RESEARCH REPORT SERIES
(Statistics #2007-9)

Tract-level, Residence-rule Independent, Coverage-error Adjustment of the American Community Survey (ACS) Using An Administrative Records Match

Elizabeth Huang
Donald Malec
Jerry Maples
Lynn Weidman

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: September 17, 2007

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Tract-level, Residence-rule Independent, Coverage-error Adjustment of the American Community Survey (ACS) Using An Administrative Records Match [1]

Elizabeth Huang, Donald Malec, Jerry Maples, Lynn Weidman Statistical Research Division
U.S. Bureau of the Census
Washington D.C. 20233

Abstract In order to reduce variance and correct for coverage of ACS estimates, there is a desire for controlling estimates of demographic characteristics 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 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 their known tract totals. By matching to 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 Census 2000 long-form, where the correct population totals are known from the Census 2000 short-form.

Keywords Tract-level Coverage Error, Administrative Records Matching, The American Community Survey, Estimation Controls

1 Introduction

The American Community Survey (ACS) has been implemented to provide continuous measurement of U.S. key demographic and socioeconomic characteristics previously only measured once a decade in the decennial census long-form. The notion of continuous measurement, in this case, applies to both time and location. Because data collection is based on current residence, the ACS will reflect how the population shifts its residences throughout the year. Due to its large sample size selected from all counties in the nation and monthly data collection, currency and geographic detail of estimates will be achieved. As with the census long-form, coverage bias can be reduced and precision increased by controlling estimates to population counts. The unprecedented level of intercensal detail provided by the ACS is not matched by population controls comparable to the once-a-decade, short-form census. Annual population controls at the tract level are not currently available. Also, residence rules for the ACS and intercensal estimates are different. Data collection for the American Community Survey (ACS) is based on a "current residence" rule. With the exception of respondents who will be at a sample address for less than two months, this rule means that respondents are included as residing, essentially, where they are enumerated. For a continuously fielded survey, as the ACS is, this residence rule will correctly account for persons with multiple residences during the year (e.g. "snow-birds" and "sun-birds" who switch their residences by season), giving these people partial year residences in proportion to the time they are at each residence. 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 "single point in time" enumeration, such as the Decennial census, because there is no opportunity to determine seasonal patterns of residence for which a respondent could, incorrectly, be counted at a short-term, seasonal residence for the entire year,

[1] *Acknowledgments:* The authors are grateful for the assistance of Deborah Wagner and James Farber for providing us with matched files based on our specifications, to Robert Hemmig for assisting us with using the replicate weighting file for the 2000 decennial long form data and to Carrie Simon for assisting us in getting permission to use administrative records. We are also grateful to comments and suggestions provided by Michael Beaghen, Patrick Cantwell and Michael Ikeda on earlier versions of this manuscript. We would also like to thank the members of the Decennial Statistical Studies Divisions' informal Small Areas Variance Estimation Research group for useful discussions at the beginning of this project.

giving too much weight to seasonal residences. Another reason to use a "usual residence" rule in a survey is that, because the Decennial census is based on usual residence, intercensal estimates of population are also based on this rule and can be used as unbiased post-stratification controls.

Besides using a new definition of residence, the ACS also pushes the limits of providing precise estimates for geographic detail down to the census tract level. Even if the ACS were based on the usual residence rule, there are currently no intercensal population estimates available at the census tract level of detail. Because of the additional differences in residence rule between the ACS and population estimates, controlling the ACS to tract level estimates of population may reduce the variance at the expense of introducing bias due to controlling the current-residence population total to the, possibly different, usual-residence population total.

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. A new, third type of residence, defined as where a person's administrative records address is located can be used. By matching the ACS to administrative records (person by person), administrative records residence will be consistent between the ACS and administrative records by design. Using an administrative records residence for controlling the ACS renders population controls to be independent of either usual residence or current residence.

Due to encouraging results reported on the coverage of administrative records of the U.S. population, as presented in Farber and Miller (2003), the use of administrative records as a population control is considered and its viability investigated. Due to both an unknown amount of matching error between the ACS and the administrative records and to undercoverage of the administrative records, statistical models are used in order to produce a final estimate with the aim of evaluating it's resulting bias effects. 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 estimated totals of persons having selected demographic characteristics to adjusted 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.

The problem of how to use independent controls as a way to adjust for coverage bias is the main motivation for this work. A secondary benefit from coverage control is variance reduction. Another method that does not adjust for coverage bias but does use administrative records to reduce variance has been proposed by Fay (2006). Fay uses an administrative record match of household addresses as a source for calculating calibration estimates to achieve variance reduction. Specifically, using the MAF as a complete address frame, only administrative records with an address that matches to the MAF are used as a source of covariates in the calibration estimator. The calibration estimates are not based upon person matches. Instead, housing unit composition is used as a covariate, whether or not the composition reflects the current residents of the household.

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 if a sampled person can be matched to an administrative record, the tract that corresponds to the their administrative record address is added to their sample record after the sample has already been collected.

Conceptually, the ACS frame and the administrative records frame can be cross-classified and partitioned

in a way to denote persons in both frames whose ACS address and administrative record address fall into the same tract, those whose addresses fall into different tracts and, also, persons in only one frame but not the other.

The method proposed here assumes that all matched cases are correct; i.e. the ACS record and AR record refer to the same person. However, it is assumed that the non-matched cases can arise in two different, but indistinguishable, ways. In one way, an ACS respondent record could, conceptually, have one or more administrative records but, due to an imperfect matching procedure, the two types of records never matched. In the other way, an ACS record may not match to 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 an ACS record with unmatchable administrative records and an ACS record that does 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.

Population controls at the tract level will be made for 312 distinct demographic groups (defined in the Appendix). For each demographic group of interest, d, define,X_{dij} to be the conceptual 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 $U_{di.}$ to be the conceptual number of persons captured in the ACS frame, with address in tract i that are not matched to an administrative record. Define $R_{d.j}$ to be, conceptually, the corresponding counts of persons out of the ACS frame but having an administrative record, so that $X_{d.j} + R_{d.j}$ is the observed total number of administrative records with an address recorded in tract j.

