School District Population Estimates Methodology Research Using Administrative Data Sources: Public School Enrollment[1]

Monique Oosse

The Population Division of the U.S. Census Bureau produces annual estimates of the total and school-age populations in school districts in compliance with Title I of the Elementary and Secondary Education Act of 1994 and, more recently, the No Child Left Behind Act of 2001. The population estimates are combined with estimates of the numbers of children in poverty[2] to allocate federal funds to school districts (NRC 2000). The Population Division is testing alternative data sources and developing new methods for producing school district population estimates in order to address some of the weaknesses of the current methodology and improve the accuracy of the estimates. This paper describes the initial efforts to apply public school enrollment data from the National Center for Education Statistics (NCES) Common Core of Data (CCD) to school district population estimates.

I. Background

School district geography presents some special challenges for the production of school district population estimates. First, there are up to three types of school districts in some states: Elementary (ESD), Secondary (SSD), and Unified (USD) School Districts. Seventeen states contain overlapping school districts where more than one school district covers the same area because each school district serves only a portion of the kindergarten through twelfth grade range. For these school districts, the census populations of school-age children (ages 5 to 17 years) are assigned to one of the overlapping school districts based on the modal grade for each age and the grades served by each school[3]. This approach constrains the numbers of school-age children in school districts to be consistent with the state and county population of school-age children in the census data and in subsequent estimates. School-age children assigned to overlapping districts in this manner and children in non-overlapping school districts are referred to as "relevant children." The logic of assigning people to a particular grade range does not apply to the non-school-age population. Therefore, the total population for overlapping school districts is counted more than once and is not consistent with total population figures for other geographic levels.

Second, fifteen states have areas that are not served by school districts (as enumerated in Census 2000). Population estimates for these "balance of county" areas must be produced to ensure consistency across estimates products. However, the public school enrollment data applied in this research do not contain information about the areas not covered by school districts, by definition, so other approaches are needed.

The 1996, 1998, 2001, 2002, and 2003 school district population estimates were produced using a synthetic ratio or shares method. The synthetic ratio method assumes that the ratio of the school district population to the county population in the most recent decennial census remains constant over time.

The formula for the April 1, 2000 estimates produced using the synthetic ratio method is:

$$P'_{sd\text{-}ct,\ 2000} = (P_{sd\text{-}ct,\ 1990} / P_{ct,\ 1990}) * P'_{ct,\ 2000}$$

Where:

$P'_{sd\text{-}ct,\ 2000}$ = April 1, 2000 school district-county part[4] population estimate (total or school-age)

$P_{sd\text{-}ct,\ 1990}$ = 1990 census school district-county part population (total or school-age)

$P_{ct,\ 1990}$ = 1990 census county population (total or school-age) for the county in which the school district part lies

$P'_{ct,\ 2000}$ = April 1, 2000 county population estimate (total or school-age) for the county in which the school district part lies

The assumption that a school district's share of the county population does not change over time fails to account for differential change among school districts or portions of school districts within counties. Two Population Division evaluations of the synthetic ratio method show that most of the inaccuracy in the school district population estimates is due to this assumption. Errors in the county-level estimates have relatively small effects on the accuracy of the school district population estimates (Miller 2001; NRC 2000; Oosse 2004). Given these findings, the research to improve the school district population estimates focuses on ways to measure and estimate different levels of population changes within counties.

The evaluations of the synthetic ratio method also found that the largest percent errors in the population estimates are associated with the smallest school districts (populations under 5,000) and those with population changes of ten percent or more over a ten-year period (Miller 2001; NRC 2000; Oosse 2004). The alternative methods developed for school district population estimates also aim to reduce the different levels of accuracy, or bias, associated with school district size, population change, and other characteristics.

II. Data

A. 1990 Census

As with other Census Bureau population estimates products, the school district population estimates must use the enumerated Census population as the base or starting population. For the research presented here, the 1990 census with 1989-90 school district boundaries is the base population. Some adjustments are made to a few of the 1990 census school district records in order to

---

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

Acknowledgements: The author thanks Jerry Maples and Lynn Weidman, both of the U.S. Census Bureau, for their valuable comments on this report.

[2]The poverty estimates are produced by the Small Area Income and Poverty Estimates (SAIPE) Program in the Housing and Household Economic Statistics Division.

[3]For example, an ESD serves kindergarten through 8th grade and an SSD serves 9th through 12th grades. The modal grade (i.e. most 9 year olds are in 4th grade) assigns children ages 5 to 13 to the ESD; and children ages 14 to 17 to the SSD (U.S. Census Bureau 2004).

[4]"School district-county parts" or "school district parts" refer to each portion of a school district in a different county, for districts that cross county boundaries, and school districts contained in only one county.

match the CCD and Census 2000 records more closely. Only the 1990 census populations tabulated for the "modified" 1990 census geography are used for the estimates described in this paper.

