Investigating the use of IRS Tax Data in the SAIPE School District Poverty Estimates [1]

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

Keywords: small area estimates, share models, logistic regression

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 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. For example, 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 estimates of children in poverty available from the Census Bureau (National Research Council, 2000). 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. In this paper we focus on estimation of the number of poor school aged children (between the ages of 5 and 17) for every school district.

The geographic boundaries of a school district do not always fall within a single county. Before estimating the number of poor children in each school district, we first split up the school districts that cross over county boundaries into school district pieces that fall within each county, i.e., pieces are defined by the intersection of county and school district geographic boundaries. Estimates are made for the school district pieces and aggregated to form the estimate for the school district. The main reason to split school districts into pieces is to make estimates for the number of poor children that are consistent with the official estimated number of poor children for the entire county. Since the SAIPE county model gives updated estimates of the number of poor children within each county, it is sufficient to develop a model which reflects the changing distribution of poor children between school district pieces within a county. All estimates from models at the school district piece level will be ratio adjusted (raked) so that the totals are consistent with the SAIPE county estimates.

In previous census years, the estimated number of poor children has been obtained from the Census long form results. In non-census years, the current methodology used to estimate the number of poor children is based on two quantities (U.S. Census Bureau, 2004). The first part is the proportion of the long form model-based estimate of the number of poor children for the school district piece to the total number of poor children in the county for the census year. This creates a within county share and the shares within every county sum up to 100 percent. The second part is the SAIPE model-based estimated number of poor children in the county for the year of interest. Thus, given a geography for a set of school districts within the county, we assume that the distribution of poor children within the county is the same as the distribution from the last census, i.e. the shares remain constant. By construction, the SAIPE school district estimates are arithmetically consistent with the SAIPE county estimates. The numbers of children in poverty are then adjusted using controlled rounding to get a result with the following properties:

1. The number of children in poverty for the

[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 pieces in a county adds up to the SAIPE estimate of the number of children in poverty in their counties.

2. The number of children in poverty in the school district pieces are integers.

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

The estimation of the number of poor children in school districts is a hard problem for two reasons. First, the best source of data has been from the census long form which has been only collected once a decade. The Current Population Survey (CPS) provides basic input data for the county model, but it is much too sparse to use at the school district level. The second problem is the lack of updated information collected at the school district piece level. None of the current ongoing national surveys are designed to collect data for all school district pieces. This may change once the American Community Survey is fully implemented.

Administrative data sources such as IRS income tax data may provide more current information about child poverty than the last census. Variables that are based on schools, such as free/reduced price lunch participation and school enrollment, can not be broken into their respective school district pieces because it is not clear how many children in each county are serviced by a school that is in a district which crosses county lines. Also, school enrollment data only takes into account children in public school and not children in private or parochial schools, or home-schooled. At this time, the IRS income tax data appears to be the best source of information for use in modeling the number of poor children in school district pieces.

From the evaluation by Bell and Kramer (1999), the current methodology for estimating the number of poor children in school districts has large relative error. The goal of this paper is to develop a refinement of the SAIPE school district poverty estimates. Specifically, we want to use the IRS income tax data instead of the most current census data to estimate the within county share of poor children. We proposed that the IRS income tax data should give more relevant information about the distribution of poor children within a county for the corresponding non-census year than the census data. In the next section, we will describe the IRS income tax data and show how the data can be made more relevant to the population of interest, poor school-aged children. Several different models will be discussed in Section 3. A comparative evaluation of the fit of the proposed models will be done in Section 4. Finally, conclusions and limitations of using the IRS income tax data will be discussed in Section 5.

2 IRS Income Tax Data

In the current production models for both state and county estimates for the number of poor, the IRS income tax data is a useful predictor (U.S. Census Bureau, 2004). In the state model, a "tax return poverty rate" and nonfiler rate both help to predict the state poverty rate, where the "tax return poverty rate" is a ratio with numerator being the number of exemptions on income tax returns with adjusted gross income below the poverty threshold (which depends on the size of the household given by the total number of exemptions on the tax return) and denominator being the total number of exemptions. In the county model, the log number of exemptions on income tax returns with adjusted gross income below the poverty threshold (same criteria as for the state model) is used in the model for the number of poor. Recently, IRS income tax data has been tabulated for school districts and school district pieces to be used for estimates at the school district level. Each return contributes a number of total exemptions and a number of child exemptions. Also, if a return has an adjusted gross income below the official poverty threshold for a family of the size implied by the number of exemptions on the return, then all of the exemptions on that return are considered to be poor exemptions. Thus, there are four main quantities recorded from each return: total exemptions, total poor exemptions, child exemptions and poor child exemptions.

