

Catalogue no. 11-522-XIE

**Statistics Canada International
Symposium Series - Proceedings**

**Symposium 2006 :
Methodological Issues in
Measuring Population Health**



2006



**Statistics
Canada**

**Statistique
Canada**

Canada

A Case Study in Using Model-Assisted Estimation to Integrate Survey and Administrative Data

Robert E. Fay¹

Abstract

This paper reports research to introduce model-assisted estimation into the American Community Survey (ACS), a large-scale ongoing survey intended to replace the long-form sample in the U.S. decennial censuses. The proposed application integrates information from administrative records into ACS estimation. The approach to model-assisted estimation restricts the use of the administrative records to adjustments to the survey weights, while retaining the data on characteristics reported by respondents in the ACS. Although the ACS is a general-purpose survey not specifically tied to health, this case study may suggest possible methodological applications in areas of health statistics.

KEY WORDS: American Community Survey; Small area estimation; Calibration estimation.

1. Introduction

The year 2005 marked the beginning of full production for the American Community Survey (ACS) in the United States. The survey replaces the long form of the U.S. decennial census, the equivalent of the long questionnaire in the Canadian census. The ACS is an ongoing survey of the U.S. population, designating approximately a 1-in-480 sample of housing units every month. The content is similar to that collected on the 2000 long form. Accumulated over a year, the designated sample of 1-in-40 yields about 3 million housing units in the U.S. The ACS has already been used to produce 2005 estimates for states, counties, and places with populations of 65,000+. Accumulated over five years, the ACS data for 2005-2009 will represent approximately a 1-in-8 sample of all U.S. housing units, and will be the basis of estimates to the same geographic detail as Census 2000.

The ACS achieves a qualitative advantage over the U.S. decennial long form by collecting the data with more experienced interviewers and obtaining generally higher rates of response. Because of cost constraints, however, the 1-in-8 designated sample over a 5-year period falls short of the overall 1-in-6 rate of the decennial census. Furthermore, the ACS selects a subsample of initial nonrespondents for personal visit, so the realized ACS sample falls short of the 1-in-8 designation. Consequently, the Census Bureau offers the ACS as a reasoned alternative to the decennial census long form, but it acknowledges that ACS estimates will have higher sampling error than recent censuses.

The research effort for ACS included a national test during 1999-2001 at a much lower sampling rate than the full-scale implementation. Since 1999, however, a version of the ACS has been conducted in 36 test counties (out of the more than 3,000 counties in the U.S.) at approximately the same sampling rates as the full-scale implementation. The data from these counties provide a basis to test the performance of ACS methods for small geographic areas, including estimation methods and their resulting variances.

A few years ago, Paul Voss and his colleagues discovered that sampling errors on ACS subcounty estimates in the test counties were larger than anticipated, even when sample size was taken into account. Their observations were further replicated (Starsinic, 2005; Fay, 2005a). Unlike the Census 2000 long-form, which applied raking-ratio estimation within relatively small subcounty weighting areas, ACS estimation used population controls only at the county or higher geographic levels. This finding, combined with the somewhat smaller sampling rates in the ACS relative to the census, heightened the importance of improving ACS estimation for subcounty areas, if possible. This

¹ U.S. Census Bureau, 4700 Silver Hill Road, Washington, DC U.S.A., 20233 (robert.e.fay.iii@census.gov).

paper represents the fourth in a series reporting on research to improve the ACS estimates (Fay 2005a, 2005b, 2006). The basic strategy includes the use of administrative records with model-assisted estimation.

A few rationales may be offered for including this case study in a conference on measuring population health. Although not primarily a health survey (it includes a small set of disability items), the ACS will offer geographically detailed demographic and economic data potentially useful to many epidemiological and other analyses of population health. Second, the methods described here to combine administrative and survey data through model-assisted estimation may find use in other applications, particularly where auxiliary administrative data on health is available.

In addition to its connection with research on the use of administrative data, the paper is tied to two other broad themes: the considerable contributions of Canadian researchers in developing and applying model-assisted estimators to large-scale problems, and the connection of this work to small area estimation.