The school district and county records are combined to create separate population counts for each portion of a school district in a different county ("school district-county parts") in the 1990 census. This allows the school district population estimates to maintain consistency with county population estimates. Population counts for the areas not covered by school districts are tabulated by counties. The assignment of relevant children to overlapping school districts in the 1990 census data is based on age-to-grade distributions[5] in a different process than the modal grade assignment described above. These differences are reflected in the subsequent estimates and may contribute to the inaccuracy of the estimates when compared with the Census 2000 population.

B. April 1, 2000 County Population Estimates

All the school district population estimates presented in this paper are constrained to equal the April 1, 2000 county population estimates in order to maintain consistency across different levels of geography, as dictated by sound demographic methodology and Census Bureau policies. Conditioning on the estimates for higher levels of geography also helps produce more accurate estimates for smaller areas, like most school districts and school district-county parts. The April 1, 2000 county population estimates were produced to evaluate the accuracy of the county estimates methodology for the 1990s. The county population estimates were produced using a cohort-component method that:

1) starts with the 1990 census county population,
2) adds the annual numbers of births for each county,
3) adds estimates of annual net internal migration for each county,
4) adds estimates of annual net international migration for each county, and
5) subtracts the annual numbers of death for each county.

These components (births, deaths, and migration) are calculated from Internal Revenue Service, Immigration and Naturalization Service, Social Security Administration, and birth and death registration (vital statistics) records for the 10-year time period since the 1990 census. County population estimates for single years of age (0 to 85), sex, race, and Hispanic origin also were produced and tabulated for ages 5 to 17 to provide county estimates of the numbers of school-age children.

C. Common Core of Data: Public School Enrollment

This paper describes and evaluates new methods for producing school district population estimates using public school enrollment data from the Common Core of Data (CCD) collected by NCES. CCD public school enrollment data are appropriate for the school district population estimates for several reasons. First, it is one of the few sources of data that is available at the school district level. The data cover "virtually 100 percent" of school districts and public schools (NCES 2001, p. 3). Second, the data are collected systematically from State Education Agencies by the Department of Education and are imputed when not available or implausible. Third, the data are collected annually and are provided to the public in a timely manner, less than two years following the October 1 reference date. CCD survey data for the 1986-87 through 2002-03 school years can be obtained from the NCES website (nces.ed.gov). Finally, these data will continue to be available in the foreseeable future. Though the CCD survey provides a lot of information about schools and school districts, the work presented in this paper uses only the school district total enrollment for 1989, 1990, 1999, and 2000[6]. Coverage analysis for the CCD data shows that at least three fourths of school-age children enumerated in the 1990 census and Census 2000 are represented in the school enrollment totals.

The use of public school enrollment data for school district population estimates has some limitations. First, school enrollment may not correspond to the place of residence and the frequency with which this occurs is not known. Second, kindergarten through twelfth grade enrollment does not include all children ages 5 to 17 and may include children of other ages. Public school enrollment rates may differ by age, grade, race, Hispanic origin, and geographic location. Finally, public school enrollment and changes in enrollment reflect the combined effects of cohort aging, internal and international migration, grade progression, student retention [related to private school enrollment (about 10 percent nationwide), home school participation (about 2 percent nationwide), charter school participation, and drop-out rates], as well as reporting errors and redistricting. For these reasons, the relationship between school enrollment for elementary ages and the total population became less stable during the 1980s, resulting in a change to the estimates method that once used school enrollment data for county migration rates (Long 1993). With all these factors, differences or changes in school enrollment may or may not correspond well to population movement and change.

In order to match the Census 2000 data more closely, some changes were made to the 1989 and 1990 CCD data to reflect fourteen future school district consolidations in 8 states.

III. Evaluation of Experimental Estimates

The Census 2000 enumerated population is the standard for determining the accuracy of the experimental April 1, 2000 estimates. The Census 2000 population data are tabulated for the 1999-2000 school district boundaries in effect during the time the Census was taken, but include modifications as documented in the Count Question Resolution[7] program. Also, in order to match the CCD and 1990 census data more closely, some changes were made to the Census 2000 data to reflect school district consolidations. Only the Census 2000 populations tabulated for the "modified" Census 2000 geography are used to evaluate the experimental estimates.

Accuracy is measured by the Mean Absolute Percent Error (MAPE). The MAPE is the sum of the absolute percent differences between each school district experimental estimate and the corresponding enumerated Census 2000 population divided by the number of school districts. Written as a formula:

$$MAPE = [\sum 100 * | (P'_{sd,\,2000} - P_{sd,\,2000}) | / P_{sd,\,2000}] / N$$

Where:

$P'_{sd,\,2000}$ = April 1, 2000 school district population estimate (total

[5]The assignment of relevant children in the 1990 census data is based on the age-to-grade distributions from the 1988, 1989, and 1990 Current Population Surveys and sample data from the 1990 census (Miller 2001).

