Using Administrative Records with Model-Assisted Estimation for the American Community Survey

Robert E. Fay[1]
U.S. Census Bureau, 4700 Silver Hill Rd., Washington, DC 20233-9001

Key Words: ACS, survey estimation, GREG, calibration estimation.

Abstract.

Full implementation of the American Community Survey (ACS) began in 2005. Among other purposes, the ACS will replace the decennial census long-form data, enabling a short-form census in 2010. A test implementation of the ACS in 36 test counties during 1999-2001 suggested that the initial ACS estimation procedure, while adequate at the county level, yields higher variances at the small-area levels of tract and block group relative to the decennial long form. Previously reported research argued the likely success of an approach combining administrative record data with a generalized regression estimator. The case for likely success was based on the R-square of the underlying regression. This paper reports detailed results on a full implementation in 34 test counties, showing the degree to which the predictions based on regression diagnostics have been confirmed.

1. Introduction

The American Community Survey (ACS) completed its first year of full implementation in 2005. The ACS replaces the decennial census long form, which had been an integral part of the last several censuses from 1940 up through Census 2000. Elimination of the long form from the 2010 census enables a redesign of decennial procedures. The full ACS implementation in 2005 was preceded by over a decade of development, including full-scale tests in 36 counties during 1999-2001 and a small-scale national sample at a reduced sampling rate beginning in 2000.

In 2000, the sampling rate of households for the census long form had been approximately 1-in-6 overall. In contrast to the "snapshot" approach of the census, the ACS designates a sample of approximately 1-in-480 households each month, accumulating to approximately 1-in-40 in a one-year period. (Because the ACS, a multi-mode survey, draws a subsample of households for the personal-visit mode, the actual sample yield is somewhat less than the designated 1-in-40.) Current plans are to publish one-year period estimates for states and for counties, places, and other geographic units of population 65,000 or more. By accumulating ACS data over three years, three-year period estimates are planned for geographic units of population 20,000 or more.

In 2000 and previous censuses, tracts (average population of roughly 4000 persons) and block groups (average population of roughly 1500 persons) were the smallest units of publication for long-form data. The ACS will also provide estimates at the tract and block-group levels, but only as five-year period estimates, based on a designated sample of approximately 1-in-8 households. The ACS as a replacement for the long form is further described in Fay (2005a) and several of the sources cited there.

As first noted by Paul Voss and his colleagues (Van Auken et al. 2004) and detailed by Starsinic (2005), tract-level sampling variances for ACS estimates are considerably larger than initially projected, whereas county-level variances generally meet design predictions. In hindsight, most of the ACS shortfall in tract-level precision appears due to the census's use of 100% counts as controls for census weighting areas, areas that typically coincided with the census tracts. During most of the decade, no analogous controls are available to the ACS. Instead, the ACS incorporates controls based on population estimates only at the county level and higher levels of geography. Thus, the 1999-2001 findings in the test counties agreed with the initially projected ACS variances at geographic levels where population controls were used in the weighting, and the findings exceeded the projections at lower levels where population controls were not used.

Last year, two papers (Fay 2005a, 2005b) outlined an approach to improving the precision of ACS tract-level estimates based on a strategy combining model-assisted estimation—specifically generalized regression estimation (GREG)—with administrative records. The basic elements of the proposal are:

1. Link administrative records to the ACS sampling frame, dropping administrative records that cannot be linked.
2. Form unweighted totals of the linked administrative record characteristics at the tract level.
3. Apply ACS sampling weights at the housing unit level to the linked administrative record data that fall into the ACS sample. The weighted estimates at this step represent unbiased (or essentially unbiased) estimates of the unweighted totals in step 2.
4. Using generalized regression estimation (GREG), calibrate the ACS sample weights so that the weighted administrative totals from the sample match the unweighted totals from step 2. (The number of constraints is allowed to vary with the size and other characteristics of each tract.)

5. Use the new weights in subsequent stages of the ACS weighting, which includes ratio and raking/ratio estimation. Although the new weights are adjusted in subsequent estimation steps, the argument is that most of the variance reduction at the tract level will be retained in the final weights.