In order to assign the income tax returns to the school district pieces, the return must first be geocoded into a census block based on the home address of the return. Not all returns are able to have their home address successfully geocoded to a census block. However, we can assign every tax return to a county, which we assume is done without error. Therefore, we can compute the geocoding rates for the various types of exemptions for each county. After all of the returns are geocoded to census blocks, or not geocoded to any block, the exemptions in the blocks are tabulated for school district pieces based on a given set of geographical boundaries.

Some school districts have overlapping boundaries. Often this is due to an area being serviced by separate elementary and secondary school districts. The process for assigning tax exemptions into school districts does not take this into account. Therefore, the same tax exemptions may be assigned to multiple school districts. Between this and the geocoding issues mentioned earlier, we cannot simply use the tabulated number of tax exemptions as given. In the next section, we will address how to modify the tabulated numbers of tax exemptions to deal with these two issues.

3 Models

Our goal is to model the number of poor children in each school district. Our unit of analysis is the school district piece. We have a collection of school district pieces ($j = 1, ..., J_i$) in county i ($i = 1, ..., I$). Data from Census 2000 long form estimates of related children aged 5-17 in families in poverty will be used to fit and evaluate the model. Since the data is from the long form, we also have an estimate of the sampling error variance. The explanatory variable will be the number of poor child tax exemptions from the IRS income tax data. We limit our analysis to counties which contain more than one school district piece. Share models obtain no information from counties that only contain a single school district piece as there is no within county variation to model. Also, we exclude school district pieces that lack any census long form data.

In each county i, there is a tabulation of geocoded exemptions for each school district piece j, $x_{ij,g}$, and a tabulation of non geocoded exemptions, $x_{i,ng}$. Before the tax exemptions can be used in any model, two issues need to be addressed. The first issue is the age range of children that is serviced by the school districts. Although no age is given for child exemptions on the tax data, we will assume that it covers the entire under 18 age group. For each school district we have the grade range from the NCES 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 to their respective grades in their shared territory. In these areas, exemptions for school-age children are allocated between districts in which they reside on the basis of the grade range of the district and the grade assigned to the child. The census long form estimates for poor children reflect the grade range of the school districts. Therefore, we need to modify the number of poor child tax exemptions to reflect the grade range of the school district. Let A_{ij} be the grade range adjustment factor for school district piece (i, j). We will assume that the grade/age distribution is uniform within county

$$A_{ij} = \text{grade range}/18.$$

For example, a unified school district with grade range K-12 will have $A_{ij} = 13/18$. Thus, $x^a_{ij,g} = x_{ij,g} \times A_{ij}$.

The second issue is how to deal with the non-geocoded poor child exemptions, $x_{i,ng}$, in each county. We must find a process to distribute these exemptions down into the school district pieces within the county. The non-geocoded exemptions are first adjusted to reflect the target population of age 5 to 17 year old children (grade range K-12), $x^a_{i,ng} = x_{i,ng} \times 13/18$. We assume that the non-geocoded exemptions are distributed among the school district pieces by a multinomial process with the probability of an exemption belonging to a particular piece being proportional to the number of relevant school aged children in that school district piece compared to the county:

$$p_{ij} = \frac{\text{School Dist Piece child pop}_{ij}}{\text{County child pop}_i}.$$

By using a multinomial process with this probability structure, we are implicitly assuming that the probability of each poor child exemption to be a non-geocoded exemption is constant within a county. We will assign the expected number of non-geocoded exemptions under the multinomial process to each piece. This gives a variable that measures the grade range appropriate number of exemptions with a correction for the non-geocoded exemptions.

$$x_{ij} = x^a_{ij,g} + (p_{ij} \times x^a_{i,ng})$$

Since we put a probability distribution on the non-geocoded exemptions, we can also compute the variance added to our exemption variable due to non-geocoded allocation:

$$Var(x_{i,j}) = \sigma^2_{xij} = x^a_{i,ng} p_{ij}(1 - p_{ij}). \qquad (1)$$

We will use this variance in number of poor child exemptions to reflect that we do not observe the true number of poor exemptions.