The next section further introduces the ACS and the estimation problem. The third section describes the use of one form of model-assisted estimation, generalized regression estimation (GREG), particularly in recent Canadian censuses. The fourth section summarizes encouraging results from the empirical studies thus far. The final discussion section suggests how the methods could be applied to other problems, including health surveys.

2. The Estimation Problem

2.1 The American Community Survey

As already noted, the monthly 1-in-480 designated sample accumulates into a 1-in-40 designated sample in one year, a 3-in-40 sample over 3 years, and a 1-in-8 sample over 5 years. Current publication plans are for 1-year period estimates for states, counties, and places of population over 65,000; 3-year period estimates for counties and places over 20,000; and 5-year period estimates for the same geographic detail as Census 2000. Census 2000 published estimates for tracts (areas of roughly 4,000 persons) and block groups (roughly 1,500 persons). Census 2000 also published estimates for places, many of which were quite small, and other units of local government.

ACS publication of 1-year estimates for 2005 has already begun, but the first 3-year period estimates for 2005-2007 will appear in 2008, and the first 5-year for 2005-2009 will appear in 2010. The complex design of ACS continues to pose challenges, however. Ongoing work includes developing methods to help users interpret 1-, 3-, and 5-year estimates. Users will also have to develop approaches to relate these period estimates to statistics from other sources, such as health data.

2.2 Variances of subcounty estimates

Of the 36 test counties in 1999-2001, 34 were sampled at rates of 3% (5 large counties, including San Francisco, California and Bronx County, New York) or 5% per year (29 counties). Both rates exceeded the 2.5% per year for the full ACS. Over 3 years, the cumulative 9% sample in the large counties approaches the ACS production rate of 12.5% for 5 years, and the 15% sample in the remaining 29 test counties exceeds it. Consequently, the 1999-2001 data for the 34 test counties provide a useful model for the performance of 5-year period estimates for the ACS. (The 2 remaining counties were sampled at about 1% and are excluded from the study.)

As noted in the introduction, the 1999-2001 data from the test counties provided the first warning that ACS tract-level variances were considerably higher than had been anticipated, even after accounting for sample size. Initially, ACS subcounty estimates, such as tract estimates, were expected to exhibit variances comparable to those for the decennial census. But in 2000 the complete count census data, collected on both the short and long forms, provided control totals for ratio-raking estimation in census weighting areas that were typically coterminous with census tracts. There is no directly comparable source of controls for the ACS: The Census Bureau's program of intercensal population estimation is unable to provide satisfactory estimates at this fine geographic level. Both in 1999-2001

and the current ACS production in 2005, intercensal population controls were only used at the county level and higher.

Arguably, estimation in past U.S. censuses is the one estimation problem most similar to the one faced by the ACS. In an attempted analogy between estimation in the census and the small-area estimation goals of the 5-year estimates for ACS, one attack on the problem would be to somehow find a new approach to produce census-like controls at the tract level in order to simulate census long-form weighting. But such an approach is likely to be problematic given previous research experience with intercensal population estimates.

Instead, the approach taken here bears a close relationship to what arguably is the second most similar estimation problem: the Canadian long-questionnaire estimation. Since 1991, Statistics Canada has used generalized regression estimation (GREG) in its census weighting which, like the U.S. census, uses complete count data in the estimation. To adapt the approach to the ACS, administrative data can be substituted for the role of complete count census data in GREG, as will be described here.

3. Model-Assisted Estimation

3.1 General theory

Generalized regression estimation is part of a larger class of model-assisted estimators. In application, there are frequent overlaps with the related class of *calibration estimators* (Deville and Särndal, 1992). The literature on model-assisted estimation is extensive; among the key references are contributions and reviews by Särndal (1984), Särndal, Swensson, and Wretman (1992), Rao (1994; 2003, ch. 2), and Fuller (2002).

General features of model-assisted estimation are critical to the logic of this application. Unlike model-based estimation, model-assisted estimation is essentially design-unbiased when applied under the conditions to be elaborated here. Under these conditions, the improvement from model-assisted estimation can be measured simply by the realized variance reduction. In general, the model used by model-assisted estimation need not be perfectly true, but substantial variance reduction occurs only if the model is highly predictive.