[6]CCD data are referenced by the starting school year in this paper (i.e. the 1989-90 school year is written as 1989).

[7]The official program to correct Census 2000 enumeration errors.

or school-age)

$P_{sd, 2000}$ = Census 2000 school district population (total or school-age)

N = Number of school districts (all districts or those with a particular characteristic)

The MAPE measures the overall accuracy of the estimates, but not the direction of the errors. Other measures of accuracy, such as the mean algebraic percent error and the root mean square error, yield the same conclusions and are not reported in this paper.

MAPEs are computed only for school districts with total or school-age populations of 30 or more in Census 2000 because population estimates for smaller numbers are particularly unreliable. About 1,400 school districts (9.3 percent) have less than 30 people and are excluded from the total population estimates MAPEs. About 1,800 school districts (11.7 percent) have less than 30 school-age children in Census 2000 and are excluded from the school-age population estimates MAPEs. Also excluded are an additional 15 districts for the total population estimates MAPEs and 17 districts for the school-age population estimates MAPEs because the districts have percent errors outside the range of -90 to 500 percent. These districts are excluded because the extremely large values impact the mean errors so much that the statistics lose meaning when comparing methods.

At best, school district population estimates can be expected to be as accurate as the county population estimates to which the school district estimates are constrained. Comparisons of the county population estimates produced for April 1, 2000 with the Census 2000 enumeration results report a MAPE of 3.8 percent for the total population, ranging from 1.5 percent for Ohio to 16.3 percent for Hawaii (Blumerman and Christenson 2002). With errors introduced by both the county population estimates (to which all school district population estimates are constrained) and the school district estimates methodology, a practical target MAPE for the total population estimates is about seven percent. Estimates of the school-age population tend to be less accurate and have higher MAPEs because it is more difficult to estimate smaller populations and populations with particular characteristics, such as age. The MAPE is about 7.5 percent for the county population estimates of the population ages 5 to 17. The target MAPE for the school-age population is about ten percent. These targets represent about a 30 percent improvement in the accuracy of the population estimates created using the synthetic ratio method, as measured in this paper.

April 1, 2000 school district population estimates are created using the synthetic ratio method to provide a baseline to evaluate the experimental estimates. The accuracy of the synthetic ratio estimates differs from previous evaluations because the 1990 census base file contains population counts for 1989-90 school district boundaries and the Census 2000 evaluation data are in 1999-2000 school district boundaries. These differences are retained deliberately because the 1989 and 1990 school enrollment data correspond to the 1990 census geography, but this introduces additional error into the estimates compared with the previous evaluations that account for school district boundary changes (Miller 2001, Oosse 2004).

Accuracy is evaluated for all school districts and by several characteristics of school districts:

1) type (ESD, SSD, or USD)
2) percent of school-age children in poverty from the 1990 census
3) population size in the 1990 census
4) ten-year intercensal (1990-2000) percent population change
5) urban-rural location, seven categories as classified by NCES
6) location in counties with and without overlapping districts

It is important for the new methods to produce accurate estimates regardless of school district type, size, location, or poverty rate and to improve upon the disparities in accuracy generated by the current method of producing school district population estimates. Evaluating the school district estimates separately for counties that contain overlapping school districts indicates the amount of error attributable to the assignment of relevant children. However, relevant children are assigned to overlapping school districts using different approaches for the 1990 census base population (age-to-grade distributions, see footnote 5) and the Census 2000 data used to measure accuracy (modal grade, see footnote 3). This is an additional source of error in all sets of estimates.

IV. Experimental Methods

Three different methods were developed to produce experimental school district population estimates using the school enrollment data. It was expected that the school enrollment data would reflect the size of and change in the school-age population more accurately than the total population. However, it is possible that school enrollment serves as an indicator of changes in the total population of school districts, so the public school enrollment data are used to produce both school-age and total population estimates.

A. School District Enrollment Shares

The first set of experimental estimates distributes April 1, 2000 county population estimates among school district-county parts according to the average 1999-2000 school enrollment[8]. This approach is a shares method similar to the synthetic ratio method, except that the more current data from the 1999-2000 CCD are used to distribute the April 1, 2000 county estimates in place of the 1990 census data. First, for school districts that cross county lines, the average school enrollment counts are distributed between the school district parts in each county according to the proportions of the school-age population in the 1990 census. Then the April 1, 2000 county population estimates of the total and school-age populations are multiplied by the school district-county part proportion of the total 1999-2000 school enrollment for the county. Finally, the school district-county part estimates are summed by school districts.