Lastly, define $X^*_{d....}$ to be an independent estimate of the total U.S. population for the demographic group in question. (Although lacking an estimate of accuracy or precision, the annual population estimates produced by the U.S. Census Bureau will be used for $X^*_{d....}$.) This adjustment for administrative record undercoverage at the tract level by a national level adjustment raises concern, as it could be that there is still significant coverage error at the tract level. However, there are no other geographic levels to adjust for coverage error of administrative record residence. The success of this assumption, in conjunctions with others, will be evaluated in section 3.

The control population is constructed by adjusting the administrative record population for undercoverage. The factor $\frac{X^*_{d....}}{X_{d..}+R_{d..}}$ is created by ratio adjusting the administrative record totals to equal the independent population estimate for each demographic group at the national level. This factor is applied to the tract level administrative record total $(X_{d.j} + R_{d.j})$ to create the control totals.

After adjusting the administrative records for undercoverage, by assumption the unmatched ACS records are represented in the administrative record count, however their intended administrative record tract is still unknown. The unknown locations are imputed based on the assumption that tract location is missing at random within the ACS tracts. Using the ACS base weights, obtain an estimate, $\hat{U}_{di.}$ 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, $\frac{\hat{X}_{dij}}{\hat{X}_{di.}}$ of the $\hat{U}_{di.}$ unmatched cases to tract j.

Let w_{kt} be the ACS base weight for person k sampled at time t (belonging to demographic group d, census tract i), then the sampled persons that match to the AR file (in AR address tract, j) receive the new weight:

$$w^{'}_{kt} = \frac{(X_{d.j} + R_{d.j})\frac{X^*_{d....}}{X_{d..}+R_{d..}}}{\sum_i(\hat{X}_{dij} + \hat{U}_{di.}\frac{\hat{X}_{dij}}{\hat{X}_{di.}})} \; w_{kt}.$$

When an ACS record is not linked to any administrative record tract, it is not readily apparent as to which population control should be used. Again, based on the missing at random within tract assumption, the expected value (over the distribution of administrative record tracts) is used.

$$w_{kt}^{'} = \sum_j \frac{\hat{X}_{dij}}{\hat{X}_{di.}} \left(\frac{(X_{d.j} + R_{d.j}) \frac{X^*_{d...}}{X_{d..}+R_{d..}}}{\sum_i (\hat{X}_{dij} + \hat{U}_{di.} \frac{\hat{X}_{dij}}{\hat{X}_{di.}})} \right) w_{kt}.$$

To avoid variability in the weights due to small sample size in tract-level demographic control 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 or if the ratio between the control total and the ACS total is too large (exceeds 3.5), it is collapsed with a pre-designated "similar" cell, in a hierarchical manner. Actually, the specifications used for the ACS at the county level (See U.S. Census Bureau (2006) section 11.4. at website: http://www.census.gov/acs/www/Downloads/tp67.pdf) are followed for collapsing cells within census tracts with the exception that six race categories are used: "White", "Black", "Indian", "Asian", "Pacific Islander" and "Multiple Race". These are the most detailed race categories available from the administrative records file. The multiple race response in the ACS sample was assigned to a single race group based on the largest minority race selected in that collapsed estimation strata (tract for this study). Before a single race group is created, any response which contains the category some other race needs to be redefined into one of the five major race groups (White, Black, American Indian and Alaskan Native, Asian, and Native Hawaiian and other Pacific Islander). The single-race response is combined with the Hispanic origin response to form weighting race/Hispanic origin groups (five Non-Hispanic race groups and Hispanics). The collapsing is done by race/Hispanic group first. Within each collapsed weighting race group, the people in sample are divided into 26 age-sex groups. If an age-sex group within a weighting race group does not have at least 10 people in sample or the ratio of the control total to the pre-controlled weighted estimate is not less than 3.5, then it is collapsed with other age-sex groups until cells of at least 10 people in sample and a ratio of less than 3.5 are formed. The basic goal of the scheme is to keep 0-17 children together when possible, men 18-54 together, women 18-54 together, and senior 55 and older together, by sex if possible. For more details on ACS collapsing rules in the weighting scheme see Asiala (2004).

In summary, a consistent residence rule between the survey frame and the control population has been achieved by using an administrative record residence. This is a third type of residence which may be different from either current- or usual-residence.

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 both matching error and coverage error.

Fortunately, a census long-form to 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. The main matching variable used was social security number, in conjunction with variables such as name, date of birth and geography. For more detail, see Farber and Miller (2003). Note that in this comparison residence rules are consistent between the survey (census long form) and its respective controls (using census short-form or

the coverage adjusted matched administrative record file). The long form population control uses the usual residence of the long form and short form. The AR control uses the AR residence obtained by matching the AR records to the long form records. Hence, one can compare variance reduction and biases caused by matching error and coverage error in isolation of residence rule bias. In this comparison, the Census 2000 short-form population totals are the target population. Hence, these totals can be assumed to be the "truth". In fact, the Census short-form totals, by demographic group, are used to control the coverage of the administrative records at the national level. This comparison can only evaluate variance reduction and bias introduced by matching to administrative records and adjusting the administrative records for their own coverage error. This comparison cannot evaluate benefits to providing tract-level coverage adjustment that is residence-definition free for a survey such as ACS that uses different residence rules for data collection and its population controls, because the long-form population estimates are based on the same residence rule as their corresponding short-form totals.

4 Results

Using the long-form file match to the administrative record file, one can obtain estimates of cross-tract 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. However, this geographic distribution of residence will likely be different with an ACS match to administrative records since the ACS uses current residence.

In this 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. In addition, 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 explicitly 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, controlling to adjusted administrative record totals, 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.