These steps will be described in more detail in subsequent sections.

An important feature of this approach is that the calibration introduced at step 4 potentially achieves a variance reduction without introducing any appreciable new bias. That is, if the survey weights produce unbiased estimates before step 4, the adjusted weights will essentially continue to do so. By the same token, step 4 is not designed to remove bias, such as bias arising from (1) coverage errors in the frame, whether through undercoverage of housing units or the presence of duplicates, or (2) coverage errors of persons within housing units in the frame. Instead, the sole objective of the GREG estimation in this proposed application is to reduce variance. Consequently, the task of assessing its merits is achieved by comparing the resulting variances using the GREG estimation with the variances without it. If the variance reduction were trivial or non-existent, the evidence would discourage application of the approach. But appreciable variance reductions are a direct measure of the benefits of the GREG step.

In sharp contrast, ratio or raking/ratio estimates potentially can reduce some forms of bias, such as bias from the frame or coverage errors of persons. Because these forms of estimation rely on controls that have their own sources of error, ratio or raking/ratio estimation may introduce new biases into the estimates. For example, post-censal population estimates figure heavily into ACS weighting, and these estimates can have appreciable error, particularly for geographic units with small populations. The amount of variance reduction from ratio or raking/ratio estimation is an important consideration, but variance estimates do not measure the potential reductions of bias or the potential introduction of new sources of bias.

The preceding argument was outlined in Fay (2005a, 2005b). As a preliminary assessment of the potential gains from model-assisted estimation, the two papers analyzed 1999-2001 data from the 36 ACS test counties. The analysis was only diagnostic, however. The proposed GREG estimation had not yet been implemented; instead, the papers reported the predictive value, expressed in the form of R-squared, for some initial regressions predicting ACS sample characteristics on the basis of administrative record data. For example, a housing-unit level regression predicted the number of persons in the ACS household on the basis of (1) an indicator variable denoting whether any administrative record persons had been linked to the household, and (2) second variable equal to the number of linked administrative record persons. Encouraging R-squared values near .5 were reported.

This paper reports results from subsequent empirical work with 34 of the 36 test counties during 1999-2001. The 34 counties each had sampling rates of 3% or 5% per year, yielding designated samples of either 9% or 15% for the 3-year period. The 3-year estimates in these 34 counties are based on sampling rates similar to the designated rates for the full-scale production ACS—12.5% for a 5-year period. (The remaining two counties were excluded because their sampling rates of 1% per year yielded much smaller samples than will be characteristic of the production ACS.)

The empirical results are based on implementing GREG at the tract level in the test counties. Variances were estimated using replication both before and after the GREG step. As will be reported here, the gains previously suggested by the R-square results indeed materialized. Thus, the model-assisted approach remains highly promising.

The paper also discusses initial results expanding the scope of the investigation beyond the tract level. Plans to publish one-year period estimates for places of 65,000+ and three-year period estimates for places of 20,000+ include many places considerably smaller than their corresponding counties. If ratio and raking/ratio estimation include controls only at the county level and higher geographic levels, then model-assisted estimation could potentially improve estimates for smaller places meeting the publication thresholds. The paper comments on initial evidence on this question.

2. Generalized Regression Estimation

The proposal specifically employs GREG estimation, which can be motivated as a special case of model assisted estimation (Särndal, Swensson, and Wretman, 1992). In many applications, GREG estimation can also be motivated as a form of calibration estimation (Deville and Särndal 1992). This section introduces both notions to the level of detail necessary for the subsequent sections of this paper. But the section will touch on only a small part of these substantial literatures, which include a number of review papers (e.g., Fuller 2002). Within the context of this ACS application, an earlier paper (Fay 2005a) reviewed key references in more detail than this section, including Särndal's (1984) argument that model-assisted estimation might be a suitable choice for some small domain estimation problems. Similarly, in *Small Area Estimation*, Rao (2003, ch. 2) also notes the potential use of model-assisted estimation in small area estimation. The review in Fay (2005a) included some theoretical results potentially of future interest in developing this ACS application. Fay (2005a) also summarized applications of GREG estimation in the Canadian population censuses of 1991, 1996, and 2001, (Bankier, Rathwell, and Majkowski 1992; Bankier, Houle, and Luc 1997; Bankier and Janes 2003), which bear a number of parallels with the potential ACS application. A subsequent paper (Fay 2005b) also reviewed basic references, but it instead attempted to emphasize previous applications of model-assisted estimation to official

statistics in the U.S.