The April 1, 2000 school district population estimates produced using the school district enrollment shares are calculated in two main steps:

Step 1: $E_{sd\text{-}ct,\ 1999\text{-}2000} = (P_{sd\text{-}ct,\ 1990} / P_{sd,\ 1990}) * E_{sd,\ 1999\text{-}2000}$

Step 2: $P'_{sd\text{-}ct,\ 2000} = (E_{sd\text{-}ct,\ 1999\text{-}2000} / E_{ct,\ 1999\text{-}2000}) * P'_{ct,\ 2000}$

Where:

$E_{sd\text{-}ct,\ 1999\text{-}2000}$ = 1999-2000 school enrollment allocated to school district-county parts

$P_{sd\text{-}ct,\ 1990}$ = 1990 census school district part population ages 5 to 17

[8]Averages of school enrollment data for 1989 and 1990 or 1999 and 2000 are used throughout this paper to generate more stable values. Also, school enrollment is reported as of October 1 for each year, so averaging the data for the two surrounding years more closely corresponds to the April 1 census reference date.

$P_{sd,\ 1990}$ = 1990 census school district population ages 5 to 17
$E_{sd,\ 1999\text{-}2000}$ = total 1999-2000 school enrollment for the school district associated with the school district-county part
$P'_{sd\text{-}ct,\ 2000}$ = April 1, 2000 school district-county part population estimate (total or school-age)
$E_{ct,\ 1999\text{-}2000}$ = 1999-2000 school enrollment allocated to school district-county parts (Esd-ct, 1999-2000), summed by county
$P'_{ct,\ 2000}$ = April 1, 2000 county population estimate (total or school-age) for the county in which the school district part lies

As there are no school enrollment records for areas not covered by school districts ("balance of county"), the ratio of the balance of county population to the county population in the 1990 census is used to allocate the April 1, 2000 county population estimates to these areas, as in the synthetic ratio method.

Table 1 shows the MAPEs for the experimental estimates of the total population created with the school district enrollment shares method (column 3) and the baseline estimates created using the synthetic ratio method (column 2). The school enrollment shares method generally does not perform as well for the total population estimates as the synthetic ratio method. The MAPE for the school enrollment shares estimates is 20.6 percent, more than twice the MAPE for the synthetic ratio estimates, 9.7 percent. The MAPEs are considerably worse for the experimental estimates for many of the categories of school districts, such as SSDs (57.3 percent for the experimental method and 9.2 percent for the synthetic ratio method) and districts with less than 5,000 people in the 1990 census (26.7 percent for the experimental method and 12.5 percent for the synthetic ratio method).

In contrast, Table 2 shows that the school enrollment shares estimates for the school-age population (column 3) are slightly more accurate than the estimates produced using the synthetic ratio method (column 2) (MAPEs of 14.3 percent and 14.6 percent, respectively). The MAPEs also are lower for the experimental method than the synthetic ratio method for most categories of school districts. For example, the MAPEs for school districts with less than 10 percent of children in poverty are 14.9 percent for the synthetic ratio method estimates and 13.4 percent for the school enrollment shares method. The average errors of the enrollment shares estimates for SSDs and for school districts that serve large cities are 44 and 28 percent more accurate, respectively, than the corresponding estimates produced using the synthetic ratio method. These are the only categories of school districts for which the experimental estimates achieve or nearly achieve the 30 percent target level of improvement.

As noted above, the assignment of relevant children may introduce additional error to the school-age population estimates. The school enrollment shares method reduces the average error for both overlapping and non-overlapping school districts. Perhaps more importantly, the experimental method reduces the difference in MAPEs between the overlapping and non-overlapping school districts by 1.6 percentage points.

Though this experimental method offers some improvement over the current method for producing school-age population estimates, it does not take advantage of the information provided by the 1990 census data. The experimental methods below apply the 1990 census data as well as the school enrollment data to produce more accurate school district estimates of the population ages 5 to 17.

B. Percent Change in School Enrollment

For the second set of experimental estimates, the percent change in the total school enrollment from 1989-1990 to 1999-2000 is applied to the corresponding school district-county part population from the 1990 census to produce preliminary total and school-age population estimates for April 1, 2000.

The formula for the preliminary population estimates is:

$$P'_{sd\text{-}ct,\ 2000} = P_{sd\text{-}ct,\ 1990} + [P_{sd\text{-}ct,\ 1990} * (E_{1999\text{-}2000} - E_{1989\text{-}1990})/E_{1989\text{-}1990}]$$

Where:
$P'_{sd\text{-}ct,\ 2000}$ = April 1, 2000 school district-county part population estimate (total or school-age)
$P_{sd\text{-}ct,\ 1990}$ = 1990 census school district-county part population (total or school-age)
$E_{1999\text{-}2000}$ = total 1999-2000 school enrollment
$E_{1989\text{-}1990}$ = total 1989-1990 school enrollment

The preliminary estimates are constrained proportionally to sum to the April 1, 2000 county estimates for the total population and the population ages 5 to 17. Constraining the total population estimates to the county population estimates for overlapping school districts is performed by multiplying the April 1, 2000 county estimates by the ratio of the sum of the 1990 census school district-county part total population (counted more than once for overlapping districts) to the independent county total population from the 1990 census. As there are no school enrollment records for balance of county areas, the ratio of the balance of county population to the county population in the 1990 census is used to allocate the April 1, 2000 county population estimates to these areas, as in the synthetic ratio method.

