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, small area estimates, generalized regression estimation, calibration estimation.

Abstract.

The Census Bureau has proposed the American Community Survey (ACS) as a replacement for the long-form data collected in the last several decennial censuses. This advance would provide more timely local data than the census and would simplify the 2010 Census by eliminating the long form. To satisfactorily replace the long form, however, ACS estimates at local levels such as census tracts must have reliability approximately comparable to long-form estimates. The initial experience for 36 test counties during 1999-2001 indicated that the variances of the ACS estimates, particularly for totals, are larger at the tract level than previously anticipated. In hindsight, census long-form estimates benefit from raking/ratio estimation to census population controls for weighting areas the size of typical tracts; through 2004, ACS estimates have not yet used population controls below the county level.

Model-assisted estimators using administrative records appear to offer a partial technical solution to this problem. This paper reports an initial investigation of how these methods could apply to small area data from the ACS, particularly estimates formed as three- or five-year averages.

1. Introduction

In 2005, the American Community Survey (ACS) moved into full production after several years of testing and development. The timing fits with the U.S. Census Bureau's plan to replace the decennial census long form with the ACS. Because the long form provided geographically detailed data, including data for census tracts and block groups, the ACS must provide estimates of comparable quality to represent a full replacement. This paper particularly concerns possible strategies for ACS estimation at the lowest geographic levels offered by Census 2000.

In developing the ACS, the Census Bureau tested a full-scale implementation in 36 counties beginning in 1999. (Equivalently, these counties are referred to elsewhere as the 31, or occasionally 30, *test sites.*) In 34 of the 36 counties, the sampling rates were set at either 3% or 5% per year, so that the ACS sample accumulated during the 3-year period 1999-2001 was approximately the same size, 9% or 15%, as the 12.5% expected to accumulate over 5 years for the production ACS at approximately 2.5% per year.

The Census Bureau sponsored four studies of the 1999-2001 test results by outside researchers, who compared specific site-level results with Census 2000. Gage (2004), Hough and Swanson (2004), Salvo, Lobo, and Calabrese (2004), and Paul Voss and his colleagues (Van Auken et al. 2004) all reported favorably on the ACS results. They found that the estimated sampling variances for ACS county estimates were generally in line with previous projections and, with some notable exceptions, the ACS and decennial long-form results presented substantively similar results. Paul Voss and others pointed out, however, that the variances of the ACS estimates at the tract level were disappointingly higher than expected.

The disappointing tract-level variances have led to a set of related research efforts to investigate estimation strategies to improve the reliability at sub-county levels. Starsinic (2005) reported on a detailed variance study of tract-level estimates for the 34 test counties, substantially expanding on the observations of Voss and his colleagues. He also reported on the variance impact of using population controls at the tract level, if such controls were available.

For the period between decennial censuses, administrative records appear the most promising source of geographically detailed auxiliary information. The Administrative Records Staff at the Census Bureau has been engaged in a multi-year effort to systematize the acquisition and processing of administrative records for statistical uses at the Census Bureau. One of their products, 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. The records can be linked to specific housing unit entries in the Census Bureau's Master Address File (MAF). (The MAF also provides the ACS sampling frame.) Judson (2000) provided an overview of the Administrative Records project.

Don Malec of the Census Bureau plans an approach to the ACS estimation problem using administrative records matched at the person level. At the time of this writing, his research is at too preliminary a stage to be summarized here. Simpler approaches involving ratios of administrative record totals over time have also been proposed.

This paper will describe another research effort using administrative records, but in considerably different manner. Specifically, matching is conducted at the housing unit level only, using only records matched to the MAF. With this approach, it is unnecessary to determine whether persons in the administrative sources are the same as reported in the ACS.

The title of this paper uses the term *model-assisted estimation* to denote the broad class of estimators considered by Särndal, Swensson, and Wretman (1992).

The estimators generally appeal to models in order to incorporate auxiliary information, yet they have valid design-based properties. For example, ratio estimation may be regarded as a form of model-assisted estimation. The terms *generalized regression* or *GREG estimators* and *calibration estimators* are used to denote specific forms of model-assisted estimation, forms that are of interest in this application.