3.1 Log Count Models

Traditionally, we model the log transformation of count data to stabilize the variance. This is equivalent to assuming that the variance of the relative percent error is constant. We model the log number of poor children using the log number of poor child exemptions.

$$\begin{aligned} \log(y_{ij}) &= \log(y^T_{ij}) + e_{ij} \\ \log(y^T_{ij}) &= \beta_0 + \beta_1 \log(x_{ij}) + m_{ij} \qquad (2) \end{aligned}$$

where x_{ij} is the number of poor child tax exemptions as defined in the previous section. The true number of poor children is y^T_{ij} (the 'T' superscript denotes the true value) but we only get to observe y_{ij} with sampling error e_{ij} with variance σ^2_{eij}. Although the sampling error variance is also estimated, we will treat it as known. The log of the true number of poor children equals the regression function in (2) plus the model error m_{ij} with variance σ^2_m.

We obtain the following fit of this model:

Parameter	Estimate	Std. Error
β_0	0.141	.0123
β_1	0.929	.0023
σ^2_m	.132	

The coefficient of variation (CV) for this model is $\sqrt{\sigma^2_m} = 36.3\%$. Thus, one standard error is about one third the magnitude of the estimate. This model had an R-squared of .9013.

We assume that y^T_{ij} comes from a log-normal distribution and transform back to the original scale:

$$\hat{Y}_{ij} = \exp(\hat{\beta}_0 + \hat{\beta}_1 \log(x_{ij}) + \hat{\sigma}^2_m/2).$$

In order to have consistency with the county estimates, we rake the school district pieces, i.e. ratio adjust, so that they sum up to the county estimates. Log count models are not additive. If a log count model at the county level is true, then a log count model based on the same covariates at a finer level, such as school district pieces, will not sum up to give consistent results at the county level in general. This is a problem with most nonlinear models. Modifications to the log count model can be made so that the school district piece and county models agree on their first two moments (mean and variance), but it is unclear how to interpret such a model.

3.2 Share Models

Share models attempt to describe the distribution of counts between the school districts within a county. These models are conditional on the observed or estimated county total. The shares are the proportion of counts that are within a specific piece of the whole. We can view the shares as the probability that each poor child in the county should be assigned to a particular school district piece. One feature of share models is that within a county, the shares add up to 100 percent. While one can put a regression model on the share (or transformation of the share), care will need to be taken to make sure that the shares remain within valid range (0 to 1) and that they are ratio adjusted to preserve their sum to 100 percent.

We will present several share type models:

1. direct share: using the actual poor child tax shares (this can be viewed as a special case of #2).
2. modeled share: exponentially weighted proportions of poor child tax exemptions

3. linear shares: linear model of the IRS shares to census shares

4. logistic shares: model the logistic transformation of the shares

5. log shares: model the log transform of the shares

On models 3-5, we will have to adjust the estimates so that they are consistent with the county totals since those models do not guarantee that the estimates match. In the next section, we will assess the various share models by comparing mean squared errors.

In the most general form, we want to estimate the number of poor children in school district piece j county i as the product of the county number of poor children and the share of the county poor for the school district piece

$$\hat{y}_{ij} = g_{ij}(x) y_{i+}^T$$

where $g_{ij}(x)$ is a model-based estimate of the share for the school district piece as a function of the IRS income tax data. In our models, the function $g_{ij}(x) = g(x_i)$ depends only on the vector of poor child tax exemptions for the pieces of a particular county. However, we do not observe the true county total poor and must base the estimate on the census county total of poor children, y_{i+}, with sampling error variance σ^2_{ei+}. This gives our general estimator the form

$$\hat{y}_{ij} = g_{ij}(x) y_{i+}. \qquad (3)$$

All of the models presented in this section fall under the general form given by (3). For models which may require raking (models 3-5), the $g_{ij}(x)$ is the estimated share after the raking process.

Let x_{sij} be the share of the number of poor child exemptions for school district piece j within county i. The direct share method models the number of poor children as follows:

$$\hat{y}_{ij} = x_{sij} y_{i+}.$$

In this model, $g_{ij}(x) = x_{sij}$ and there are no parameters to estimate. The variance of the census estimate will be used for model evaluations in the next section.