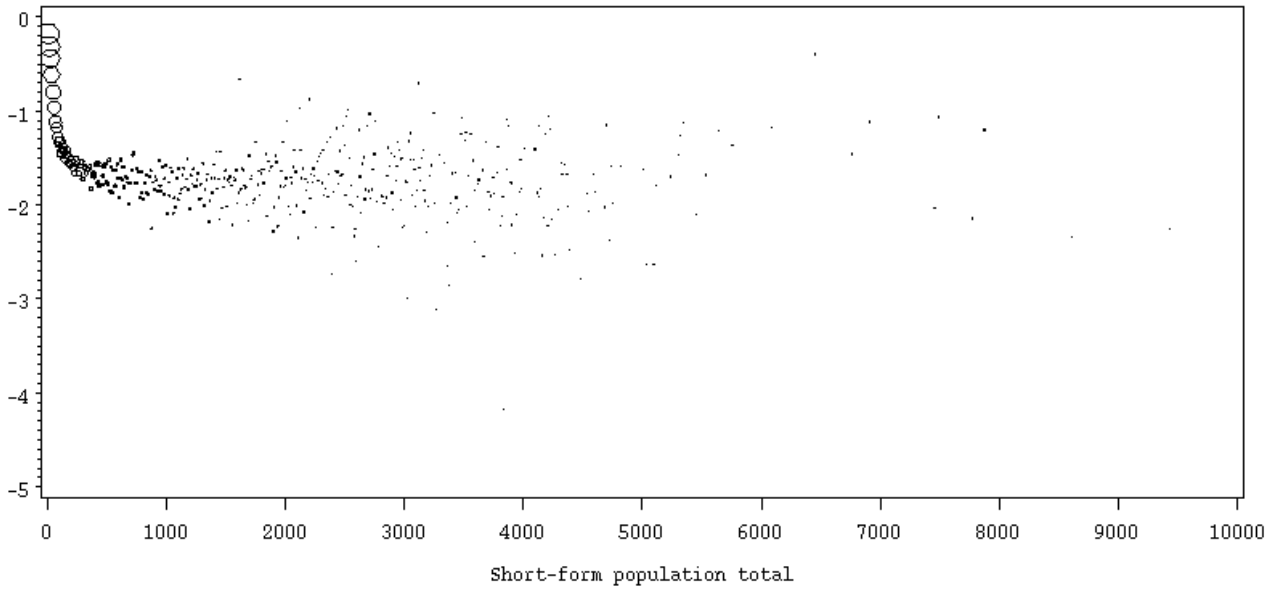
4.1 Results for Asians

Estimates for total population, sex, thirteen age groups, six race and Hispanic origin groups were evaluated, the results were typically the same, with the smaller groups exhibiting more variability. For illustration, estimates of total Asian population by tract are presented first. Only a brief summary of the evaluations for each of the remaining subgroups follows in section 4.2 since the results are basically the same. The plots in Figures 1 and 2 summarize results for Asians from each of the census tracts in the U.S. (approximately 65,000 in number). They present median values for tracts grouped along the x-axis according to population. Specifically, tracts are in the same group if their corresponding short-form population totals are the same, to the nearest ten.

The relative gain, per tract, in terms of variance reduction is measured by the natural logarithm of the variance estimate of the administrative tract-level control estimator minus the natural logarithm of the

variance estimate of the census short-form 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 to census short-form totals at the county level. These values are plotted against the true population of Asians (obtained from the census short-form) to evaluate the effect of population size on the relative precision of the estimates. As can be seen in Figure 1, using tract level administrative controls can greatly reduce variance. The bubbles, whose size reflects the relative number of tracts at a specific x-y coordinate are included to give an impression of how the entire population of tracts is being affected. In summary, over the 53,504 tracts with an Asian present, the median natural logarithm relative reduction in variance is -1.04 (65% reduction of variance).

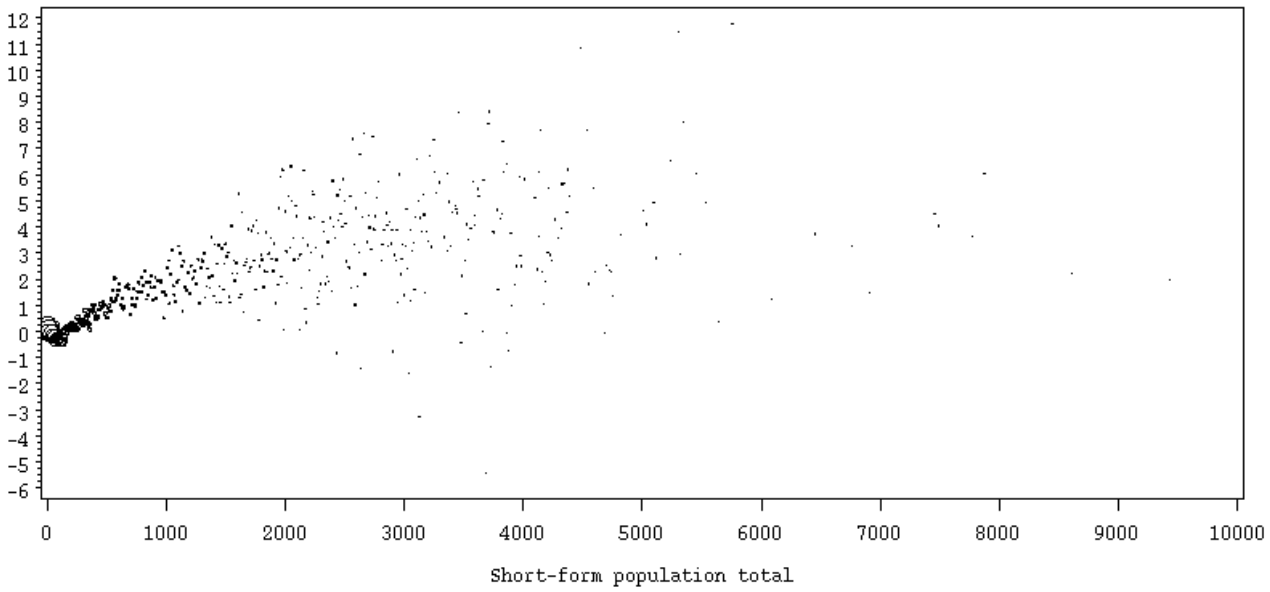
Figure 1: *Tract Comparisons of Estimated Total Asians vs. Census 2000 Tract-level Asian Population: Natural logarithm of Variance of tract-level administrative control estimates minus natural logarithm of Variance of county-level census short-form control estimates (Note: scatter plot grouped by local median values and bubbles represent relative number of tracts that comprise a median)*