When a model-assisted estimator is applied to a problem with adequate sample size, its bias is generally negligible relative to its sampling variance, whether or not the underlying model holds for the population. Consequently, objectively assessing the performance of a model-assisted estimator essentially reduces to a problem in variance estimation. In contrast, bias is an important component of error, sometimes the primary component, for *model-based* or *model-dependent estimators,* including many often considered in small domain estimation. Thus, model-assisted estimation provides a feasible route to exploit the auxiliary information available from administrative records, even in the absence of a complete accounting for systematic differences between the administrative data and the ACS population.

The general purpose of the research effort is to apply or adopt existing theory and practice to the ACS rather than to advance the estimation literature. For this reason, the paper will first outline the ACS problem and its specific constraints before identifying potentially applicable theory. Section 2 will describe publicly available long-form data from Census 2000, to suggest what users of small-area data might expect from the ACS. Section 3 will describe ACS sampling and the current plan to release estimates based on 3- and 5-year averages in addition to 1-year estimates. Section 4 will then describe the current ACS estimation methods in sufficient detail to show where additional estimation steps could be incorporated to improve tract-level estimates. Section 5 notes some of the extensive theoretical literature, and the following section summarizes aspects of experience from the Canadian Census closely related to the ACS problem. The final two sections report on initial empirical work and discuss related implications.

2. What Census 2000 Offers Users of Small-Area Long-Form Data

Although the sampling rates for Census 2000 purposefully varied depending on geographic considerations, overall the census sampled approximately one in six households to receive a long form rather than the short form. The short form asked only the most basic person characteristics—age, sex, ethnicity, race, and household relationship—such as those mandated by the apportionment of the House of Representatives and the Voting Rights Act. It also asked only a few housing unit characteristics, such as tenure and occupancy status. The long form asked both the short-form items and also more detailed characteristics such as marital status, education, journey to work, disability, ancestry, and income.

The Census Bureau's web site (www.census.gov) offers the public a wealth of information, including data from Census 2000. To illustrate, this section describes information that I directly gathered from site, without privileged access to internal information. Readers are encouraged to replicate my findings.

The primary gateway to Census 2000 (http://www.census.gov/main/www/cen2000.html) displays a variety of paths into small-area census data. For example, users unfamiliar with geographic terms such as *tract, block group,* and *block* can find a geographic glossary (http://www.census.gov/geo/www/tiger/glossry2.pdf) explaining that counties are divided into tracts, which are further divided into block groups and then into blocks. Tracts "generally have between 1,500 and 8,000 people, with an optimum size of 4,000 people." Block groups "generally contain between 600 and 3,000 people, with an optimum size of 1,500 people." The census block is the basic unit of census geography, and blocks vary widely in population. But to protect confidentiality, only short-form (100%) data are released at the block level. The block group is the lowest geographic level of release for long-form data, and a similar policy is planned for the ACS.

SF 3 provides geographically detailed long-form information. "Summary File 3 consists of 813 detailed tables of Census 2000 social, economic and housing characteristics compiled from a sample of approximately 19 million housing units (about 1 in 6 households) that received the Census 2000 long-form questionnaire." (http://www.census.gov/Press-Release/www/2002/sumfile3.html). SF 3 offers many tables down to the block-group level (P and H tables), but others only to the tract level only (PCT and HCT tables).

Users can readily observe population discrepancies between the 100% and sample estimates at the block-group level; for example, in tables P1 and P3 of SF 3. Because tracts were frequently used as weighting areas, however, often there is exact agreement at the tract level. Table 1 shows the results for the tract where I now live.

Table 1 Comparison of Census 2000 100% and Sample Counts by Block Group, Tract 1045.01, Montgomery County, MD.

Block Group	100 %	Sample
BG 1	799	822
BG 2	1063	992
BG 3	796	787
BG 4	806	863
Tract total	3464	3464

Source: U.S. Census Bureau, American Fact Finder, Tables P1 and P3, SF 3, obtained 13 May 2005.

Additional long-form data is available from SF 4, but only down to the tract level. "The sample data are presented in 213 population tables (matrices) and 110 housing tables,

identified with 'PCT' and 'HCT', respectively. Each table is iterated for 336 population groups: the total population, 132 race groups, 78 American Indian and Alaska Native tribe categories (reflecting 39 individual tribes), 39 Hispanic or Latino groups, and 86 ancestry groups."