Model-Assisted Estimation. Consider a population with values $y_1,\ldots,y_N$. Consider the estimation of the population total by $\hat{Y}$ based on a sample s drawn according to probabilities π_i. Let $W_i^{(0)}$ denote initial weights, either based on the inverse probability of selection, $W_i^{(0)} = \pi_i^{-1}$ or, more generally, weights based on π_i^{-1} adjusted by some early steps of estimation. (In the ACS application, the $W_i^{(0)}$ include household noninterview adjustments, as described in Fay 2005a.) Suppose there are auxiliary data $X = [x_{pi}]$, where x_{pi} represents the value of the pth auxiliary variable out of P and the ith unit out of N in the domain. Assume the auxiliary data are known for the complete population. Let $\hat{Y}^{(0)} = diag(W^{(0)})y$, $\hat{Y}^{(0)\prime}1_n = \sum_s y_i W_i^{(0)}$, and $\hat{X}^{(0)} = x\, diag(W^{(0)}) = [W_i^{(0)} x_{pi}]$.

One expression for the GREG estimator is given by

$$\hat{Y}_{rg} = \hat{Y}^{(0)\prime}1_n + \hat{B}'(X1_N - \hat{X}^{(0)}1_n) \qquad (1)$$

where

$$\hat{B} = (\hat{B}_1,...,\hat{B}_P)' = \left(\sum_s W_i^{(0)} x_i x_i'/\sigma_i^2\right)^{-1} \sum_s W_i^{(0)} x_i y_i/\sigma_i^2 \qquad (2)$$

Different forms of the estimator result from different specifications for $\sigma_i^2 > 0$. Särndal, Swensson, and Wretman (1992) motivate the estimator based on a model ξ for the underlying population, where each y_i is the realization from a random variable Y_i with expected value $E_\xi(Y_i) = \sum_{p=1}^{P}\beta_p x_{pi}$, and variance $\sigma_i^2 > 0$. Equation (2) accounts for the joint roles of the model (through σ_i^2) and design probabilities in estimating the regression.

In their book, Särndal, Swensson, and Wretman (1992) develop GREG estimation within the larger class of model-assisted estimators. Although not obvious from expressions (1) and (2), they show algebraically that a given set of auxiliary data $X = [x_{pi}]$, weights, $W^{(0)}$, and $\sigma_i^2 > 0$ implies a set of new weights W. With the new weights, W, the regression estimator $\hat{Y}_{rg}$ for any Y is given by the weighted estimate of Y for the sample cases. Thus, GREG estimation results in a weighting adjustment that is independent of the choice of Y. (This relationship is show in eq. (4) below.)

Calibration Estimation. In calibration estimation (Deville and Särndal 1992), preliminary weights, such as Horwitz-Thompson weights, are adjusted to calibrate sample estimates to known population totals, subject to a loss or penalty function on the degree of difference from the preliminary weights. One form of loss function leads directly to regression estimation, but other forms of calibration estimation result from different loss functions.

In general, the calibrated weights, $g_i W_i^{(0)}$, satisfy the P constraints $\hat{X}^{(0)}g = X1_N$ subject to minimizing the quantity $L = \sum_s \pi_i^{-1}(g_i - 1)^2/q_i$. The specific choice $q_i = 1/\sigma_i^2$ leads back to GREG given by (1) and (2).

Thus, an appropriate choice of parameters connects model-assisted estimation and calibration estimation. In calibration estimation, a primary emphasis is to match the new weighted estimates to a set of specified population constraints. In the proposed application to ACS, the estimator calibrates the weights to match totals that result from the linkage of the administrative records to the ACS frame. These totals will not be published, however, and are merely a means toward an end—variance reduction in the ACS estimates.