Table 1 shows the MAPEs for the total population for the experimental estimates created using the percent change in school enrollment (column 4). The second set of experimental method estimates have an average error of 12.6 percent for all school districts, which is less accurate than the synthetic ratio method estimates MAPE of 9.7 percent. The MAPEs for the synthetic ratio estimates of the total population also are more accurate than the experimental estimates for most categories of school district characteristics. However, the experimental estimates are more accurate than the synthetic ratio method for school districts with 20,000 people or more and for school districts located in large and mid-size cities. The synthetic ratio method also performs well for the larger school districts, so the improvement offered by the experimental estimates is relatively small.

The MAPEs for the experimental estimates created using the percent change in public school enrollment are much higher than the MAPEs for the synthetic ratio estimates for ESDs, SSDs, and overlapping school districts. These are the school districts for which it is more difficult to create accurate estimates, regardless of the method. These results indicate that changes in school enrollment, particularly for school districts that serve only part of the kindergarten through twelfth grade range, do not represent changes in the total population very well.

Table 2 shows the MAPEs for the school-age population for the experimental estimates based on the percent change in school enrollment (column 4). For all school districts, the MAPE for the experimental estimates is 13.3 percent, more accurate than the MAPE for the synthetic ratio estimates, 14.6 percent (column 2) and the MAPE for the first set of experimental estimates, 14.3 percent (column 3). The experimental estimates also are more

accurate than the traditional method for most characteristics of school districts. The largest improvements are for the SSDs, school districts with populations of 20,000 or more, and school districts in large cities or urban fringes of large cities. For these categories, the MAPEs for the experimental estimates are at least 25 percent lower than the MAPEs for the synthetic ratio method estimates, achieving or nearly achieving the 30 percent criteria for substantial improvement over the current method. However, the improvements are relatively small for the school district categories that are hardest to estimate. For example, the traditional estimate MAPE for school districts in rural areas is 16.8 percent and the MAPE for the experimental estimates is only 0.6 percentage points lower, 16.2 percent.

Both the experimental estimates and the synthetic ratio method estimates are less accurate for school districts in counties with overlapping districts compared with school districts in counties with no overlapping school districts. For the second set of experimental estimates, the MAPE is 16.8 percent for overlapping districts and 10.6 percent for non-overlapping districts. Similarly, the mean errors for ESDs and SSDs are almost twice as high as the mean errors for USDs. This shows that the assignment of relevant children contributes to the inaccuracies of the estimates, regardless of the estimate method.

C. Regression Models

The third set of experimental estimates uses school enrollment and 1990 census data as independent variables in regression models. The Census 2000 enumerated school-age and total populations are the dependent variables. The natural log of the variables generally produces more accurate estimates than the unadjusted forms of the variables, particularly for smaller school districts. Several models were tested using combinations of school enrollment and census data as independent variables and three of the most accurate models are discussed below.

For example, the regression coefficients are applied to the 1990 census and school enrollment data to produce the preliminary estimates of the population ages 5 to 17 as follows:

$$Ln(P'_{sd,\,2000}) = \beta_0 + \beta_1 * Ln(P_{sd,\,1990}) + \beta_2 * Ln(E_{1999})$$

Where:

$Ln(P'_{sd,\,2000})$ = natural log of the April 1, 2000 school district population estimate (total or school-age)

$Ln(P_{sd,\,1990})$ = natural log of the 1990 census school district population (total or school-age)

$Ln(E_{1999})$ = natural log of the 1999-2000 school district enrollment

$\beta_0 = 0.14410$, the intercept value fit for the regression model

$\beta_1 = 0.42582$, the regression coefficient calculated for $Ln(P_{sd,\,1990})$

$\beta_2 = 0.56942$, the regression coefficient calculated for $Ln(E_{1999})$

The preliminary estimates for each model are controlled proportionally to the April 1, 2000 county population estimates. As there are no school enrollment records for balance of county areas, the ratio of the balance of county population to the county population in the 1990 census is used to allocate the April 1, 2000 county population estimates for these areas, as in the synthetic ratio method.