Sampled at roughly 1-in-6, the typical block group of 1500 people in Census 2000 might provide only 250 sample persons (or fewer because of nonresponse) and roughly 100 sample housing units. By the standards of typical household surveys, these are small samples. But providing users data at this level of detail allows them to aggregate SF 3 data to approximate other local areas if the geographically coarser tracts are unsuitable. (U.S. readers are encouraged to find the boundaries of their own tracts and block groups.)

Some of the tables of SF 3 appear designed particularly for geographic aggregation. P33, for example, shows aggregate minutes of commuting to work for workers not working at home. The aggregate for my own block group 2, 10,000 minutes (exactly—the estimates for block groups 1, 3, and 4 are 10,865, 11,025, and 11,220, respectively), is an interesting statistic, but one that probably finds more use when aggregated across areas and then divided by an appropriate denominator.

3. The ACS as a Replacement for the Long Form

Although data collection for the census generally spans more than three months, most users are not substantially misled by interpreting the census data as if it were a snapshot of the population on April 1. The census design intends that it measure the population as of April 1; and the reference period for income items is the previous calendar year regardless of when the census response is obtained.

In contrast to the decennial census, the ACS is an ongoing monthly survey, with new samples of housing units selected each month at the approximate rate of 1 in 480. Aggregated over time, the ACS data provide data similar to the census long form, but averaged over time rather than as a snapshot. The late Charles (Chip) Alexander deserves credit for developing much of the intellectual foundation for the ACS. Before his untimely death in 2002, he described the overall ACS strategy (Alexander 2002a, 2002b). Other descriptions are widely available, including one for local government officials (U.S. Census Bureau 2004a).

ACS samples are selected from the Census Bureau's Master Address File (MAF). The MAF was closely linked to Census 2000 data collection, but now is continually updated to reflect changes in the housing inventory. Data collection for each monthly ACS sample is spread out over 3 months. To simplify the details somewhat, the first month of data collection is allocated to mail response; the second, to telephone; and the second and third, to personal followup visits through Computer Assisted Personal Interviewing (CAPI). In the second month, a telephone interview is attempted for any household not responding in the first month, when a satisfactory telephone number is available. In the second and third months, the remaining sample is subsampled at an overall rate of 1 in 3 for personal visit, including households where telephone follow up was unsuccessful. The personal visit is the only opportunity to determine whether a unit sampled from the frame is definitely occupied, vacant, or should be deleted (currently non-existent or ineligible), because almost all responses during the mail and telephone phases are from occupied units.

Instead of a fixed reference period, the ACS employs a moving reference window, so that those answering by mail are generally reporting for a different reference period than persons from the same selected sample who fail to respond by mail and are interviewed in personal follow-up (CAPI) one or two months later. Consequently, the data collected for any one month does not have a strict probability basis, and there are currently no plans to release monthly estimates from the ACS.

Averaged over a year, the 1-in-480 monthly sample becomes approximately 1-in-40, a much smaller sample that the 1-in-6 long form. The current plan is to release one-year ACS estimates only for areas with population 65,000 or more. Similarly, a threshold of 20,000 persons is planned for release of three-year averages. Tract and block-group estimates will be released only for the five-year averages.

4. Current ACS Weighting

The specifications for ACS weighting (U.S. Census Bureau 2002) detail the estimation steps for programming purposes. A more accessible summary (U.S. Census Bureau 2004b) was written to provide an overview of ACS estimation steps for a broader audience, including external peers in the statistical profession. The summary (p. 3) divides ACS estimation into 15 steps, shown in fig. 1.

Like the decennial censuses, the ACS collects data for both housing units and persons, and the estimation process results in separate weights for each housing unit and person in sample. Broadly speaking, the estimation steps begin with housing unit weights representing the inverse probability of selection, including subsampling for CAPI. In step 5, a high-level ratio estimate is applied to level out the effect of variations in response rates across the three modes. Steps 6-9 adjust for household noninterviews during the CAPI phase; the logic of these steps is complex, but the strikingly low nonresponse in ACS (approximately 3%) mitigates the impact of the adjustments. Step 10 computes and applies a ratio estimate to estimated housing controls at the county level.