GREG in the Census of Canada. As previously noted, Bankier and his colleagues have published papers corresponding to each implementation of GREG estimation for the Census of Canada in 1991, 1996, and 2001. Fay (2005a) summarized their notation, which has some advantages in complex applications.

For example, Bankier and Janes (2003) link their GREG application both to the model-assisted and calibration literatures, but they express most relationships in the form of weighted estimates. In complex applications, including the ACS, GREG estimation may follow other steps of estimation that have adjusted the survey weights. Consequently, this approach more readily shows how the estimator is calculated from the weighted estimates based on previous estimation steps.

Bankier and Janes (2003) stipulate a function L for calibration estimation in the following form:

$$L = (g - 1_n)'\hat{V}(g - 1_n) \qquad (3)$$

where the matrix $\hat{V}$ should be symmetric and positive definite. For a given $\hat{V}$, they expressed the result of minimizing (3) as

$$g = 1_n + \hat{V}^{-1}\hat{X}^{(0)}(\hat{X}^{(0)\prime}\hat{V}^{-1}\hat{X}^{(0)})^{-1}(X1_N - \hat{X}^{(0)}1_n) \qquad (4)$$

Note that (4) employs the matrix of weighted characteristics, $\hat{X}^{(0)}$. Using the standard argument in the literature, Bankier and Janes (2003) also remark that any characteristic estimated by $\hat{Y} = \hat{Y}^{(0)}g = \sum_i g_i W_i^{(0)} y_i$, can be written in the standard form of a regression estimator

$$\hat{Y}_{rg} = \hat{Y}^{(0)} 1_n + \hat{B}'(X 1_N - \hat{X}^{(0)} 1_n) = \hat{B}' X 1_N + \hat{e}^{(0)} 1_n \quad (5)$$

where

$$\hat{B} = (\hat{X}^{(0)} \hat{V}^{-1} \hat{X}^{(0)'})^{-1} \hat{X}^{(0)} \hat{V}^{-1} \hat{Y}^{(0)'} \quad (6)$$

and $\hat{e}^{(0)} = [W_i^{(0)} e_i]$ is a 1 x n vector of weighted residuals, $e_i = y_i - \hat{B}' x_i$. Regardless of the characteristic *Y*, $\hat{B}$ given by (6) is consistent with (4), thus emphasizing the connection between regression and calibration.

The mathematically equivalent expressions in (5) provide two characterizations of the regression estimator. (Both forms appear often, such as in Särndal, Swensson, and Wretman 1992) The first (identical to (1)) shows the estimator as the sum of the Horwitz-Thompson estimator and a regression correction based on the differences between the population and weighted sample *x*'s. In the second, regression predictions for the population are adjusted by a correction based on weighted residuals. Either equation can be used to show why the estimators are asymptotically unbiased.

3. Possible Application to ACS

The ACS Frame. Leading up to Census 2000, the Census Bureau has maintained a Master Address File (MAF) for all housing unit addresses in the U.S. (Although not in scope for the ACS, the MAF includes business addresses as well.) The MAF was the basis of Census 2000. The file is updated on an ongoing basis, including a revision based on Census 2000.

The ACS sample is drawn from the MAF as its single frame. In contrast, most other household surveys conducted by the Census Bureau—including the Current Population Survey—employ multiple frames, including both list and area sampling. In a sense, the use of the MAF as a single frame brings the ACS closer to a "textbook" sampling application.

A complicating detail is that two new extracts from the MAF are used each year as the ACS sampling frame in order to remain representative of the current housing unit inventory. Over a three-year period, six such extracts are used. This feature can be reflected in the application of GREG estimation.