For all the models tested, the regression coefficients and R-square values to measure the explained variation indicate that the 1990 census population has a very strong relationship with the Census 2000 population and the addition of school enrollment makes a much smaller contribution to the models. In fact, the estimates for the total population are more accurate for the regression model with only the log of the 1990 census total population as the independent variable. The MAPE for the regression model using only the 1990 census is 9.7 percent and the MAPE for the model using the 1990 census and 1999-2000 school enrollment is 11.2 percent (see Table 1, columns 5 and 6). The addition of the natural log of the 1999-2000 school enrollment increases the accuracy of the total population estimates very slightly for school districts in large and mid-size cities, districts with populations of 20,000 or more, and with 1990-2000 total population change of ten percent or more. The accuracy of the estimates decreases with the addition of school enrollment data for all other characteristics of school districts.

The MAPEs for the total population estimates created with the synthetic ratio method and the estimates created with the regression model using only the 1990 census are the same, 9.7 percent (Table 1, columns 2 and 5). The regression model estimates are better than the synthetic ratio estimates for ESDs, districts with at least 10 percent of children in poverty as measured in the 1990 census, districts with less than 5,000 people, districts with ten-year total population change of less than five percent, and districts located in rural areas or small towns. However, the differences are relatively small, a result of using the same data in both sets of estimates.

In contrast with the results for the total population, including the school enrollment data in the regression model produces estimates of the school-age population that are more accurate than the regression model estimates based only on the 1990 census data. For all school districts, the MAPE is 14.6 percent for the estimates produced from the regression model with the 1990 census school-age population as the only independent variable (Table 2, column 5) and the MAPE is 12.5 percent for the regression model estimates with the 1990 census school-age population and the 1999-2000 school enrollment data as the independent variables (Table 2, column 6). This improvement occurs for most characteristics of school districts and is quite large for some categories, such as districts with less than ten percent of children in poverty and school districts that experienced school-age population growth of ten percent or more during the 1990s .

The estimates created using the regression model with 1990 census and school enrollment data are at least 25 percent more accurate than the estimates based on the synthetic ratio method for SSDs, districts that serve large and mid-size cities, school districts with 20,000 people or more, and districts that experienced population growth of 10 percent or more. The estimates for these categories achieve or nearly achieve the target level of improvement over the current method. However, these are some of the school districts for which the current method produces more accurate estimates and the improvement in accuracy by the regression method results in greater disparity between the estimates for large and small districts and for districts in urban and rural areas. The improvement in the estimates for SSDs and overlapping school districts suggests that use of the school enrollment data mitigates some of the inaccuracy introduced by the assignment of relevant children to overlapping school districts.

In an attempt to produce better estimates for small school districts, separate regression coefficients are calculated for different sizes of school districts, as determined by the 1990 census total population. Estimates of the total and school-age population

also are produced based on separate regression coefficients for school district type, locale, and poverty. The best set of estimates, and the only set of estimates more accurate than those discussed above, are from the different regression models for the three school district types (Elementary, Secondary, and Unified).

The overall MAPE of 10.8 percent for the regression model estimates of the total population is still higher than the MAPE of 9.7 percent for the synthetic ratio method estimates (see Table 1, columns 2 and 7). However, the regression model estimates that apply different coefficients for the three school district types are slightly more accurate than the synthetic ratio method estimates for school districts with 20,000 people or more, districts with population growth of 10 percent or more, and districts that serve large and mid-size cities or urban fringes of mid-size cities.

The estimates of the school-age population calculated from regression models with different coefficients for each school district type have a MAPE of 12.4 percent (Table 2, column 7), slightly better than the MAPE of 12.5 percent for the regression model estimates with the same set of coefficients for all school districts (Table 2, column 6). The improvement occurs for most categories of school districts and is largest for ESDs, SSDs, school districts that experienced population decline of five percent or more from 1990 to 2000, and school districts that serve urban fringes of large cities. Though the overall improvement is minimal, calculating different regression coefficients for the three school district types may reduce the amount of bias in the school-age population estimates. For example, the regression model estimates reduce the range of MAPEs between overlapping and non-overlapping school districts and between USDs, ESDs, and SSDs compared with the synthetic ratio method estimates. Improving the accuracy of the school-age population estimates where relevant children are assigned to overlapping school districts is one of the keys to producing more accurate school-age population estimates.

V. Summary

This paper presents some of the experimental methods developed for using public school enrollment data to produce school district population estimates. None of the experimental methods consistently improve the accuracy of the total population estimates compared with the current method. Though all of the experimental methods offer some improvement in the overall accuracy of the school-age population estimates, they do not achieve the goal of reducing the mean errors by 30 percent. Furthermore, the MAPEs for the school-age population estimates for small school districts and districts with relatively large population changes are still considerably higher than the target MAPE of ten percent.