Comparison of mean-squared errors in figure 2, however, indicates that any savings in precision is mostly lost due to increasing bias in the tract-level estimates. Although there is some reduction in MSE using tract-level controls for some small-size tracts, most estimates indicate little gain. The bubbles are included to give an impression of how the entire population of tracts is being affected. In summary, over the 53,504 tracts with an Asian present, the median natural logarithm relative reduction in variance is -1.04 (65% reduction of variance).Specifically, over the 53,504 tracts with an Asian present, the median natural logarithm relative MSE reduction is -.07. In addition, 52% of the tracts representing 41% of the total Asian population had a lower MSE using county-level census short-form controls.

The next four figures provide a comparison of the administrative records controls by separately looking at the effects of using administrative records versus census short-form controls and, also, the effects of

Figure 2: *Tract Comparisons of Estimated Total Asians vs. Census 2000 Tract-level Asian Population: Natural logarithm of MSE of tract-level administrative control estimates minus natural logarithm of MSE of county-level census short-form control estimates (Note: scatter plot grouped by local median values and bubbles represent relative number of tracts that comprise a median)*

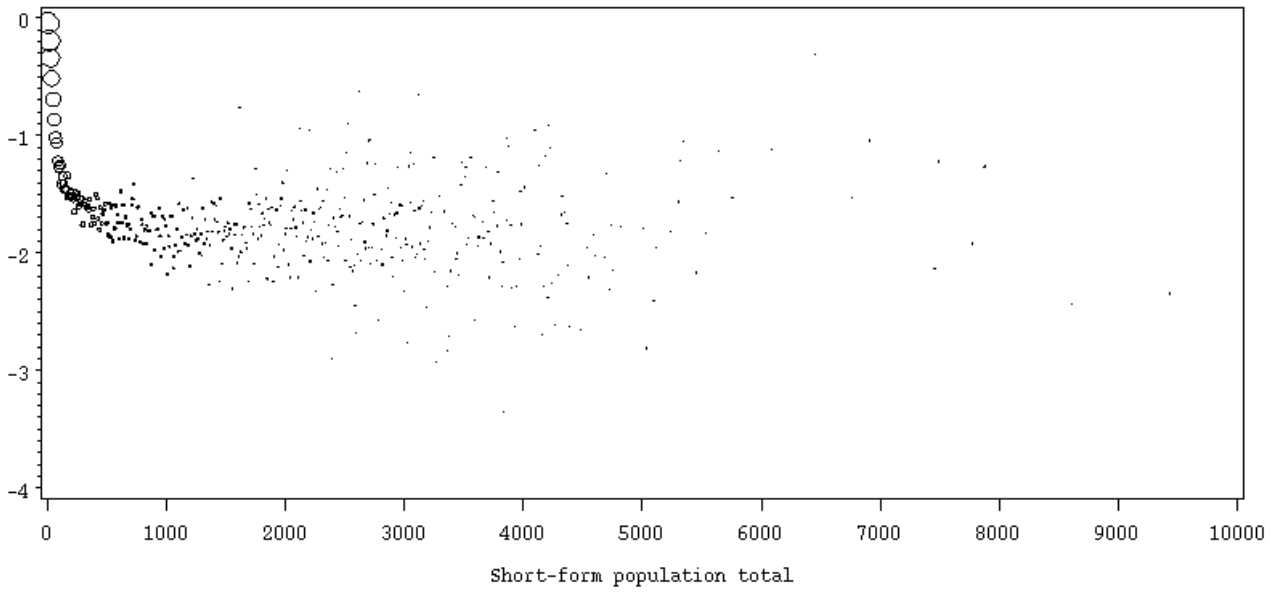


controlling at the tract versus county level. Specifically, figures 3 and 4 provide a comparison of variances and mean squared errors based on using administrative record controls at either the tract level or at the county level. This comparison eliminates differences due to using short-form totals and focuses on gains in controlling at a smaller geographic level.

Figures 5 and 6 provide a look at the effects of using administrative records versus census short-form records as controls at the county level.

As can be seen, the reduction in variance is comparable whether or not control is to tract estimates based on administrative records or on short form totals (Figures 1 and 3). The comparison of MSE based on tract-level control to either administrative record or short-form county level controls yields roughly equal results (Figures 2 and 4) suggesting that carrying down control to the administrative record tract level is the main reason for bias. Some bias still seems to be attributable to just using the administrative records approach instead of controlling to the census short-forms, however. Comparing the MSE error between administrative record and census short-form county control (Figure 6) shows that using the administrative records typically results in a larger MSE. Figure 5 exhibits a greater reduction in variability using administrative record controls for small populations. Since this is not accompanied by a corresponding reduction in MSE, it's cause will not be investigated here.