Multiple Modes. Each month, the approximately 1-in-480 sample of housing units drawn from the MAF is mailed an ACS questionnaire. (There are also provisions for unmailable addresses.) A window of about one month is allowed to accumulate the questionnaires from the approximately half of the sampled housing units that return questionnaires by mail. The next two months include telephone interviewing, when possible, and a personal visit followup for a subsample. The subsampling rate for personal visit is approximately 1-in-3 overall, but the rates vary locally depending on expected response to the first two modes. By design, the designated sample in a given month can respond over a period of three months depending on the mode of interview.

Estimation Steps. A previous paper (Fay 2005a) listed 15 steps in the estimation process in order to indicate a possible placement for GREG. Here, the estimation steps are summarized by the following:

(1) Initial estimation steps reflecting the inverse of the probability of selection, including subsampling for personal visit, and housing-unit noninterview adjustments. At this stage, ACS sample cases have only housing-unit weights.
(2) The proposed GREG estimation at low-level geography, such as census tracts. Weights remain at a housing-unit level.
(3) Ratio estimation for housing unit counts at a higher geographic level, such as counties. Weights remain at a housing-unit level.
(4) Conversion of housing-unit weights into person-level weights, ratio and raking/ratio estimation of the person weights, and adjustments to align the housing unit weights for partial consistency with the person-level weights. The result is a set of housing-unit and person-level weights.

Estimation steps (3) and (4) adjust ACS weights to be consistent with external controls. Consequently, early placement of the proposed GREG estimation at (2) allows it to improve tract-level estimates without interfering with the goal of imposing the external controls in steps (3) and (4).

Administrative Record Data. The Census Bureau has used administrative records for many years, including in economic censuses and surveys and in post-censal population estimation. More recently, staff members in the Administrative Records Research Branch in the Data Integration Division (and its organizational predecessors) have consolidated the acquisition and processing of administrative records for statistical uses at the Census Bureau. The Master Address File Auxiliary Reference File compiles a census-like portrayal of the population covered by the administrative records, showing the basic demographic characteristics of persons and their households. Most of these households have been linked to the MAF.

General Approach. The approach uses only administrative records from the Master Address File Auxiliary Reference File that are linked to the versions of the MAF used as the sampling frame for the ACS. The remaining records from the Master Address File Auxiliary Reference File are ignored in this approach. At the level of the individual MAF entry, the administrative record data are used to define a set of *x* variables, such as the number of persons in the household or the number of persons age 0-17. An unweighted sum of these *x* variables forms the *X* of eq. (3) and other expressions in the previous section. Because all ACS sample cases are drawn from the MAF, it is possible to apply the ACS sample weights to the *x* variables for housing

units in the ACS sample. The weighted sum, $\hat{X}^{(0)}$, should theoretically be an unbiased estimate of X, an observation that supports the claim that the GREG in this application is asymptotically unbiased.

Remark on Household Noninterviews. The noninterview adjustments in (1) precede the GREG at (2), so the preceding argument assumes that the noninterview adjustments are effectively unbiased. If they are not, however, the GREG may actually reduce their bias to some degree. For example, if housing units with only one person are more inclined not to respond, the use of number of persons as an x variable in the GREG may reduce the bias somewhat. Fortunately, household nonresponse, about 3%, remains a relatively small problem in ACS.

4. Preliminary Application to Tracts in the Test Counties

Five populous test counties—San Francisco, CA; Broward, FL; Lake, IL; Bronx, NY; and Franklin, OH—were sampled at 3%, and 29 test counties were sampled at 5%. The ACS sample in these test counties was drawn from six MAF extracts. As in Fay (2005a), only the records from the 2000 Master Address File Auxiliary Reference File were matched to each of the six ACS frames. (The discussion section will return to how administrative record data for different years could be used in production.) As intuition would suggest, the correlations between the ACS and 2000 administrative data are highest for 2000 ACS data. But the correlations are nonetheless sizeable for 1999 and 2001 ACS data (Fay 2005a).

The complexities of the frame can be reflected in the calculation of the unweighted X used in the GREG. Over the three-year period, the matrix of weighted characteristics, $\hat{X}^{(0)}$, estimates a weighted average of the six individual frames: each of the three years, 1999-2001, are weighted equally, but within each year, the first extract is used as the frame for three months and the second for nine. The weighted sum should therefore be estimated by weighting three of the extracts by 1/12 and the other three by 3/12 = 1/4.