In his recent book on small area estimation, Rao (2003, ch. 2) reviews model-assisted estimation as a possible form of small area estimation. In his notation, consider a sample, s , of elements, j , from a population, and a set of initial weights, w_j , possibly equal to the inverse of the probability of selection ($=\pi_j^{-1}$) or a similar design-based weight. Initially, estimates of a population total for a characteristic, Y , are estimated by $\hat{Y} = \sum_s w_j y_j$. Population totals for auxiliary data, $X=(X_1, \dots, X_p)^T$ are known, but they are also estimated by $\hat{X} = \sum_s w_j x_j$. The GREG estimator takes the form

$$\hat{Y}_{GR} = \hat{Y} + (X - \hat{X})^T \hat{B} \quad (1)$$

where

$$\hat{B} = (\hat{B}_1, \dots, \hat{B}_p)^T = \left(\sum_s w_j x_j x_j^T / c_j \right)^{-1} \sum_s w_j x_j y_j / c_j \quad (2)$$

for constants $c_j > 0$.

The preceding formula appear to depend on the choice of the characteristic, Y , but GREG can be expressed in the form of an adjustment, $g_j(s)$, to the initial weight, w_j , giving $w_j^*(s) = g_j(s)w_j$, where

$$g_j(s) = 1 + (X - \hat{X})^T (\sum w_j x_j x_j^T / c_j)^{-1} x_j / c_j \quad (3)$$

With this adjustment, the new weights are *calibrated* to the auxiliary data, in the sense that the weighted survey estimate agrees with the known population total,

$$\sum_s w_j^* x_j = X. \quad (4)$$

3.2 Estimation for the Canadian long questionnaire

In 1991, Statistics Canada replaced the raking-ratio estimation used in the 1986 census with GREG estimation, which was progressively refined in 1996 and 2001. Michael Bankier and his colleagues (Bankier, Rathwell, and Majkowski, 1992; Bankier, Houle, and Luc, 1997; Bankier and Janes, 2003) developed the details of the implementation.

The complete counts obtained from the short and long questionnaires constituted the x variables. The basic work unit for census estimation was the weighting area, which was further divided into dissemination areas representing the lowest level of intended publication. In 2001, the weighting areas contained an average of about 1865 occupied private dwellings, and the average dissemination area contained 239 (Bankier and Janes, 2003). In each census, a two-step estimation approach was used combining a first step to achieve approximate calibration of some components of X for each dissemination area followed by a second step to exactly calibrate as many components of X as feasible at the weighting area level.

As an early step in the research, Fay (2005a) reviewed the Canadian methods and results in more detail to extract general principles potentially useful to the ACS problem.

3.3 Transplanting the methods to ACS

Briefly, the estimation for the Canadian census can be mapped onto the ACS problem through the following connections:

1. Replace the complete count X by administrative data for the same year as the ACS collection, including determining the x variables for the ACS sample cases.
2. Calibrate the weights to X , but do not publish statistics directly from the administrative data.
3. Insert the GREG estimation early in the estimation process, without altering any of the subsequent ACS weighting steps at the county level.
4. Expect GREG to improve the variances of many ACS estimates \hat{Y} at the tract level.

A critical feature in this application is the importance of the ACS frame: the Master Address File (MAF). The MAF is an ongoing inventory of all housing units in the U.S. It is periodically updated for changes in the housing unit inventory. The MAF was updated several times in preparation for Census 2000, and then updated after Census 2000 to incorporate the census results. Housing units are assigned a permanent MAFID stable across all versions. In sampling for ACS, two different versions of the MAF are used as frames each year in order for the ACS sample to closely reflect the current housing universe.

The administrative data used in this project are one of the products from a multi-year effort at the Census Bureau to systematize the processing and use of administrative data from multiple sources, including the Social Security Administration and the Internal Revenue Service. With a success rate approaching 80-90%, the administrative data are linked to the MAF. The data present a census-like picture of the composition of each household, with known or imputed values for the age, sex, race, and ethnicity of each member. The files used in this project do not contain names, Social Security numbers, any income-related data, or even an indication of whether the data came from a tax return or another source instead.