The next model is a generalization of the direct share method. An exponentially weighted share is modeled as follows:

$$y_{ij} = \frac{x_{ij}^{\beta}}{\sum_j x_{ij}^{\beta}} y_{i+} = \frac{\exp(\beta \log x_{ij})}{\sum_j \exp(\beta \log x_{ij})} \times y_{i+} \qquad (4)$$

Note that this model simplifies to the direct share method when $\beta = 1$. This model can also be derived from creating a share model from the log count model (2). By model construction, the shares within a county sum up to 100 percent, thus no adjustment will need to be made. The log likelihood function for this model, conditioning on the number of poor children in the county, assume that the shares are the multinomial probabilities $g_{ij}(x)$

$$\begin{aligned} \text{loglike} &= \sum_i \sum_j y_{ij} \log g_{ij}(x) \\ &= \sum_i \sum_j y_{ij} \log \frac{\exp(\beta \log x_{ij})}{\sum_j \exp(\beta \log x_{ij})}. \end{aligned}$$

This model can easily be extended to include additional covariates. One feature that this model (4) lacks is an intercept term in the $\exp(\cdot)$ part. An intercept would factor out and cancel with the same term in the denominator, so it would be unidentifiable. The parameter β is estimated by maximum likelihood. For Census 2000 and income year 1999 income tax data, the estimate for β was $\hat{\beta} = 1.01745$ with a standard error of .0454. Note that $\beta = 1$ is within the 95% confidence interval for this model, suggesting that the direct share method is a plausible model given the class of models under (4).

The next three models use the census share, or a transformation of the share, as the response variable. These models do not have the constraint that the predicted shares should add up to 100 percent. Therefore, we will need to ratio adjust the shares after the predictions have been made. The three responses to be modeled are the linear share, log share and logistic share. The linear and log share model both have the potential for having estimated shares outside of the 0 to 1 range, while the logistic share model will keep all of the predicted shares within the valid range. Let y_{sij} be the share of census number of poor children for school district piece j within county i. The linear, log and logistic share models are:

$$\begin{aligned} y_{sij} &= \beta_{a0} + \beta_{a1} x_{sij} + m_{aij} + e_{aij} \\ \log y_{sij} &= \beta_{b0} + \beta_{b1} \log x_{sij} + m_{bij} + e_{bij} \\ \text{logit}(y_{sij}) &= \beta_{c0} + \beta_{c1} \text{logit}(x_{sij}) + m_{cij} + e_{cij} \end{aligned}$$

where the m_{ij}'s are the model error terms and the e_{ij}'s are the appropriate sampling error terms whose variances are obtained by Taylor series linearlization (Wolter, 1985). The results of the regressions with standard errors in parentheses are

Model	β_0	β_1	σ_m^2	R^2
linear	.0001	.9666	.0001	.9142
	(.0001)	(.0022)		
log	-.1034	.9304	.116	.9116
	(.0069)	(.0022)		
logistic	-.1189	.9344	.162	.9039
	(.0075)	(.0023)		

When taking the census sampling variance into account, the linear share model is very close to the direct share estimator. The models for the log and logistic transformations of the share show a flattening effect, $\beta_1 < 1$ shrinking towards equal shares for each piece which results when $\beta_1 = 0$.

To construct the whole school district estimate under any of the models, we multiply the estimated share, $g_{ij}(x)$ by the county estimated number of poor children and sum over the school district pieces that composes school district k:

$$\hat{y}_k = \sum_{(i,j) \in SD_k} g_{ij}(x) y_{i+}.$$

4 Model Evaluations

In this section we compare the predictions of number of poor children in school districts, $\hat{y}_k$, from the models to the 2000 census long form estimates, y_{ij}. In order to compare school districts of vastly different sizes ranging from under 20 to near 1 million, we will compare the differences between the log of the model estimate and the log of the census estimate. In addition to greatly reducing the size effects on the error structure, the differences in logs can be loosely interpreted as percent errors, $(\hat{y}_k - y_{\text{cen},k})/y_{\text{cen},k}$, when the differences are small (Bell and Kramer 1999).

In our MSE evaluation, there are 3 sources of error that we should account for in our analysis:

1. census long form sampling error
2. allocation error for IRS non-geocoded exemptions
3. model error.