Figure 3: *Tract Comparisons of Estimated Total Asians vs. Census 2000 Tract-level Asian Population: Natural logarithm of Variance of tract-level administrative control estimates minus natural logarithm 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)*



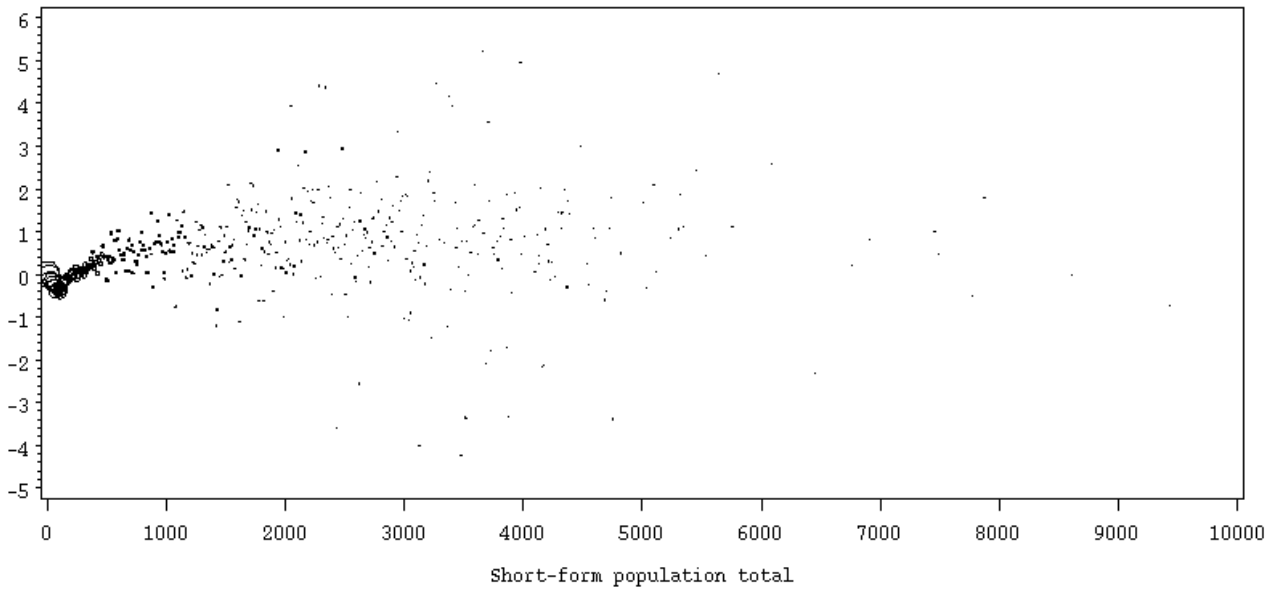
4.1.1 Further Evaluation of Bias for Asians

To help evaluate possible causes of the bias apparent in Figure 2, the relative bias of the estimates ('rbiasm1t' on figures 7 and 9) is specified as the value of $(\hat{t}_0 - t_0)/t_0$, and is plotted against two potential sources of bias: 1) the amount of collapsing (COLRATE) and 2) the amount of nonmatches (imputations) to administrative records (IMPRATE).

The collapsing for each demographic group in a collapsed cell measures the proportion of sample weight in the cell that belongs to individuals that were from a different demographic group. If a small group is collapsed into a larger group, then the members of the smaller group would have a larger collapsing rate, while the members of the larger group would have a smaller collapsing rate. If a cell was not collapsed, then its collapsing rate would be zero. The control cells are based on AR tracts and then treated as an individual level characteristic when evaluating population estimates for census tracts. This metric can indicate if collapsing is smoothing too many different, in terms of control factors, demographic groups together. The imputation rate measures the proportion of sample weight in a control cell belonging to long-form records that were non-matches to the administrative records. This metric can help assess whether the assumption that the AR tract addresses are missing at random within Census tract is problematic.

First, it is necessary to define the proportion of weight in an AR control cell (within an AR tract) that does not have the same demographic characteristics as d (race, Hispanic origin, sex and age group). $\hat{X}_{d.j} = \sum_i (\hat{X}_{dij} + \hat{U}_{di.} \frac{\hat{X}_{dij}}{\hat{X}_{di.}})$ is the estimated total in AR tract j and demographic d. Let d^* be the set of

Figure 4: *Tract Comparisons of Estimated Total Asians vs. Census 2000 Tract-level Asian Population: Natural logarithmLog of MSE of tract-level administrative control estimates minus natural logarithm 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)*



demographic groups collapsed together to form a control cell, so that the proportion of sample weight in the collapsed cell not belonging to demographic group d is

$$c_{dj} = 1 - \frac{\hat{X}_{d.j}}{\sum_{d \in d^*} \hat{X}_{d.j}}. \tag{1}$$

Note that if the control cell had no collapsing, then $c_{dj} = 0$. The collapsing proportions are associated with the individual and are averaged over the sample used to make the population estimates of interest, e.g. for ACS residences in census tract, i, the population estimate of the amount of collapsing for demographic group d is:

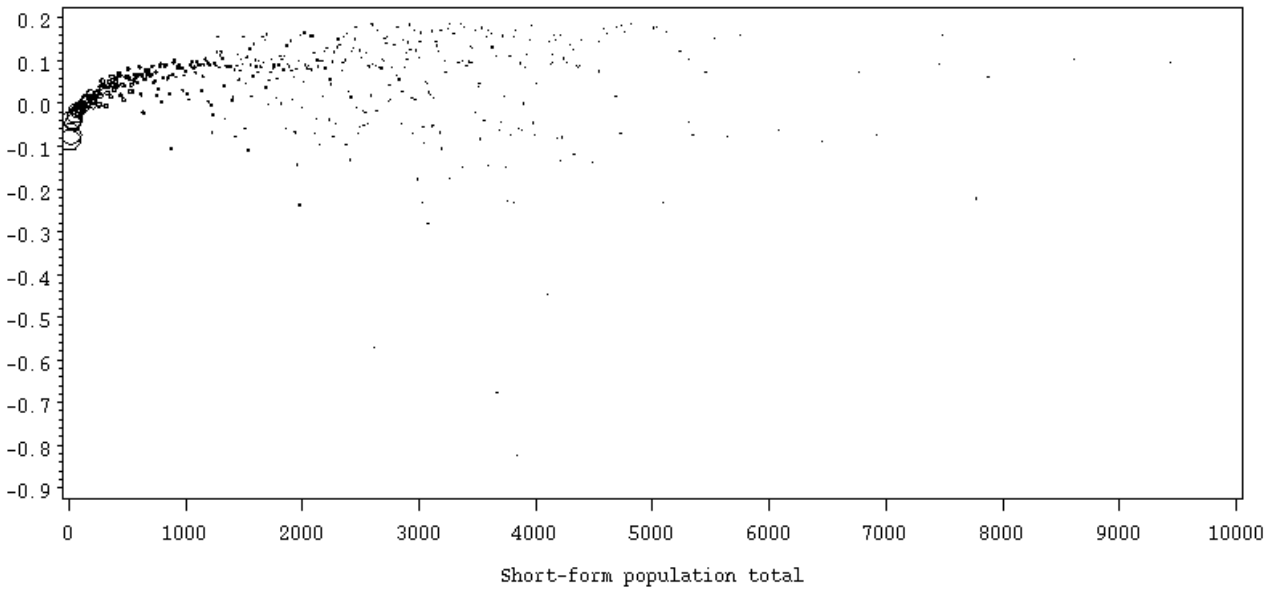
$$\text{COLRATE}_{di} = \frac{\sum_j (\hat{X}_{dij} c_{dj} + \hat{U}_{di.} \frac{\hat{X}_{dij}}{\hat{X}_{di.}} c_{dj})}{\hat{X}_{di.} + \hat{U}_{di.}} = \frac{1}{\hat{X}_{di.}} \sum_j \hat{X}_{dij} c_{dj}. \tag{2}$$

To define the imputation rate for demographic d for AR tract j, let

$$\text{impAR}_{dj} = \frac{\sum_{d' \in d^*} \sum_i (\hat{U}_{d'i.} \frac{\hat{X}_{d'ij}}{\hat{X}_{d'i.}})}{\sum_{d' \in d^*} \hat{X}_{d'.j}} \tag{3}$$

where, again, d^* is the set of demographic groups collapsed together to form a control cell for demographic

Figure 5: *Tract Comparisons of Estimated Total Asians vs. Census 2000 Tract-level Asian Population: Natural logarithm of Variance of county-level administrative control estimates minus natural logarithm of Variance of county-level census short-form control estimates (Note: scatter plot grouped by local median values and bubbles represent relative number of tracts that comprise a median)*



group d. Then, the imputation rate is computed by averaging impAR_{dj} over the sample in the census tract.

$$\text{IMPRATE}_{di} = \frac{(1 + \frac{\hat{U}_{di.}}{\hat{X}_{di.}}) \sum_j \hat{X}_{dij} \text{impAR}_{dj}}{\hat{X}_{di.} + \hat{U}_{di.}} = \frac{1}{\hat{X}_{di.}} \sum_j \hat{X}_{dij} \text{impAR}_{dj} \tag{4}$$

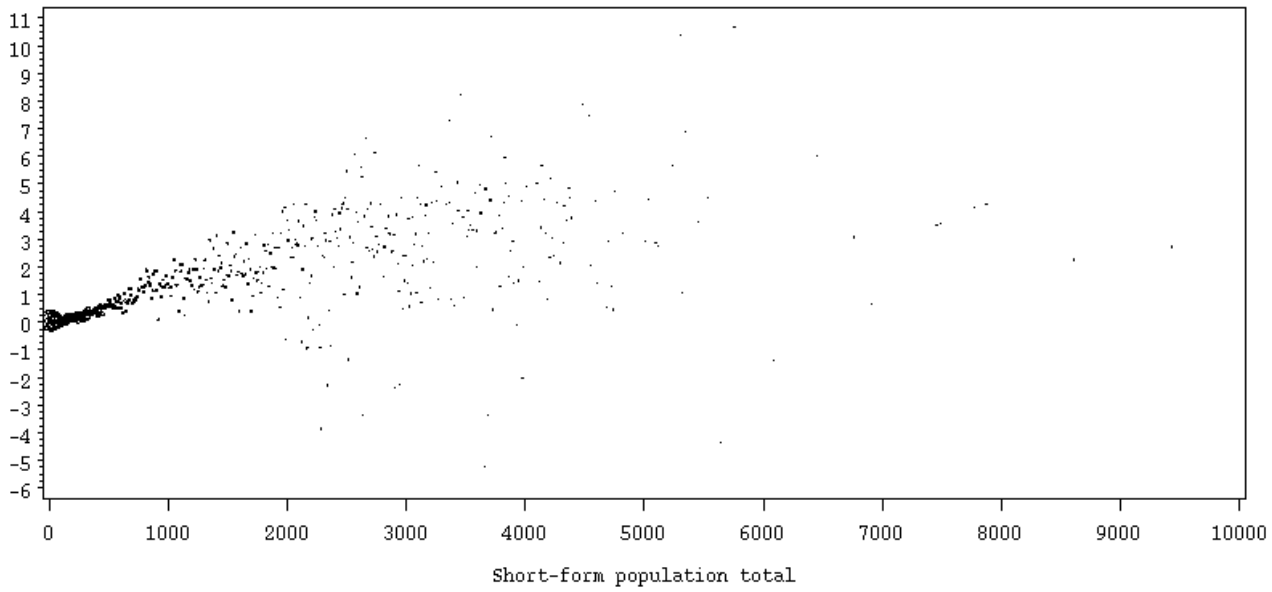
As can be seen in Figure 7, the bias in adjusting estimates of total Asians becomes more apparent only when there is a large amount of collapsing. Also, it can be seen that a large number of tracts have a large amount of collapsing. There does also appear to be a pattern of non-zero bias for tracts with little collapsing. Without further evaluation, it cannot be determined how much is due to variability based on a small number of tracts and how much is a genuine effect.

The amount of collapsing is generally related to the total number of Asians in a tract, as evidenced in Figure 8. However, some relatively large tracts may still contain a fair amount of collapsing.