Some census tracts include no housing units, and some small tracts in these counties had no ACS sample hits. Additionally, some tracts are so small that they would be expected to yield few ACS sample cases. For purposes of this study, GREG estimation was not attempted in tracts with less than 300 housing units in the merged frames. In the 34 counties, a total of 186 tracts with ACS sample cases fell into this category. Because the purpose of the GREG estimation is to reduce variance where possible without adding appreciable bias, it is logical to skip the GREG step in tracts where its application could be problematic. (Remark: When estimation steps are intended to reduce bias as well as variance, such as the ratio and raking/ratio estimators in steps (3) and (4) in the preceding section, some form of adjustment in small tracts, perhaps by collapsing small tracts together, would seemed to be necessary.) The unadjusted small tracts were not dropped from the analysis but instead included in the variance totals for both with and without GREG estimation.

For purposes of this study, three different forms of GREG were implemented: (1) a GREG where the regression included only a constant term, (2) a GREG with x variables based on sex and broad age groups, and (3) a GREG including x variables for race and/or ethnicity in sufficiently diverse tracts, as well as the same age/sex variables in (2).

Constant Regression. Simply including $x=1_N$ achieves a tract-level consistency between the sample and frame. This approach approximates the one of Starsinic (2005), who assumed a tract-level housing unit control. If the frame provided an exact inventory of housing units, then the approaches would be identical. In fact, the ACS frame includes a relatively small percentage of MAF units that, when followed up in personal-visit mode, turn out not to be valid housing units. Thus, the GREG provides an estimate of the number of valid housing units, but one that still has a variance because invalid units in the frame are detected only for the personal visit subsample.

Age/Sex Regression. For purposes of this study, the age distribution was divided into the broad groups 0-17, 18-29, 30-44, 45-64, and 65+. Based on the demographic information from the 2000 Master Address File Auxiliary Reference File, a set of 8 x variables is defined by

x1 = 1
x2 = 0-17 M+F,
x3 = 18-29 M+F
x4 = 30-44 M
x5 = 30-44 F
x6 = 45-64 M
x7 = 45-64 F
x8 = 65+ M+F

a reduced 4-variable alternative is based on

x1 = 1
x2 = 0-17 M+F,
x3 = 18-44 M+F
x4 = 45+ M+F

and a 2-variable regression is based on

x1 = 1
x2 = total admin persons.

Note that all three equations calibrate total administrative record persons. The 1-variable regression used only

x1 = 1

constraining the weights to agree with the number of units in the frame.

For this study, an initial attempt was made to fit the 8-variable regression model. Two conditions were checked: that the matrix inversions in eq. (4) could be computed algebraically, and that none of the resulting weights were less than 0.5. (Fay (1995a) reviewed a related literature on alternative approaches to preventing negative weights and similar constraints, but these approaches have not yet been investigated.) For a given tract, if either condition failed for

the 8-variable regression, then the 4-variable regression was tried, and the two conditions were again checked. The cycle was repeated for the 2-variable regression. The 1-variable regression used as the last resort.

Race/Ethnicity. Four broad race/ethnicity groupings were considered in addition to the outcome of the age/sex calibration: Hispanic origin, non-Hispanic Black, non-Hispanic White, and all other. The possible regressions added one or two variables with the number of persons in the following categories in the administrative record data:

Hispanic, non-Hispanic Black (two variables)
Hispanic
Black
Other

In many tracts, however, the results for the age/sex GREG were used without any term involving race or ethnicity.

Variance Estimation. In general, ACS variances are estimated through replication, using 80 replicate weights. For this analysis, variances before GREG estimation were computed from the replicate weights available at the end of the noninterview adjustments. For each tract, the full sample was used to select the version of the regression for GREG, and the selected regression was implemented using each of the 80 sets of replicate weights, producing 80 new sets of replicate weights reflecting GREG estimation.

Results. Table 1 summarizes the use of regression variables in the most complex version of the GREG.