Distinct person-level weights first appear at step 12—until then the estimation steps are applied at the housing-unit level only. At the end of step 10, only (1) interviewed households (including those during the CAPI phase), (2) vacant housing units and those occupied by non-residents (identified during the CAPI phase), and (3) units that should be deleted (also identified during the CAPI phase) will have

positive weights.

Step 11 begins the process of assigning initial weights to persons from the current values of the household weight; the person weights are adjusted to controls in step 12. A single household member is selected as a *principal person*, such as the wife in a jointly headed wife/husband household or the household head in most other situations. The weight of the principal person in an occupied household, and the process is begun a second time at step 10. The last steps—14, where weights are rounded to integers; and 15, where outliers are identified and down-weighted—would have little direct impact on tract- and block-group estimates.

The Major Steps to the Weighting Process

1. Prepare the files for weighting.
2. Swap housing units for disclosure avoidance.
3. Form the collapsed estimation strata
4. Apply the base weights and CAPI sub-sampling weights.
5. Apply a monthly adjustment to make the total weighted number of responses agree with the actual weighted mail out each month. (Monthly sample factor).
6. Apply a non-interview factor (1) by tract and building type.
7. Apply another non-interview factor (2) by month and building type.
8. Apply another non-interview factor for CAPI cases only using month and building type.
9. Apply a non-interview factor (mode bias factor) by tenure, month, and marital status.
10. Control the housing unit (HU) counts to a larger geographic level.
11. Form the population control weighting cells.
12. Apply the HU weights to all people in a HU and control their weights to the population controls.
13. Apply the principal person weight to the HU and apply the housing unit controls again.
14. Round the housing unit and person weights.
15. Identify and down-weight outliers.

Fig. 1. The steps of ACS weighting as summarized in 2004. A potential place to locate weighting adjustments to use administrative records is between steps 9 and 10 or 10 and 11. For more detail on these steps, see U.S. Census Bureau (2004b). Some aspects of the weighting—such as the preparation (1), collapsing (3), cell formation (11), and rounding (14)—involve details that do not appreciably affect tract and block-group reliability.

This analysis suggests two alternative placements for an additional step or steps designed to improve sub-county estimates through model-assisted estimation. Without affecting the current non-interview adjustment, model-assisted estimation could be implemented either between steps 9 and 10, as a step 9a, or between steps 10 and 11, as a step 10a. The preliminary results presented in this paper take the second approach, building on the weights available from step 10.

If the estimation is placed at 10a, then the model-assisted estimation could be implemented either

- by adjusting the single household weight, leaving step 11 to assign the household weights to persons and adjust them, or
- by estimating tract-level population estimates, applying the housing unit weights to the persons, and using raking-ratio estimation to produce person level weights for step 11.

The second option could contribute to further weight variation among household members, so the research will favor attempting to achieve a single household weight at the end of step 10a.

5. Generalized Regression and Calibration Estimation

The overlapping literatures for generalized regression and calibration estimation include a substantial number of papers, and even careful reviews (Fuller 2002, Rao 1994) do not trace the priority of individual researchers' contributions in complete detail. This section emphasizes a few key references to serve as an indication of the general growth of the field.

Generalized regression estimation for finite samples can be motivated through (1) ties to linear regression or (2) as a specific member of the class of calibration estimators. As an example of the first approach, Särndal, Swensson, and Wretman (1992) explain the regression estimator in their Chapter 6 by showing its connections both to linear regression generally and to the difference estimator. They consider the estimation of a population total $\hat{Y}$ for a population with values $y_1,\ldots,y_N$ based on a sample s drawn according to probabilities π_i. There are auxiliary data $X = [x_{pi}]$, where x_{pi} represents the value of the pth auxiliary variable out P and the ith unit out of N in the domain. For simplicity, assume the auxiliary data are known for the complete population. Let $W_i^{(0)} = \pi_i^{-1}$, $\hat{Y}^{(0)} = diag(W^{(0)})y$, $\hat{Y}^{(0)\prime} 1_n = \sum_s y_i / \pi_i$, and $\hat{X}^{(0)} = x\, diag(W^{(0)}) = [W_i^{(0)} x_{pi}]$.