Figure 9 exhibits a relationship between bias and the proportion of long-form cases that did not match to an administrative record. Unlike Figure 7, where the largest bias is represented by a relatively large number of tracts (as indicated by the size of the bubble), there are fewer tracts where a large proportion of Asians need to be imputed due to administrative records non-matches.

In this case, the relative amount of unmatched cases does not appear to be related to tract size, as seen in Figure 10.

Figure 6: *Tract Comparisons of Estimated Total Asians vs. Census 2000 Tract-level Asian Population: Natural logarithm of MSE of county-level administrative control estimates minus natural logarithm of MSE of county-level census short-form control estimates (Note: scatter plot grouped by local median values and bubbles represent relative number of tracts that comprise a median)*



4.2 Results for Selected Demographic Groups

The following presents a summary of the evaluation for estimates for total population grouped by sex, thirteen age groups, six race groups and Hispanic origin. As mentioned in section (4.1), the results follow, more or less, those for Asians. Figures 11, 13 and 15 provide boxplots summarizing the distribution of relative natural logarithm variance for demographic groups. The center line in each box is the median, the box edges the first and third quartiles and the ends 5% and 95% percentiles. (Unlike, figures 1-6, the following figures summarize individual tracts and are not smoothed by grouped medians.) When they were evaluated, the results were typically the same, with the smaller groups exhibiting more variability. Figures 12, 14 and 16 summarize the differences in MSEs. As with Asians, there is no clear gain in coverage adjustment at the tract level over the county level.

4.2.1 Further Evaluation of Bias for Selected Demographic Groups

Figure 17 shows how the relative bias is affected by both the amount of collapsing and the amount of administrative record imputation. This figure is produced by taking each tract-level demographic group and cross-classifying it by the amount it was collapsed and imputed. [Note: Figure 18 provides a frequency plot of what the imputation rate and collapsing rate is for these tract-level demographic groups]. It can be seen that most do not have the very extreme biases shown in figure 17. Even though all demographic groups are on the following plot, there is a clear pattern that either extreme degrees of collapsing or imputation will

Figure 7: *Evaluation of Tract-Level Bias of Tract-Level Administrative Record Controls for Asians: Estimated bias versus proportion of tract members collapsed. (Note: scatter plot grouped by local median values and bubbles represent relative number of tracts that comprise a median - grouped by the nearest hundredth colrate)*

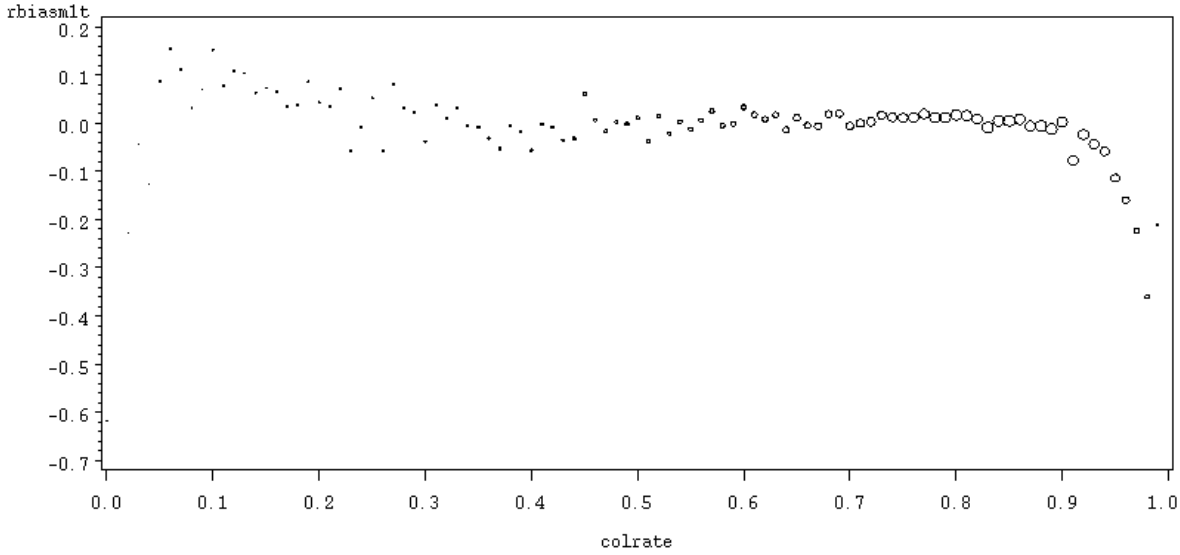


Figure 8: *Evaluation of Tract-Level Bias of Tract-Level Administrative Record Controls for Asians: Proportion of tract members collapsed versus total number of Asians.*

