American Community Survey (ACS) Variance Reduction Of Small Areas Via Coverage Adjustment Using An Administrative Records Match *

Elizabeth Huang[1], Donald Malec[1], Jerry Maples[1], Lynn Weidman[1]
U.S. Bureau of the Census[1]
Washington D.C. 20233

Abstract

In order to reduce variance and correct for coverage of ACS estimates, there is a desire for population control at the tract level. Currently, intercensal population controls are based on "usual residence" and are not available at the tract level, while the ACS produces intercensal estimates at the tract level that are based on "current residence". This project proposes a way to use new controls, obtained by matching the ACS sample to an administrative records file, and then controlling sample estimates of administrative record tract counts to known tract totals. By matching administrative records addresses, this procedure achieves a consistent residence rule between sample and control. To evaluate the effects of coverage error in the administrative records and matching error with the sample, the procedure is applied to the 2000 Census long-form, where the correct population totals are known.

1 Introduction

The American Community Survey (ACS) has been implemented to provide continuous measurement of key U.S. demographic and socioeconomic characteristics previously measured only once a decade via the decennial census long-form. (See U.S. Census Bureau (2003) for more details.)

As with the census long-form, coverage bias can be reduced and precision increased by controlling population estimates to population counts. Due to the unprecedented level of intercensal detail provided by the ACS, population controls comparable to the once-a-decade, short-form census controls are not currently available. Also, the residence rules are different. The American Community Survey is based on a "current residence" rule. With the exception of respondents who will be in a sample address for less than two months, this rule means that respondents are included as residing, essentially, where they are enumerated. Most other population surveys, including the Decennial census, are based on a "usual residence" rule where the "main residence" of the person is counted as their only residence for an entire year. This rule is especially useful in a one-time enumeration, such as the Decennial census, because there is no opportunity to determine seasonal patterns of residence and a respondent could, incorrectly, be counted at a short-term, seasonal residence for the entire year, giving too much weight to seasonal residences.

The following outlines a way to construct new population controls which refer to the same definition of residence. This is possible because the ACS sample can now be matched to an administrative record (AR) file. However, due to matching error between the ACS and the administrative records and due to undercoverage of the administrative records, statistical models are used in order to produce a final estimate. The administrative file used is the STARS 2000 Person Characteristic File (PCF), which was developed by combining and unduplicating a number of administrative record files in an attempt to cover the population of the United States. For more details and a report on an evaluation of administrative record coverage see Farber and Miller (2003).

The proposed method ratio-adjusts to control totals from administrative records in order to correct for coverage error and to reduce variance. It may be worth noting that part of the method is, mechanically, the same as any type of post-stratification. The only difference is that the post-strata membership of the sample is not available until determined by a match to the administrative records.

2 The Method

The method is based on post-stratified estimation. Usually post-strata are based on information already collected in a survey, e.g., a person's demographic class and the geographic area in which they live. In this case, a person's administrative record location (at the tract level) is added to the sampled respondent's record after the sample has already been collected.

Conceptually, the ACS frame and the administrative records frame can be crossed-classified and partitioned in a way to denote persons in both frames with the same or different tract address and, also, persons in one frame but not the other.

The method proposed here assumes that all matched cases are correct. However, it is assumed that the non-matched cases can arise in two different, but indistinguishable, ways. In one way, an ACS respondent could, conceptually, have one or more administrative records but, due to an imperfect matching procedure, a match is never made. In the other way, an ACS respondent may not have an administrative record, because the AR file does not cover the ACS universe, so that there is no match.

By assuming that some records are unmatchable, it becomes impossible to distinguish between ACS records with unmatch-

*This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. The authors would like to thank Deborah Wagner and James Farber for developing and producing the matched 2000 Decennial Census long-form/2000 Person Characteristic File.

able administrative records and ACS records that do not have any corresponding administrative record. Further assumptions must be made or more information needs to be available in order to use the AR data as a control. By adjusting the administrative records for undercoverage, the ACS universe is now nested within the adjusted administrative records. We further assume that the administrative record tracts for the unmatchable ACS records are distributed identically to the matchable ACS records within current residence tract.

For each demographic group of interest, define,X_{IMij} to be the number of persons captured in the ACS frame, with address in tract i that are matched to an administrative record with address in tract j. Define $X_{IUi.}$ to be the number of persons captured in the ACS frame, with address in tract i that are not matched to an administrative record. Defining X_{OMij} and $X_{OUi.}$ to be the corresponding counts of persons out of the ACS frame but having an administirative record, so that $X_{.M.j} + X_{.U.j}$ is the total number of administrative records with an address recorded in tract j. Lastly, define $X_{....}$ to be an independent estimate of the total U.S. population for the demographic group in question. (Although lacking a complete error profile, the annual demographic estimates produced by the U.S. Census will be used for $X_{....}$.)