They introduce the regression estimator (p. 225) by

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

where

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

They 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 σ_i^2. Equation (2) accounts for the joint roles of the model (through σ_i^2) and design probabilities in estimating the regression. The balance of Chapter 6 elaborates the general theory of regression estimation. Chapter 7 links the general theory to common applications, including ratio estimation, simple regression, and multiple regression. Calibration estimation is virtually unmentioned in their development.

Both Rao (1994) and Fuller (2002) credit Deville and Särndal (1992) for introducing the terms *calibration estimation* and *calibration weights.* Preliminary weights, such as the 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.

The calibrated weights, $g_i W_i^{(0)}$, satisfy the *P* constraints $\hat{X}^{(0)} g = X 1_N$ subject to minimizing the quantity $L = \sum_s \pi_i^{-1} (g_i - 1)^2 / q_i$. Setting $q_i = 1/\sigma_i^2$ leads back to (1) and (2).

Subsequently, authors have often combined both the original motivation for regression estimation with the calibration characterization. Bankier and Janes (2003) follow this approach, but they express most relationships in the form of weighted estimates. They stipulate a function *L* for calibration estimation in the following form:

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

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

$$g = 1_n + \hat{V}^{-1} \hat{X}^{(0)} (\hat{X}^{(0)'} \hat{V}^{-1} \hat{X}^{(0)})^{-1} (X 1_N - \hat{X}^{(0)} 1_n) \quad (4)$$

Note that (4) employs the matrix of weighted characteristics, $\hat{X}^{(0)}$. Using the standard argument in the literature, Bankier and Janes 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

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

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 (3), thus emphasizing the connection between regression and calibration.

The mathematically equivalent expressions in (5) provide two characterizations of the regression estimator. 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.

Using somewhat different notation, Bankier, Rathwell, and Majkowski (1992) introduce regression estimation in a similar manner, again using the weighted matrix, $\hat{X}^{(0)}$. Their approach to the mathematical exposition is advantageous with respect to describing their two-step regression estimator in the Canadian census, but it presents some challenges in relating their exposition to other important papers in the literature. For this reason, both approaches have been presented here.

Although the matrix $\hat{V}$ in (3) should be symmetric and positive definite, mathematical considerations by themselves do not dictate a single choice. For simplicity, the discussion here will be limited to diagonal matrices—examples in the literature tend to be of this form, and a reason to consider more elaborate matrices for the ACS application is not yet evident. The choice $\sigma_i^2 = \sigma^2$ in (2)—equivalent to q_i = 1 in the Deville and Särndal (1992) formulation and $\hat{V} = diag(\pi^{-1}) = diag(W^{(0)})$ in (3), (4), and (6)—gives a model-independent weighting of the sample data to produce a consistent estimator of the unweighted regression in the population. Särndal, Swensson, and Wretman (1992) and Fuller (2002), among other authors, discuss this case separately. In spite of its natural foundation, this approach is not necessarily the method of choice in all situations. Särndal, Swensson, and Wretman (1992, Result 6.5.1, pp. 231-232) describe a wider set of assumptions on $\sigma_i^2 = \lambda' x_i$, for a constant vector λ satisfying $\lambda' x_i > 0$, leading to useful mathematical simplifications.

For a given characteristic *Y*, it is possible to ask what specific value of $\hat{B}$ in (1) would minimize var($\hat{Y}_{rg}$). The *optimal estimator* (Rao 1994), apparently due to Montanari in 1987, is also described by Särndal, Swensson, and

Wretman (1992, pp. 239-242) and by Fuller (2002). Unlike the previous approaches, which yield results consistent with a single set of *g*-weights (4), weights producing the optimal estimator for one characteristic *Y* are likely to be inconsistent with the optimal estimator for another.

Estevao and Särndal (2004) reported both theoretical and empirical results comparing calibration estimation with two general forms of regression estimation: one estimating regression coefficients separately for each domain and one "borrowing strength" by estimating the regression coefficients on the basis of data from the entire sample. They emphasized that all three versions were nearly design-unbiased at the domain level, but their findings point to a distinct advantage to calibration estimation and that borrowing strength is the least likely alternative. These recommendations deserve consideration in the ACS application, but two practical aspects of the ACS application may affect the choice. First, the primary objective is not necessarily to achieve consistency between the ACS sample data and the administrative record data, since the administrative record data will not be published at the geographic detail that will be used in the estimation. Second, the potential application involves relatively small sample sizes that introduce challenges beyond the scope of the Estevao and Särndal work.