The most promising approach applies regression coefficients calculated from the 1990 census and 1999-2000 school district enrollment with the Census 2000 school district population as the dependent variable. There are some concerns, however, about using the regression model approach for the official Census Bureau school district population estimates in the future. First, the regression model requires the assumption that the relationship between census and school enrollment data for the current decade is the same as for the 1990s. Second, the mean errors for the regression model estimates are about three times higher at 30 percent for one state, Minnesota, than the synthetic ratio method estimates (results not shown). Introducing a method with the potential for such a large amount of error for just one state may unfairly affect the distribution of federal funds. Similarly, all three experimental methods reduce some of the bias for the school-age population estimates, but estimates for larger school districts and those with moderate population changes are still more accurate than estimates for smaller school districts and those with more extreme population changes. This also has the potential to adversely impact the distribution of federal funds. Lastly, any new method to produce school district population estimates must be evaluated and coordinated with changes in the methodology for producing school district poverty estimates to avoid artificially constructing relationships between the two estimates, particularly if both programs change to a regression model approach.

This preliminary research may be extended in several ways to improve the accuracy of school district population estimates. First, other CCD survey data, such as grade-specific enrollment, may be tested, particularly for improving the accuracy of the estimates for overlapping school districts. Second, the CCD data may be supplemented by private school enrollment data from the NCES Private School Survey to provide more complete coverage of the school-age population. Third, other administrative data sources, such as IRS tax return extracts, may be applied to school district population estimates, modeled after the county population estimates program. Future research to improve the accuracy of the school district population estimates may focus on regression models, as they may be the best approach for combining the information from different administrative data sources.

References

Blumerman, Lisa M. and Matthew Christenson. 2002. "The Population Estimates of the 1990s: Close to the Mark?" Paper presented at the Annual Meeting of the American Sociological Association, Chicago, Illinois, August 16-19, 2002.

Long, John. 1993. "Postcensal Population Estimates: States, Counties, and Places." Population Division Working Paper # 3.

Miller, Esther R. 2001. "Evaluation of the 1990 School District Level Population Estimates Based on the Synthetic Ratio Approach." Population Division Working Paper # 54.

National Center for Education Statistics. 2001. "Documentation to the NCES Common Core of Data Local Education Agency Universe Survey: School Year 2000-2001."

National Research Council. 2000. Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology. Washington, DC: National Academy Press.

Oosse, Monique. 2004. "Evaluation of April 1, 2000 School District Population Estimates Based on the Synthetic Ratio Method." Population Division Working Paper # 74.

U.S. Census Bureau. 2004. "2000 Overview of School District Estimates." Small Area Income and Poverty Estimates. <http://www.census.gov/hhes/www/saipe/school/sd00over.html >

Table 1. Mean Absolute Percent Errors (MAPE) for April 1, 2000 Estimates of the Total Population Created with the Synthetic Ratio Method and Five Experimental Methods

School District Characteristics	Number of School Districts* (col. 1)	Mean Absolute Percent Errors: Synthetic Ratio Method (col. 2)	School District Enrollment Shares, 1999-2000 (col. 3)	Percent Change in School Enrollment 1989-90 to 1999-2000 (col. 4)	Regression Models: Natural Log of 1990 Census Total Population (col. 5)	Regression Models: Natural Logs of 1990 Census Total Population and Average 1999-2000 School District Enrollment: One Model (col. 6)	Regression Models: Natural Logs of 1990 Census Total Population and Average 1999-2000 School District Enrollment: by School District Type (col.7)
Total population	14,030	9.7	20.6	12.6	9.7	11.2	10.8
School District Type: 1990 Census							
Unified School District	10,758	8.4	15.1	9.4	8.5	9.2	8.9
Elementary School District	2,703	14.9	35.3	23.4	14.4	17.3	16.8
Secondary School District	529	9.2	57.3	23.5	9.4	20.1	17.2
Balance of County	24	38.3	38.2	38.8	41.1	38.5	37.7
Percent of Children in Poverty: 1990 Census							
Under 10%	4,212	10.4	21.9	12.2	10.5	11.3	10.6
10% to 24.9%	6,746	8.9	19.9	12.2	8.8	10.9	10.5
25% or More	2,916	10.0	19.8	13.7	9.9	11.6	11.3
School District Total Population: 1990 Census							
Under 5,000	6,459	12.5	26.7	18.5	12.3	15.5	15.2
5,000 to 19,999	4,836	7.5	15.5	8.1	7.7	8.0	7.5
20,000 or more	2,735	7.1	15.4	6.9	7.3	7.2	6.5
Percent Change in Total Population, 1990-2000							
Decrease more than 10%	1,007	28.4	32.1	33.1	25.0	29.8	28.3
Decrease 5.1% to 10%	999	9.0	22.3	13.9	8.0	14.4	13.8
Decrease up to 5% to Increase up to 4.9%	4,002	5.2	19.0	10.0	5.0	8.5	8.3
Increase 5% to 9.9%	2,048	5.0	18.7	9.4	5.3	6.9	6.8
Increase of 10% or more	5,974	11.4	20.1	11.9	12.2	11.0	10.5
Urban-Rural Location: 1989 CCD Data							
Large City	152	10.0	18.2	9.3	10.6	10.2	9.4
Mid-Size City	684	9.8	17.5	9.4	10.0	9.5	8.9
Urban Fringe, Large City	1,242	8.1	23.1	9.5	8.3	8.9	8.2
Urban Fringe, Mid-Size City	863	8.9	17.6	9.9	8.9	8.9	8.1
Large Town	208	7.8	18.3	8.9	7.8	8.5	8.2
Small Town	4,106	8.6	18.3	10.7	8.5	9.6	9.1
Rural, Outside MSA	6,617	10.6	22.2	15.1	10.5	13.1	12.8
School District Geography: 1990 Census							
Overlapping	6,285	11.9	27.9	17.2	11.7	14.8	13.7
Not Overlapping	7,745	8.0	14.7	8.9	8.1	8.4	8.4

