)]^2$$

summing over all school districts subscripted by *sd*. We use $log(x+1)$ to deal with the occasional zeros in the data as these would distort the comparisons and arise only for the smallest of school districts, with virtually no effect on the mid-sized and larger districts. In addition to looking at the overall MSE for the estimators, we also want to consider categorizing the school districts by child population size (small, medium and large) given by the Census 2000 population counts, according to the categories given in Table 2.

Table 2 - School District Size Categories

Category	Size	frequency
Small	1-499	4424
Medium	500-1999	4877
Large	2000+	5007

Table 3 - Mean Squared Difference by Size categories Estimating IY 1999

Size	Off.	Tax poor share	Tax pov rt. (a)	Tax pov rt. (b)
All	.39	.34	.32	.31
Small	.71	.63	.62	.58
Med.	.36	.32	.29	.29
Large	.15	.11	.09	.09

Table 3 presents mean squared differences for the different estimators by size of school districts. Our goal is to identify an estimator which uses the IRS income tax data and performs better, i.e. has lower MSDiff, than the current official method, when compared against Census 2000 estimates.

Comparing across all school districts, we see that all the estimators using IRS information performed better than the official method. The tax poverty rate share using IRS data in constructing the population estimates had the lowest MSDiff overall and for all size categories. The difference between this estimator and the tax poverty rate share estimator that uses the official child population estimate was small, however.

An analogous analysis can be done by reversing the estimators in time, taking Census 2000 as the "previous census," estimating 5-17 poverty for school districts in IY 1989, and comparing the estimates against the 1990 Census child poverty estimates for school districts. This gives us a second time point to compare the performance of the alternative school district 5-17 child poverty estimators.

One major difference between the 1989 and 1999 IRS income tax data is the overall non-geocoding rates, 27% and 15% respectively. One possible reason for this large difference is that the income 1989 income tax data had to be mapped on to 1999-

Table 4 - Mean Squared Difference by Size categories Estimating IY 1989

Size	Off.	Tax poor share	Tax pov rt. (a)	Tax pov rt. (b)
All	.50	.53	.53	.50
Small	.98	1.01	1.02	.96
Med.	.39	.42	.41	.39
Large	.17	.23	.22	.22

2000 geography and the time lag may have made this more difficult. The 1990 census block geography did not exactly correspond to the 2000 census block geography and addresses can change or be removed over time.

The mean squared differences for the estimates of Census 1989 are shown in Table 4. Again, the tax poverty rate share method (b) is the best of the IRS income tax data based methods. However, this estimator is not uniformly better than the official method, specifically, it does worse for the large school districts. Overall and for small and medium size school districts, the tax poverty rate share model does about the same or only slightly better than the official method. At this time, it is not clear why the estimators using IRS income tax data do not work as well for estimating IY 1989 as they do for IY 1999. It may have something to do with differences between geocoding rates of the 1989 and 1999 IRS income tax data. The *non-geocoding rates* in 1989 and 1999 were, 27% and 15% respectively, and it may be that the higher non-geocoding rate in 1989 compromised the estimators for IY 1989 that made use of the IRS data. We plan to investigate this possibility. (Note: A possible reason for the large difference in geocoding rates is that the IY 1989 tax data had to be mapped onto 1999-2000 geographic boundaries, and the time lag betwen the reference year for the data and that of the boudaries may have made the geocoding more difficult. The 1990 census block geography did not exactly correspond to the 2000 census block geography and addresses could change or be removed over time.)

5. Discussion

We presented various models that make use of the number of poor child tax exemptions in constructing estimates of the number of poor children in each school district. Estimates were constructed and their accuracy assessed by comparing them to estimates from the Census 2000 and 1990 Census long form results. Results of Maples and Bell (2004) showed the IRS tax data to be very useful in modeling the Census 2000 long form child poverty estimates, and the evaluation results presented here also show improvements in estimates for IY 1999 from the methods that make use of the 1999 tax data. However, the results are not as clear when evaluating estimates for IY 1989 against the 1990 census child poverty estimates. Overall, then, the methods that use IRS income tax data show promise, but more work is needed to understand why IY 1989 income tax data does not predict Census 1990 results as well as the official method.

There are still issues that need to be addressed regarding use of the IRS tax data at the school district level. First, some counties have a large percentage of non-geocoded tax exemptions. This makes it difficult to use the tax data for these counties. By distributing the non-geocoded exemptions proportionally to the child population estimates, we are making assumptions about the geocoding process. Mainly, we assume that all child exemptions have an equal probability of not being geocoded regardless of which school district piece of the county they belong to. Note that the variable used to proportionally allocate the non-geocoded exemptions, the child population count. Improvements in the geocoding process can greatly reduce the errors that arise from allocating the non-geocoded exemptions. Another issue with the tax data is that we do not know the non-filing rates for school district pieces within the county. Any difference between the true population count and the number of tax exemptions can be due to any combination of non-geocoded exemptions and non-filing. Our models implicitly assume that the non-filing rates are constant throughout the county.

One issue not addressed in this paper is the estimation of variance for these estimators. The estimators are subject to multiple sources of error.

First, there is error in the estimate of child population in a school district piece. Second, there is error in the county estimate of the number of poor children. Third, there is error associated with having to allocate the non-geocoded exemptions from the IRS income tax data. Finally, there is the error in estimating the school district piece to county poverty shares for 5-17 children.

References

Maples, J. and Bell, W. (2004), "Investigating the use of IRS Tax Data in the SAIPE School District Poverty Estimates," Proceedings of the American Statistical Association.

U.S. Census Bureau (2005), "Small Area and Income Estimates - School District Estimates," http://www.census.gov/hhes/www/saipe/, accessed on Oct 11, 2005.

Fay, R. E. and Herriot, R. A. (1979) "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, **74**, 269-277.