The control population is constructed as follows: Adjust for undercoverage of the administrative record population using the factor: $(X_{.M.j} + X_{.U.j})\frac{X_{....}}{X_{.M..}+X_{.U..}}$. Using the ACS base weights, obtain an estimate, $\hat{X}_{IUi.}$ of the unmatched ACS records in tract i, and assign them to an administrative record address tract as if they were missing at random within sampled address tract i, i.e. assign the fraction, $\frac{\hat{X}_{IMij}}{\hat{X}_{IMi.}}$ of the $\hat{X}_{IUi.}$ unmatched cases to tract j.

Calling w_{kt} the base weight for person k sampled at time t the sampled persons that match to the AR file receive the new weight:

$$w_{kt}^{'} = \frac{(X_{.M.j} + X_{.U.j})\frac{X_{....}}{X_{.M..}+X_{.U..}}}{\sum_i \hat{X}_{IMij} + \hat{X}_{IUi.}\frac{\hat{X}_{IMij}}{\hat{X}_{IMi.}}} w_{kt}.$$

When an ACS record is not linked to any administrative record tract, an average weight based on the distribution of AR tracts among the matched cases in the same current residence tract is used:

$$w_{kt}^{'} = \sum_j \frac{X_{IMij}}{X_{IMi.}} \left(\frac{(X_{.M.j} + X_{.U.j})\frac{X_{....}}{X_{.M..}+X_{.U..}}}{\sum_i \hat{X}_{IMij} + \hat{X}_{IUi.}\frac{\hat{X}_{IMij}}{\hat{X}_{IMi.}}} \right) w_{kt}.$$

To avoid variability in the weights due to small sample size in tract-level control demographic cells, a collapsed cell procedure similar to that used at the county level for ACS has been implemented. Basically , if the sample size in a demographic cell is less than 10, it is collapsed with a pre-designated "similar" cell, in a hierarchical manner.

3 An Evaluation Using the 2000 Decennial Census Long-Form

As outlined above, the administrative record matching method was constructed to adjust for coverage errors without introducing bias caused by using different residence rules. However, the administrative record matching method cannot be considered completely successful unless its inclusion can provide substantial variance reduction without appreciatively introducing new biases due to matching errors.

Fortunately, a long-form administrative record match file is available for making a comparison between estimates using long-form population controls and estimates using matched administrative record population controls. By matching long-form returns to administrative records, tract level controls based on AR residence can be constructed; estimates and their estimated variances can be made. Note that in this comparison between controlling the long-form to the Census short-form and controlling the long-form to a matched administrative record file, residence rules are consistent between the survey and its respective control. Hence, one can compare variance reduction and biases caused by matching error in isolation of residence rule bias. In tracts where there are few seasonal residences, this comparison should yield information about current residence. This comparison can only evaluate variance reduction and bias introduced by matching to administrative records. This comparison cannot evaluate benefits to providing tract-level coverage adjustment that is residence-definition free because the long-form population estimates are assumed to be unbiased estimates of their short-form totals.

4 Results

Using the long-form file match to the administrative record file, one can obtain estimates of cross residence characteristics using the long-form base weights. At the national level, it is estimated that 27% of the long-form population cannot be matched to administrative records at the individual level. Of those that do match, 86% have their long-form address and their administrative record address in the same census tract, 8% have their two addresses in the same county but in different tracts, slightly more than 3% have their addresses in the state but different counties and slightly less that 3% have their administrative and census addresses in different states.

In this initial analysis, estimates of population totals and their estimated variances are made for selected demographic groups. Since the controls are also for population counts, albeit for totals of administrative records, it is expected that gains in precision will be more apparent for these estimates than for other population characteristics such as income, etc. However, since the actual population totals from the short form are available, the estimates of population allow an estimate of bias.

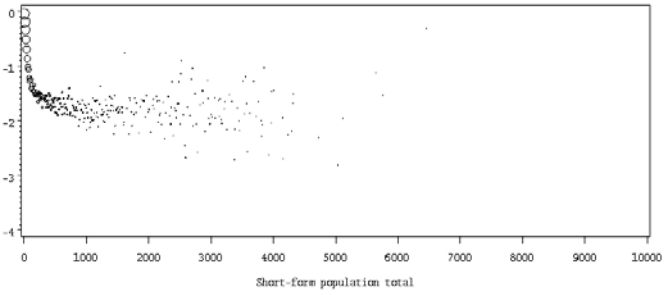
All variance estimates are made by using replicates created explicity for long-form variance estimation (Gbur and Fairchild, 2002). Since the target values for population estimates, say t_0 are available, the value of $(\hat{t}_0 - t_0)^2$ is used as an unbiased estimate of Mean Squared Error (MSE).