* The numbers shown are the number of school districts included in the MAPE calculations for the synthetic ratio method estimates. These numbers may differ slightly for the experimental estimates because of the criteria for excluding outliers.

Table 2. Mean Absolute Percent Errors (MAPE) for April 1, 2000 Estimates of the School-Age Population Created with the Synthetic Ratio Method and Five Experimental Methods

School District Characteristics	Number of School Districts* (col. 1)	Mean Absolute Percent Errors: Synthetic Ratio Method (col. 2)	School District Enrollment Shares, 1999-2000 (col. 3)	Percent Change in School Enrollment 1989-90 to 1999-2000 (col. 4)	Regression Models: Natural Log of 1990 Census Total Population (col. 5)	Natural Logs of 1990 Census Total Population and Average 1999-2000 School District Enrollment: One Model (col. 6)	by School District Type (col. 7)
School-Age Population	13,650	14.6	14.3	13.3	14.6	12.5	12.4
School District Type: 1990 Census							
Unified School District	10,678	12.3	11.9	11.0	12.3	10.6	10.7
Elementary School District	2,424	21.9	24.2	21.6	21.8	19.9	19.3
Secondary School District	520	28.5	16.0	21.5	28.7	15.6	14.1
Balance of County	12	46.5	46.6	48.6	49.2	50.2	47.4
Percent of Children in Poverty: 1990 Census							
Under 10%	4,122	14.9	13.4	12.1	15.0	11.8	11.8
10% to 24.9%	6,579	14.1	14.3	13.3	14.1	12.6	12.4
25% or More	2,850	14.8	15.3	14.5	14.8	13.0	12.9
School District Total Population: 1990 Census							
Under 5,000	6,079	19.3	20.4	19.2	19.3	17.5	17.2
5,000 to 19,999	4,836	11.3	9.9	9.1	11.3	9.1	9.0
20,000 or more	2,735	10.2	8.3	7.6	10.2	7.7	7.5
Percent Change in School-Age Population, 1990-2000							
Decrease more than 10%	1,894	22.4	20.0	19.7	21.7	19.0	18.3
Decrease 5.1% to 10%	1,023	8.6	12.8	10.4	8.3	10.0	9.4
Decrease up to 5% to Increase up to 4.9%	2,732	8.0	12.4	9.6	7.9	9.6	9.4
Increase 5% to 9.9%	1,371	8.7	12.1	10.3	8.8	9.4	9.2
Increase of 10% or more	6,630	17.3	14.1	14.1	17.6	12.9	13.0
Urban-Rural Location: 1989 CCD Data							
Large City	152	14.5	10.4	8.5	14.5	10.2	10.3
Mid-Size City	680	11.6	10.4	8.9	11.7	8.7	8.9
Urban Fringe, Large City	1,242	12.1	10.9	8.1	12.0	9.3	8.8
Urban Fringe, Mid-Size City	863	12.4	12.2	10.3	12.4	10.3	10.2
Large Town	207	11.5	11.1	9.3	11.5	9.3	9.4
Small Town	4,060	12.7	12.8	11.7	12.7	10.9	10.8
Rural, Outside MSA	6,345	16.8	16.7	16.2	16.9	14.9	14.8
School District Geography: 1990 Census							
Overlapping	5,966	18.3	17.0	16.8	18.3	14.8	14.5
Not Overlapping	7,684	11.8	12.1	10.6	11.8	10.8	10.7

* The numbers shown are the number of school districts included in the MAPE calculations for the synthetic ratio method estimates. These numbers may differ slightly for the experimental estimates because of the criteria for excluding outliers.