Knowing the relative sizes of these errors can tell us the impact of the census long form sampling error and IRS allocation error compared to the modeling error. For models that do not automatically have their shares sum up to 100 percent (linear, log and logistic shares), we will evaluate the raked (ratio adjusted) version of the estimates. Our goal is to provide estimates for whole school districts, so that will be the unit of analysis for the MSE evaluation. An MSE evaluation at the school district piece level would be dominated by the numerous extremely small pieces which have a huge relative error. By using whole school districts, many of these extremely small pieces will be merged back with larger pieces, damping their large relative errors.

To understand how the three sources of error form the MSE, we will decompose the error into its component parts. Let $y_{ij} = y_{ij}^T + \epsilon_{yij}$ be the estimate from the census long form which is measured with sampling error, ϵ_{yij} . Similarly, let $x_{ij} = x_{ij}^T + \epsilon_{xij}$ be the IRS poor child tax exemptions for school district piece (i,j) (derived from Section 3) where x_{ij}^T is the true underlying number of poor child tax exemptions for the school district piece. We assume that the process in which we allocate the non-geocoded tax exemptions is an unbiased estimate of the true number of non-geocoded exemptions that should have been assigned to a particular school district piece. Let n_k be the total number of whole school districts. Using a first order Taylor series approximation:

$$\begin{aligned} MSE &\equiv \frac{1}{n_k} \sum_k (\log y_k - \log \hat{y}_k)^2 \\ &= \frac{1}{n_k} \sum_k (\log \sum_{(i,j) \in SD_k} y_{ij} - \log \sum_{(i,j) \in SD_k} g_{ij}(x) y_{i+})^2 \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n_k}\sum_k (\log \sum_{(i,j)\in SD_k} (y_{ij}+\epsilon_{yij}) \\
&\quad - \log \sum_{(i,j)\in SD_k} g_{ij}(x^T+\epsilon_x)(y^T_{i+}+\epsilon_{yi+}))^2 \\
&\approx E\left[\log \sum_{(i,j)\in SD_k} y_{ij} - \log \sum_{(i,j)\in SD_k} g_{ij}(x) y_{i+}\right]^2 \qquad (5) \\
&\quad + f_1(\text{census sampling variance}) \\
&\quad + f_2(\text{non-geocoding allocation variance})
\end{aligned}
$$

where $f_1(\cdot)$ is the following function depending on all of the census sampling error variances for the school district pieces:

$$
\begin{aligned}
&f_1(\text{census sampling variance}) \\
&= \frac{\sum_{(i,j)\in SD_k} \sigma^2_{yij}}{\left[\sum_{(i,j)\in SD_k} y^T_{ij}\right]^2} \\
&+ \frac{\sum_{(i,j)\in SD_k} [g_{ij}(x^T)]^2 \sigma^2_{yi+}}{\left[\sum_{(i,j)\in SD_k} g_{ij}(x^T) y^T_{i+}\right]^2} \\
&-2 \times \frac{\sum_{(i,j)\in SD_k} g_{ij}(x^T)\sigma^2_{yij}}{\sum_{(i,j)\in SD_k} y^T_{ij} \sum_{(i,j)\in SD_k} g_{ij}(x^T) y^T_{i+}}
\end{aligned}
$$

add $f_2(\cdot)$ is the following function depending on the variances of the allocation of non-geocoded tax exemptions to school district pieces:

$$
\begin{aligned}
&f_2(\text{non-geocoding allocation variance}) \\
&= \frac{\sum_{(i,j)\in SD_k} \sigma^2_{xij}[g'_{ij}(x^T) y^T_{i+}]^2}{\left[\sum_{(i,j)\in SD_k} g_{ij}(x^T) y^T_{i+}\right]^2} \quad .
\end{aligned}
$$

MSE Results

Model	Mean Sq. Error	Census var	Model pred err.	CV
log count	.389	.198	.181	42.5%
direct	.358	.195	.152	39.1%
exp wt	.355	.201	.144	38.0%
linear	.363	.194	.159	39.9%
log	.383	.194	.179	42.3%
logistic	.383	.194	.179	42.4%

The expectation in (5) represents the error in the model predictions of the true log number of poor children if there were no allocation of non-geocoded tax exemptions. This measure takes out the contribution of census sampling error from the MSE. The MSE contribution due to the allocation of the non-geocoded tax exemptions was .009 for all models which accounted for 2.3% to 2.5% (not shown in table) of the total MSE in the models. One criteria for selecting a "best" model is to chose the model which has the lowest model prediction error variance.