6. Lessons from the 1991-2001 Canadian Censuses

To weight their 1-in-5 census sample in 1986, Statistics Canada employed a raking/ratio estimator similar to the estimator used in the U.S. decennial censuses. In 1991, the agency moved to a generalized regression approach, which it refined in applications in 1996 and 2001. A series of papers by Michael Bankier and his colleagues (Bankier, Rathwell, and Majkowski 1992; Bankier, Houle, and Luc 1997; Bankier and Janes 2003) document this work. One of the several parallels between the Canadian and ACS applications is the interest in use of the estimator at very low levels of geographic detail in place of high-level estimates for the overall sample. Over two decades ago, Särndal (1984) argued the potential usefulness of regression estimators for small domain estimation when the sample size was moderately large in the domains.

Many users of the U.S. census long-form data, with its complexities of differing household and person weights, may be surprised to learn that the Canadian censuses since 1991 achieve a single household weight that may be used for all persons in the households. The advantages in interpretability of the estimates are obvious. (Interpretability is one of the six dimensions of quality recognized by Statistics Canada (2002, 2003).)

Several aspects of the Canadian experience are particularly relevant to the ACS estimation problem. Like the U.S., publication areas in Canada include very small areas nested within larger (but still comparatively small) areas. Enumeration areas (EAs), averaging 249 households in 1986, were combined into larger weighting areas (WAs), with approximately 7 EAs in each. Roughly speaking, approximately 50 households would have fallen into the sample in the average EA and 350 households in an average WA. For purposes of comparison, an ACS sampled at 1-in-480 monthly would yield data for roughly 75 households at the block-group level and 200 households at the tract level. (The actual yield for ACS is somewhat less because of subsampling for personal visit followup.) Consequently, the problem of estimation for ACS block groups is approximately as challenging as Canadian EAs, and at the tract level the ACS has roughly half the available sample as a Canadian WA.

The 1986 raking/ratio estimation was implemented at the WA level only. In 1991, a two-step method was introduced to achieve partial calibration of the weights at the EA level followed by relatively standard calibration at the WA level. Weights from the two-step process can be represented by $W_i = g_i\, g_i^{(A)} W_g^{(0)}$, where $g_i^{(A)}$ denotes the first-step adjustment computed at the EA level and g_i denotes the calibration at the WA level. The first-step adjustments $g_i^{(A)}$ were based on averaging *g*-weights from two different EA-level calibrations, each fit to half the desired EA-level constraints. This novel approach brought the EA-level estimates closer to most constraints, although without achieving full agreement. At the WA level, the second step resulted in full agreement with most of the WA constraints. Consistent with previously noted Result 6.5.1 of Särndal, Swensson, and Wretman (1992), Bankier and his colleagues used $\lambda = 1_P$ in the expression $\sigma_i^2 = \lambda' x_i$ to give $\hat{V} = diag(\hat{X}^{(0)'} 1_P)$ for 1991 and 1996.

Bankier, Rathwell, and Majkowski (1992) detailed the approaches used to discard constraints to avoid linear dependence and prevent extreme solutions. For the 1996 census, Bankier, Houle, and Luc (1997) refined the 1991 methods to discard constraints.

In 2001, Bankier and Janes (2003) shifted the approach to what they termed a *pseudo-optimum estimator*, as an approximation to the optimum estimator. They used $\hat{V} = diag(W_i/(W_i - 1))$ in place of the 1991 and 1996 expression and reported evidence that this approach reduced the number of constraints that would be dropped. They also described a "cherry picking" approach to select the best estimator from a set of 10 based on different parameters used in the algorithm to discard constraints.

As should be clear from this review of the theoretical literature and the Canadian census experience, the model-assisted approach encompasses a variety of refinements that will merit investigation. The first goal for ACS is the proof of concept: demonstration that a specific model-assisted estimator can achieve substantial small area improvements. Once possible improvement is demonstrated, then the task of evaluating alternatives for further improvement can

begin.