Table 1. Use of regression variables in tract-level GREG implementation. Cell counts give the number of tracts in the 34 ACS test counties for the GREG implementation attempting both age/sex and race/ethnicity variables.

	Housing count in frame			
	0-299	300-999	1000-1999	2000+
Age/sex				
No GREG	186	0	0	0
1-var	0	18	10	3
2-var	0	33	34	23
4-var	0	83	161	85
8-var	0	130	757	913
Total	186	264	962	1024
Race/ethnicity				
Hisp, Bl	0	0	28	128
Hisp	0	30	173	250
Bl	0	13	123	102
Other	0	5	44	105
None	186	216	594	439
Total	186	264	962	1024

Table 2 displays results, separately for the 3% counties and 5% counties, comparing the sum of tract-level variances after step (1) to the sum of tract-level variances after GREG estimation, step (2). In general, the results are promising. For housing units, the initial work of Starsinic (2005), approximated here by the constant-term regression, achieves a substantial reduction in housing-unit statistics, a gain that the more complex GREG alternatives only marginally improve upon. Replicating Starsinic's findings, the constant-term regression also reduces variance for the estimated number of persons, but the more complex GREG alternatives achieve greater reductions. The constant-term regression makes only modest improvements for the detailed demographic categories, while the GREG alternatives are again successful.

The race/ethnicity regressions improve somewhat for tract-level variances of race and ethnic groups beyond the GREG based only on age/sex. Notably, incorporating race/ethnicity into the GREG does not produce overall increases in variance for the other characteristics, compared to the age/sex GREG.

5. Discussion

This research can be justly termed work in progress, but the results presented here already carry some potential lessons for similar applications. Two features are of particular interest:

1. The proposed application imbeds GREG as one of the steps in a complex estimation that also includes ratio and raking/ratio estimation at later steps. The goal of the added GREG step is variance reduction for small domains.
2. Although the application works with administrative records, the problem of records that do not match the survey is addressed by first matching to the frame. For purposes of GREG estimation, the population totals are based on the administrative records that can be matched. This approach supports claims of asymptotic unbiasedness.

Although it is not certain that this is the first such application with these two features, the literature reviews in Fay (2005a, 2005b) did not point to a predecessor.

As noted in the introduction, this effort began with the goal of improving the reliability of 5-year estimates for tracts. A separate initial investigation supports the notion that GREG could be beneficial for some of the subcounty publication areas that will be released as 1-year or 3-year estimates. The 1-year estimates will be for places and other geographic entities with population of 65,000 or more. The 3-year estimates will be for entities with population of 20,000 or more.

All estimates in the paper employ the administrative record data for 2000. A more appropriate version for 5-year estimates would use 4 years worth of administrative data. For each of the first 4 years, the administrative data could be matched to the frame for its corresponding year. For the last year of the 5-year period, the administrative data from the preceding year could be matched. This approach allows for a 1-year lag in obtaining administrative record data. Because the estimator is model-assisted rather than model-based, use of administrative data with a 1-year lag will incur a small loss of precision. But the model-assisted approach does not face significant problems of bias. Generally, any model-based approach would face the difficult problem of assessing the impacts of bias if it used lagged data, impacts

that could be large.

The empirical results demonstrate feasibility rather than optimality. Later research can expand the search for effective selection of variables and for strategies for insuring positive weights with desirable variance properties. Other aspects of the agenda outlined in Fay (2005a) also remain to be addressed, such as improving the reliability of block-group estimates as well as tracts.

A number of other research questions remain open. Parallel analyses are now planned with subsequent years of the ACS in the test counties, including the use of administrative record data from 2001-2004.

Note: (1) This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Michael Beaghen and Doug Olson provided helpful comments on an earlier version.

References

Bankier, M.D., Rathwell, S., and Majkowski, M. (1992), "Two Step Generalized Least Squares Estimation in the 1991 Canadian Census," *Proceedings of the Survey Research Methods Section,* American Statistical Association, pp. 764-769.