The basic estimator proposed in Section 2 will be compared to the same basic estimator that controls to the county level instead of the tract level. This comparison will help determine the effects of controlling below the county level.

Although estimates for total population, sex, thirteen age groups, six race groups and Hispanic origin were evaluated, the results were typically the same, with the smaller groups exhibiting more variablity. For illustration, estimates of total Asian population by tract are presented. The following paragraphs summarize results from each of the census tracts in the U.S. (approximately 65,000 in number). The following plots present median values based data grouped along the x-axis.

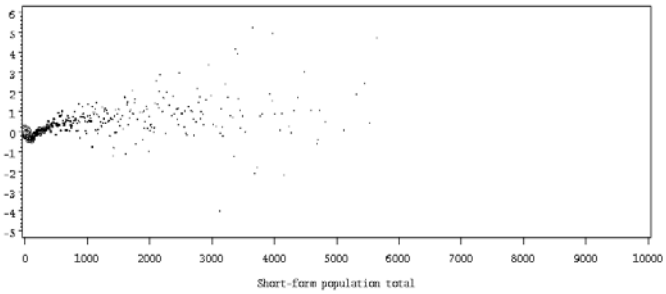
The relative gain, per tract, in terms of variance reduction is measured by the log of the variance of the administrative tract-level control estimator minus the log variance of the administrative county-level control estimator. A negative value indicates that the estimator controlled to the administrative tract total has a smaller variance than the corresponding variance based on the estimator controlling at the administrative county level. These values are plotted against the true population of Asians (obtained form the census short-form) to evaluate the effect of population size on the relative precision of the estimates. As can be seen, using tract level administrative controls can greatly reduce the variance. Over the 53,504 tracts with an Asian present, the median log relative reduction in variance is -1.04.

Figure 1: *Tract Estimate Comparisons of Total Asians vs. Actual Tract Population: Log of Variance of tract-level administrative control estimates minus Log of Variance of county-level administrative control estimates (Note: scatter plot grouped by local median values and bubbles represent relative number of tracts that comprise a median)*



Comparison of mean-squared errors, however, indicates that any savings in precision is mostly lost due to increasing bias in the tract-level controls. Although there is some reduction in MSE error using tract-level controls for some small-size tracts, most estimates indicate little gain. Specifically, over the 53,504 tracts with an Asian present, the median log relative Mean squared error reduction is -.07. In addition, 52% of the tracts had a lower MSE using administrative records controls, representing 41% of the total Asian population.

Figure 2: *Tract Estimate Comparisons of Total Asians vs. Actual Tract Population: Log of MSE of tract-level administrative control estimates minus Log of MSE of county-level administrative control estimates (Note: scatter plot grouped by local median values and bubbles represent relative number of tracts that comprise a median)*



5 Discussion

This initial analysis of a method that uses person-level administrative record matches and tract-level administrative record counts as controls indicates that the tract-level controls do not do any better than county level controls, in terms of lowering mean squared error.

Additional work which may be beneficial would be to control the estimates based on tract-level administrative controls to the corresponding county-level administrative controls. This may be beneficial since the tract-level controls dramatically lowered the variance but had a larger bias (as evidenced by comparable MSE between the tract-level and county-level administrative controls). Another useful comparison would be to look at estimates besides population estimates. Doing so would allow direct comparison between estimates using short-form tract-level population controls with estimates using administrative record tract-level controls. Lastly, more analysis of the results presented in this paper should be attempted with the aim of finding possible causes of the biases in the administrative record tract-level controlled estimates such as the cell collapsing rules and the way the unmatchable cases were handled.

6 References

Farber, James and Esther Miller (2003), "Matching Census 2000 to Administrative Records," Proceedings of the Survey Research Methods Section, 2003, American Statistical Association (Alexandria, VA).

Gbur, Phillip M. and Lisa D. Fairchild (2002). "Overview of the U.S. Census 2000 Direct Variance Estimation."Proceedings of the Survey Research Methods Section, 2002, American Statistical Association (Alexandria, VA).

U.S. Census Bureau (2003),American Community Survey Operations Plan,Release 1.
www.census.gov/acs/www/Downloads/OpsPlanfinal.pdf