7. Methods

The goal of the research is to integrate administrative record data into ACS estimation, specifically data from the Master Address File Auxiliary Reference File. For the 36 ACS test counties, an extract has been drawn from this file with characteristics for year 2000. The data set includes demographic characteristics such as age, sex, race, and ethnicity, but omits any sensitive items concerning income or benefits. The data set also does not include individual names or Social Security numbers.

Because the administrative data are compiled to provide an approximation to a census population, the analysis also will compare the performance of administrative record data with 100% data from Census 2000. Both data sets will be evaluated as predictors of the observed ACS data.

In general, variance estimation in the ACS has been implemented through replicate weighting, using 80 replicate weights. The most readily available replicate weights reflect the final weighting, but replicate weights have also been created for some intermediate steps. With the assistance of ACS staff, I have linked together data giving

1. Final ACS household and person characteristics;
2. Data for deletes and other cases with positive weights at estimation step 10;
3. The weights and replicate weights for estimation step 10;
4. Census geography down to the tract and block-group level;
5. An indicator of whether the ACS housing unit appeared in the final Census on the basis of the MAF 12-digit ID;
6. For ACS cases matching Census 2000 MAFIDs, basic housing and person characteristics;
7. An indicator of whether the ACS housing unit can be linked to the administrative data in 2000 through a MAFID; and
8. For ACS cases linked to administrative data, person characteristics from the administrative records.

8. Initial Exploration

The goal, variance reduction for ACS small area estimates, cannot be achieved through regression estimation unless at least a moderately strong regression is obtained for some important ACS characteristics. Fortunately, initial results are quite promising, and this section will summarize some of them.

One task of ACS estimation is to estimate the total valid housing units. MAF counts at the tract and block-group level are readily available, but some units on the MAF are not valid housing units. Invalid housing units are only ascertained during the CAPI field followup in the third month.

Many estimates of totals for characteristics will be correlated with total population. A second estimation task of interest is to estimate the total population at the tract or block-group level.

Preliminary unweighted regressions were fit to ACS data for valid housing units and for population. The two regressions used a common set of census predictors:

1. An intercept term
2. An indicator if not matched to the census
3. If matched, an indicator if occupied
4. The census number of persons

and administrative data predictors:

1. An intercept term
2. An indicator if administrative record persons present
3. The administrative record number of persons

The census has the advantage of one more available predictor, because it is possible to distinguish addresses on the MAF matched to a census vacant housing units from unmatched ones. The administrative data in a single year does not distinguish vacant from non-existent. (In the future when more than one year of administrative data will be available, whether a unit in the MAF was matched to the administrative persons in a previous year may help to predict whether a unit is valid in a given year.)

Table 2 gives R^2 values for the preliminary unweighted regressions to predict valid housing units. Comparison of the first two columns shows that the indicator variable for matched to the census contributes substantially to the prediction. The performance of StARS variables does not reach census levels but is respectable. The last column shows that the addition of StARS data adds little to the predictions from census data alone. In 2001, the ACS sample was drawn from an updated MAF based on 2000 census results, possibly accounting for the higher R^2, .194, in 2001 than the .189 in 2000 for the first regression. In the other 3 regressions, R^2 values peak in 2000.

Table 2 R^2 values from unweighted fit to ACS valid housing unit status in 36 test counties, 1999-2001.

	4-var. census[a]	3-var. census[b]	3-var. adrec[c]	6-var. cen+adrec[d]
1999-2001	.178	.116	.076	.182
1999	.160	.105	.070	.162
2000	.189	.131	.086	.194
2001	.194	.112	.071	.199

Note: [a] 4-variable census regression using 1) an intercept term 2) not matched to census 3) occupied 4) # persons.
b 3-variable census regression using 1) an intercept term 2) occupied 3) # persons
c 3-variable adrec regression using 1) an intercept term 2) occupied 3) # persons
d 6-variable census+adrec regression using 1) an intercept term 2) not matched to census 3) census occupied 4) # census persons 5) Adrec occupied 6) # Acdrec persons.

Table 3 gives R^2 values for similar regressions predicting ACS numbers of persons. The R^2 values of .5 or better suggest that variances for tract and block group estimates of the number of persons could be reduced by about half, a highly encouraging result. The predictions are strongest in 2000, the reference year for both the census and administrative data, but they continue to be strong in the adjacent years. Again, census data are better predictors than the administrative data, and administrative data add little to the census predictions. Nonetheless, with an R^2 of .522 in 2000, the administrative data are as successful a prediction of the current year as the census data are of an adjacent year.