Bankier, M., Houle, A.-M., and Luc, M. (1997), "Calibration Estimation in the 1991 and 1996 Canadian Censuses," *Proceedings of the Survey Research Methods Section,* American Statistical Association, pp. 66-75.

Bankier, M. and Janes, D. (2003), "Regression Estimation of the 2001 Canadian Census," *Proceedings of the 2003 Joint Statistical Meetings on CD-ROM,* American Statistical Association, pp. 442-449.

Deville, J. and Särndal, C.-E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association, 87,* 376-382.

Fay, R.E. (2005a), "Model-Assisted Estimation for the American Community Survey," *Proceedings of the 2005 Joint Statistical Meetings on CD-ROM,* American Statistical Association, pp. 3016-3023.

_____ (2005b), "Potential Applications of Model-Assisted Estimation to Demographic Surveys in the U.S.," paper presented at the Federal Committee on Statistical Methodology Research Conference, available from www.fcsm.gov/05papers/Fay_IIIC.pdf.

Fuller, W.A. (2002), "Regression Estimation for Survey Samples," *Survey Methodology, 28,* 5-23.

Rao, J.N.K. (2003), *Small Area Estimation,* John Wiley, New York.

Särndal, C.-E. (1984), "Design-Consistent Versus Model-Dependent Estimation for Small Domains," *Journal of the American Statistical Association, 79,* 624-631.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling,* Springer-Verlag, New York, NY.

Starsinic, M. (2005), "American Community Survey: Improving Reliability for Small Area Estimates," *Proceedings of the 2005 Joint Statistical Meetings on CD-ROM,* American Statistical Association, pp. 3592-3599.

Van Auken, P.M., Hammer, R.B., Voss, P.R., and Veroff, D.L. (2004), "American Community Survey and Census Comparison, Final Analytical Report, Vilas and Oneida Counties, Wisconsin; Flathead and Lake Counties, Montana," unpublished report dated March 5, 2004, available at http://www.census.gov/acs/www/AdvMeth/acs_census/lreports/vossetal.pdf.

Table 2 Preliminary percent reduction in estimated tract-level variance from three possible GREG estimation strategies in 34 ACS test counties, 1999-2001. Reductions are shown separately for 5 large counties sampled at approximately a 3%/year rate. The remaining 29 counties were sampled at 5%/year. All estimated variances are for the estimated totals of each characteristic. The fourth and last columns provide the range of reductions for the set of counties.

	% reduction in 3%/year counties				% reduction in 5%/year counties			
	Const. term regr.	Age /sex regr.	Age /sex, race /ethn regr.	Range for age/sex, race/ethn regr.	Const. term regr.	Age /sex regr.	Age /sex, race /ethn regr.	Range for age/sex, race/ethn regr.
Housing units	90	91	91	90 - 93	87	88	88	54 - 97
Occupied hu's	69	74	74	67 - 80	62	69	69	36 - 78
Total persons	47	67	68	63 - 74	42	66	66	46 - 73
Males 0-17	13	39	40	36 - 47	11	40	41	6 - 54
Females 0-17	13	39	40	35 - 45	11	41	41	-5 - 60
Males 18-29	10	26	27	21 - 31	9	26	26	-6 - 38
Females 18-29	11	28	28	24 - 33	10	28	28	0 - 39
Males 30-44	14	40	39	21 - 49	11	42	42	22 - 58
Females 30-44	17	46	46	40 - 55	13	47	47	24 - 55
Males 45-64	8	43	43	31 - 53	6	47	47	27 - 61
Females 45-64	11	45	46	38 - 56	6	49	49	27 - 60
Males 65+	1	25	25	20 - 31	-2	31	31	13 - 47
Females 65+	3	29	29	24 - 40	-2	35	35	18 - 47
Hispanic	21	33	50	22 - 52	23	39	48	-54 - 62
Non-Hisp Black	22	33	46	35 - 50	16	32	43	-27 - 61
Non-Hisp White	26	43	51	36 - 59	24	45	51	-18 - 71
Other races	10	30	44	3 - 67	9	19	26	-154 - 81