In this project, the administrative data that are successfully linked to the MAF/ACS frame are used as the source of the unweighted total, X , in eq. (1), and similarly the weighted sample cases for the interviewed ACS cases are the basis of \hat{X} . The GREG step is inserted just after the non-interview adjustments into the sequence of estimation steps. Because ACS response rates are quite high, it is possible to argue that \hat{X} is effectively an unbiased estimate of X . The remaining ACS estimation steps then follow, so the inclusion of the GREG step does not prevent the final ACS weights from agreeing with previous controls at higher-level geographic areas such as counties.

The basic premise behind the variance reduction strategies is that the administrative data provide a statistical prediction, not an exact determination, of the demographic composition of the ACS households. The linkage is done at the MAFID level, and there is no attempt to determine whether specific persons on the administrative file match the reported characteristics in the ACS households. Variance reduction does not depend on consistent matching, but it does require that the administrative data furnish an effective statistical prediction of the demographic composition of the household.

4. Preliminary Results

4.1 Data and Methods

The results presented here summarize findings presented previously (Fay 2006). As previously noted, the 1999-2001 data in 34 ACS test counties were treated as the approximate equivalent of 5 years of ACS production data. In the smallest tracts (< 300 housing units), the non-interview weights were left unadjusted (that is, $g_j(s) = 1$). GREG estimation was carried out separately within each of the other 2250 tracts.

In the intended future application, administrative data for the corresponding years will be used. Ideally, the test should have incorporated administrative data for 1999-2001, but only administrative data for 2000 was readily available for this preliminary study. (In fact, a comparable administrative record file for 1999 does not exist.) As emphasized in the previous section, variance reduction depends the success of the administrative data as a statistical prediction; in this case, the statistical prediction required extended to the ability of the administrative data to predict both one year in the past and one in the future as well as for the same year.

For this study, the choice $c_j = 1$ was made for eq. (2) and (3) without investigating alternatives. In each of the tracts where GREG was implemented, the x variables included a vector of 1's, ensuring that the adjusted weights, $w_j^*(s) = g_j(s)w_j$, would be calibrated to the tract-level total of units in the sampling frame. Up to 7 age-sex variables were included (persons 0-17, 18-29, 65+; males 30-44 and 45-64; and females 30-44 and 45-64). As detailed in Fay (2006), reduced age-sex groupings were used when collapsing was necessary. Similarly, race/ethnicity variables for Hispanic, non-Hispanic Black, and other races (including Asian and American Indian) were available and included when sufficient diversity supported their use. Collapsing rules were based on requirements that the inversion in eq. (2) and (3) not be singular and that no negative weights be computed. Minimum sizes were also specified to support separate race or ethnicity variables.

Two simpler alternatives were considered: one based only on a constant term, and a second excluding race or ethnicity variables. The constant term regression is equivalent to a ratio estimate to the frame total for the number of housing units in the tract, approximating an alternative previously investigated by Starsinic (2005). The intermediate regression evaluates the separate contribution of the race and ethnicity variables.

4.2 Results

Variance reductions were assessed by comparing variance estimates summed over tracts, t , that is, $\sum_t Var(\hat{Y})$ compared to $\sum_t Var(\hat{Y}_{RG})$. Table 1 shows the results separately for the 5 3% counties and the 29 5% counties.

Variances were estimated through replication. For the GREG step, the choice of x variables in a given tract was held fixed, but all other aspects of the GREG adjustment were replicated.

Table 1 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% per year rate. The remaining 29 counties were sampled at 5% per year. All estimated variances are for the estimated totals of each characteristic.

	3% counties			5% counties		
	Const. term regr.	Age/sex regr.	Age/sex, race/ethn regr.	Const. term regr.	Age/sex regr.	Age/sex, race/ethn regr.
Housing units	90	91	91	87	88	88
Occupied hu's	69	74	74	62	69	69
Total persons	47	67	68	42	66	66
Males 0-17	13	39	40	11	40	41
Females 0-17	13	39	40	11	41	41
Males 18-29	10	26	27	9	26	26
Females 18-29	11	28	28	10	28	28
Males 30-44	14	40	39	11	42	42
Females 30-44	17	46	46	13	47	47
Males 45-64	8	43	43	6	47	47
Females 45-64	11	45	46	6	49	49
Males 65+	1	25	25	-2	31	31
Females 65+	3	29	29	-2	35	35
Hispanic	21	33	50	23	39	48
Non-Hisp Black	22	33	46	16	32	43
Non-Hisp White	26	43	51	24	45	51
Other races	10	30	44	9	19	26