Table 3 R^2 values from unweighted fit to number of ACS persons in 36 test counties, 1999-2001.

	4-var. census	3-var. census	3-var. adrec	6-var. cen+adrec
1999-2001	.556	.554	.475	.573
1999	.537	.536	.485	.560
2000	.638	.636	.522	.654
2001	.492	.492	.416	.508

9. Discussion

Regression estimation shows considerable promise as a means to improve the ACS small area estimates in a "nearly design-unbiased manner."

Although the highest priority for this research is to improve the estimation methods for three- and five-year averages from the ACS, the resulting data sets may provide a basis for research on unit nonresponse. Fuller (2002) reviews the use of regression estimation as an approach.

Note: (1) This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau. I wish acknowledge assistance from several colleagues, including Stephen Ash, Mark Asiala, James Farber, and Donald Luery.

References

Alexander, C.H. (2002a), "Still Rolling: Leslie Kish's 'Rolling Samples' and the American Community Survey," *Survey Methodology, 28,* 35-41.

Alexander, C.H. (2002b), "A Discussion of the Quality of Estimates from the American Community Survey for Small Population Groups," unpublished document available at http://www.census.gov/acs/www/Downloads/Bibliography/THE%20QUALITY%20OF.doc (on 26 Apr 2005).

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.

Estevao, V.M. and Särndal, C.-E. (2004), "Borrowing Strength Is Not the Best Technique Within a Wide Class of Design-Consistent Domain Estimators," *Journal of Official Statistics, 20,* 645-669.

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

Gage, L. (2004), "Comparison of Census 2000 and American Community Survey 1999-2001 Estimates, San Francisco and Tulare Counties, California," unpublished report available at http://www.census.gov/acs/www/AdvMeth/acs_census/lreports/gage.pdf.

Hough, G.C. Jr. and Swanson, D.A. (2004), "The 1999-2001 American Community Survey and the 2000 Census Data Quality and Data Comparisons, Multnomah County, Oregon," unpublished report dated March 9, 2004, available at http://www.census.gov/acs/www/AdvMeth/acs_census/lreports/hough_swanson.pdf .

Judson, D. (2000), "The Statistical Administrative Records System: System Design, Successes, and Challenges," draft paper downloaded 11 July 2005, from http://www.niss.org/affiliates/dqworkshop/papers/judson-background.pdf.

Rao, J.N.K. (1994), "Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage," *Journal of Official Statistics, 10,* 153-165.

Salvo, J., Lobo, P., and Calabrese, T. (2004), "Small Area Data Quality: A Comparison of Estimates, 2000 Census and the 1999-2001 ACS, Bronx, New York Test Site" unpublished report dated March 5, 2004, available at http://www.census.gov/acs/www/AdvMeth/acs_census/lreports/SalvoLoboCalabrese.pdf

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), "Comparison of American Community Survey and Census 2000 Long Form Variance Estimates," paper to be presented at the Joint Statistical Meetings, Minneapolis, MN, 7-11 August 2005.

Statistics Canada (2002), Statistics Canada's Quality Assurance Framework, Catalogue No. 12-586-XIE, http://www.statcan.ca/english/freepub/12-586-XIE/free.htm.

_____ (2003), Statistics Canada Quality Guidelines, Fourth Edition, October 2003, Catalogue No. 12-539-XIE, http://www.statcan.ca/english/freepub/12-539-XIE/free.htm

U.S. Census Bureau (2002), "Specifications for Weighting the 2001 ACS HU Sample (ACS-W-4B)," unpublished memorandum from Charles Alexander, Jr. to James Lewis, prepared by Mark Asiala and Gregg Diffendal, dated November 26, 2002.

_____(2004a), "American Community Survey: A Handbook for State and Local Officials," available at http://www.census.gov/acs/www/Downloads/ACS04HSLO.pdf (downloaded 10 May 2005).

_____ (2004b), "Overview of ACS Supplemental Survey Design and Weighting Description," unpublished memorandum from David L. Hubble to Rajendra P. Singh, dated July 14, 2004.

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.