In the comparison of models, we included the log count model with the share models discussed in Section 3. For the log count model, shares were created using the model estimates. Comparing model prediction error variances and CVs, we see that the exponential weighted share performs better than the other share models, although the direct share and linear share model are also good candidates. Also, the models for the non linear transformation of the shares have larger CVs. The census sampling error variance accounts for 50% to 55% of the total MSE between the census long form values and model estimates. Thus, the census sampling error variance is a large component of the overall error structure. The high variance on the smaller school districts makes it difficult to develop a good predictive model for them.

5 Discussion

We have presented a variety of models using the number of poor child tax exemptions in estimating the number of poor children in each school district given by the Census 2000 long form questionnaire. Clearly the IRS tax data for income year 1999 has shown, even through just one variable, to be very informative about the census long form child poverty. One would assume that in non census years that the IRS data would also be informative about that particular year's child poverty distribution within counties. If we can assume that the relationship between true census poverty shares and IRS child poverty shares is stable over time, then the IRS data could be a great benefit in the estimation of poor children in school districts between censuses. The main advantage to using the IRS tax data is that the information in the tax data is more timely than the last census. As we move away from the census year, the census data becomes more out of date as the distribution of poor children may change.

There are still problems and issues that need to be addressed when using the IRS data. First, some counties have a large percentage of non-geocoded tax exemptions. By distributing those exemptions proportionally by child population counts, we are making several assumptions about the geocoding process. Mainly, we assume that all child exemptions in the same county have an equal probability of not being geocoded. Also, the variable used to proportionally allocate the non-geocoded exemptions comes from the census and rather should be based on population figures that are updated yearly. Currently there is a parallel project to impove school age child population figures for school district pieces (Oosse 2004). Another potential problem is the implicit assumption about the distribution of age for poor child exemptions. By multiplying the age range of a school district piece to the count of poor child exemptions, we are assuming that the ages of the children represented in those exemptions are equally allocated between the ages of 0 and 17. Distributions of single year age at the county level could be used to better adjust the number of poor child exemptions to match the appropriate age/grade range for a school district piece.

Using only from Census 2000, we cannot evaluate the quality of the models providing updated estimates. There is work in progress to tabulate the 1990 Census and IRS income tax data for income year 1989 to the school district boundaries of 2000. In this setting we can estimate our models from 1990 Census and 1989 IRS data to create a predictive model and carry it forward using IRS income tax data from income year 1999 to estimate the number of poor school age children in 1999. We can then compare to Census 2000 results to check accuracy for both new models and old update scheme.

The models presented here are rather simple, yet highly predictive. Additional variables from both IRS tax data and other sources should be considered to reduce the model error variance. Variables such as total number of child exemptions and total exemptions on tax returns with adjusted gross income below the poverty threshold are currently available in the IRS tax dataset. Variables collected on whole school districts such as school enrollment and free/reduced lunch participation may be useful in these models, but at this time it is unclear how to appropriately incorporate them into a school district piece model. Finally, as the American Community Survey goes to full implementation, the data it produces could be a rich source of information about child poverty at the school district level. As this data becomes available in the future, it should be evaluated for its potential in modeling poor children at the school district level.

References

Bell, William R. and Kramer, Matthew (1999). "On School District Error Decomposition (7/8/99 draft)," unpublished report, Statistical Research Division, U.S. Census Bureau.

National Research Council (2000). *Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology.* Panel on Estimates of Poverty for Small Geographic Areas, Constance F. Citro and Graham Kalton, editors. Committee on National Statistics. Washington, D.C.: National Academy Press.

Oosse, Monique (2004) "School District Population Estimates Methodology Research Using Administrative Data Sources: Public School Enrollment," *2004 Proceedings of the American Statistical Association*, Section on Government Statistics [CD-ROM], Toronto, Canada: American Statistical Association.

U.S. Census Bureau (2004). "Small Area and Income Estimates - School District Estimates", http://www.census.gov/hhes/www/saipe/index.html, accessed on May 3, 2004.

Wolter, K. (1985). **Introduction to Variance Estimation**, Springer-Verlag.