Source: Fay (2006)

The results in Table 1 for the constant term regression replicate the earlier finding of Starsinic (2005) that a simple ratio adjustment of the survey weights to the housing unit totals at the tract level would improve estimates, particularly of housing unit characteristics. The more complex GREG forms make only modest improvements for housing variables. For demographic variables, however, the variance improvements from the GREG step substantially exceed those from the simple ratio adjustment.

The selective addition of race or ethnicity variables from the administrative file into X improves the overall variance of the ACS estimates for these groups within tract. Variances for the other characteristics studied remain essentially unchanged.

Finally, results are quite similar for the 3% and 5% counties, suggesting that the variance improvements hold stably over a range that includes the sampling rates for the production ACS.

5. Discussion

5.1 Future research

For purposes of application to the ACS, a number of extensions of this research would further address issues of the potential utility of this approach. The study focused on (1) one period, 1999-2001, (2) the administrative records for 2000 only, and (3) tracts as the one subcounty geographic unit of interest. Within the 34 test counties, investigations

are underway now using ACS data for 1999-2005, administrative record data for 2000-2005, and for places, minor civil divisions (such as townships) and other sub-county geography in addition to tracts. For 5-year estimates, the tract would remain the basic weighting area, but expansion of the research to 3- or 1-year estimates employs weighting areas formed from higher-level geography, such as places above a specified size.

The specific implementation of GREG, although quite promising, could doubtless be refined along a number of dimensions, including the selection of variables, how many variables to include in a given situation, and whether an alternative choice for c_j would improve the overall performance. Some variant of the two-step procedure used in the Canadian censuses, which approximately controls some characteristics at the dissemination area level and exactly controls more at the weighting area level, could be attempted for the analogous relationship between block groups and tracts in the U.S. for the 5-year ACS estimates.

5.2 Extensions to other applications

Textbook formulas for the GREG, including those in Rao (2003), state the necessity to know the population totals for auxiliary data, $X=(X_1, \dots, X_p)^T$ as well as the specific auxiliary values, x_j , for the sample cases. Mathematically, knowledge of the individual values of the auxiliary data for the nonsample cases is not required. Practically, however, combining administrative record data with survey data often entails some form of matching. If the matching is imperfect then the auxiliary totals contain values that might not be represented in the sample. This undermines claims of asymptotic unbiasedness.

In this application, the problem was addressed by matching the administrative records to the entire frame (the MAF) to validate the assumptions by constructing the totals $X=(X_1, \dots, X_p)^T$ that could be estimated with the ACS sample.

A similar or equivalent approach could be used with auxiliary data related to health when ambiguities of matching auxiliary data could occur. In many applications, the conditions that support this application—a list frame as the sampling frame and an ability to link most of the auxiliary information to the frame—may not be met. The empirical results here suggest, however, that when favorable circumstances present themselves, the application of GREG may yield substantial variance improvements.

References

- Bankier, M., A.-M. Houle, and M. Luc (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 D. Janes (2003), “Regression Estimation of the 2001 Canadian Census”, *Proceedings of the 2003 Joint Statistical Meetings on CD-ROM, American Statistical Association*, pp. 442-449.
- Bankier, M. D., S. Rathwell, and M. Majkowski (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.
- Deville, J. and C.-E. Särndal (1992), “Calibration Estimators in Survey Sampling”, *Journal of the American Statistical Association*, 87, pp. 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, Arlington, VA, available from www.fcsm.gov/05papers/Fay_IIC.pdf.

- _____ (2006), "Using Administrative Records with Model-Assisted Estimation for the American Community Survey", paper presented at the Joint Statistical Meetings, Seattle, WA.
- Fuller, W. A. (2002), "Regression Estimation for Survey Samples", *Survey Methodology*, 28, pp. 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, pp. 624-631.
- Särndal, C.-E., B. Swensson, and J. Wretman (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.