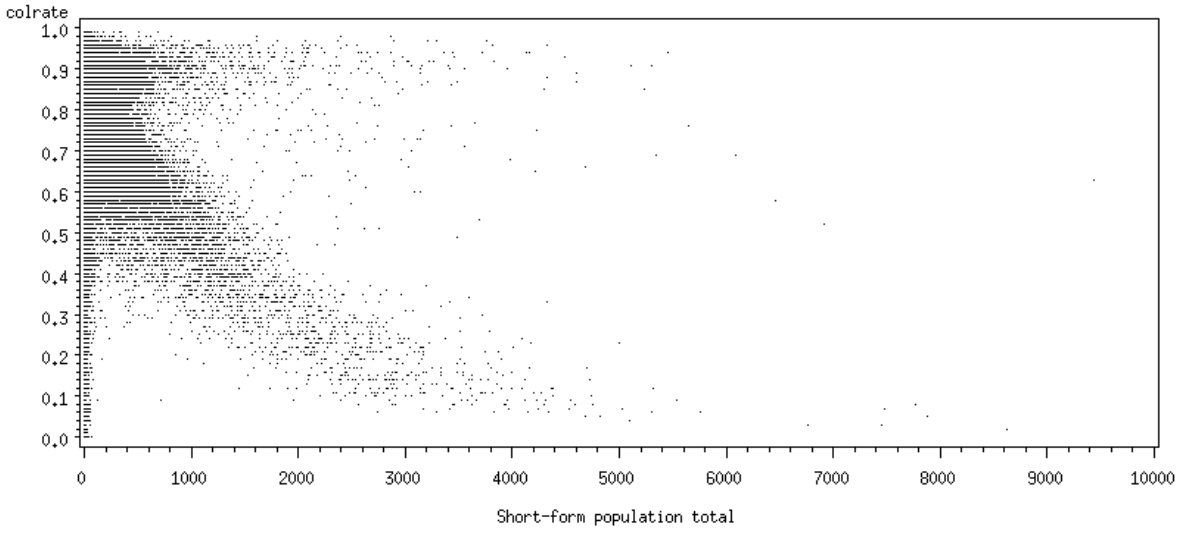


Figure 9: *Evaluation of Tract-Level Bias of Tract-Level Administrative Record Controls for Asians: Estimated bias versus proportion of tract members who do not match to an administrative record. (Note: scatter plot grouped by local median values and bubbles represent relative number of tracts that comprise a median - grouped by the nearest hundredth imprate)*

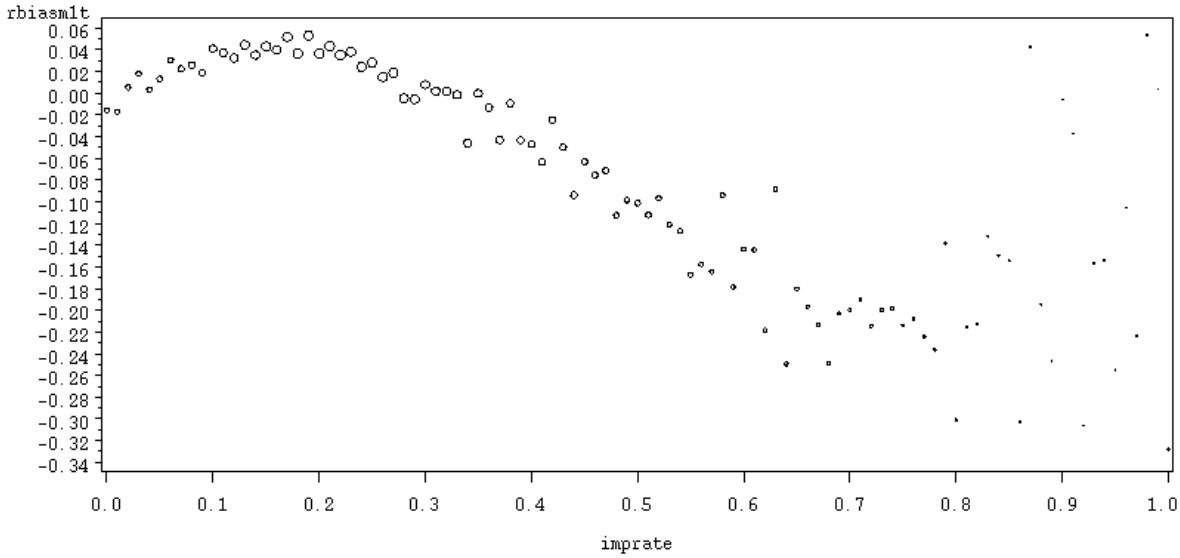


Figure 10: *Evaluation of Tract-Level Bias of Tract-Level Administrative Record Controls for Asians: Proportion of tract members who do not match to an administrative record versus total number of Asians.*

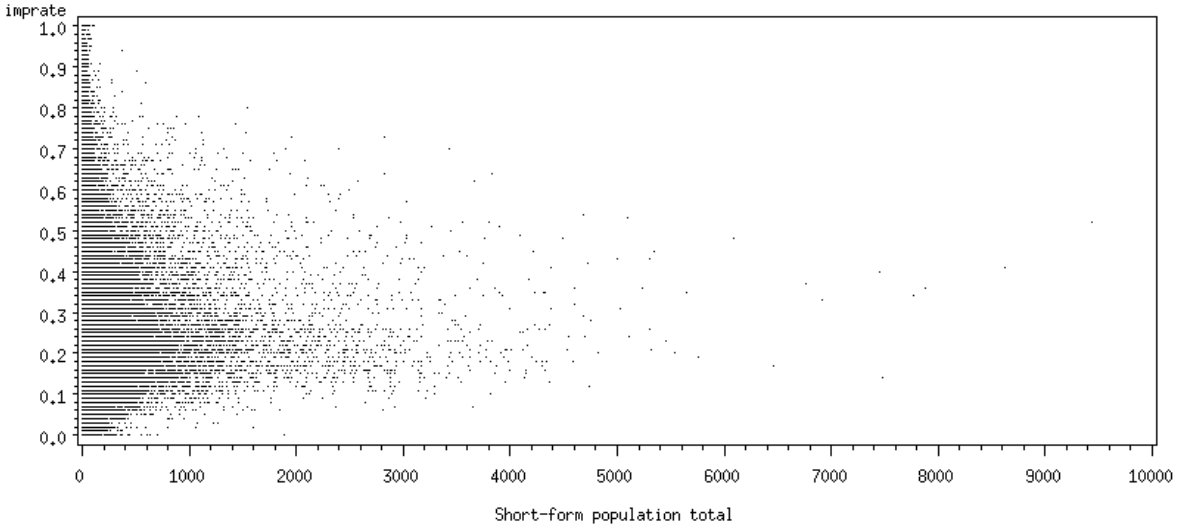


Figure 11: *Natural logarithm of Variance of tract-level administrative control estimates minus natural logarithm of Variance of county-level census short-form control estimates for all tracts for selected demographic groups (Note: distribution of tract differences summarized by boxplots) - see note at appendix for definition of demographic groups*

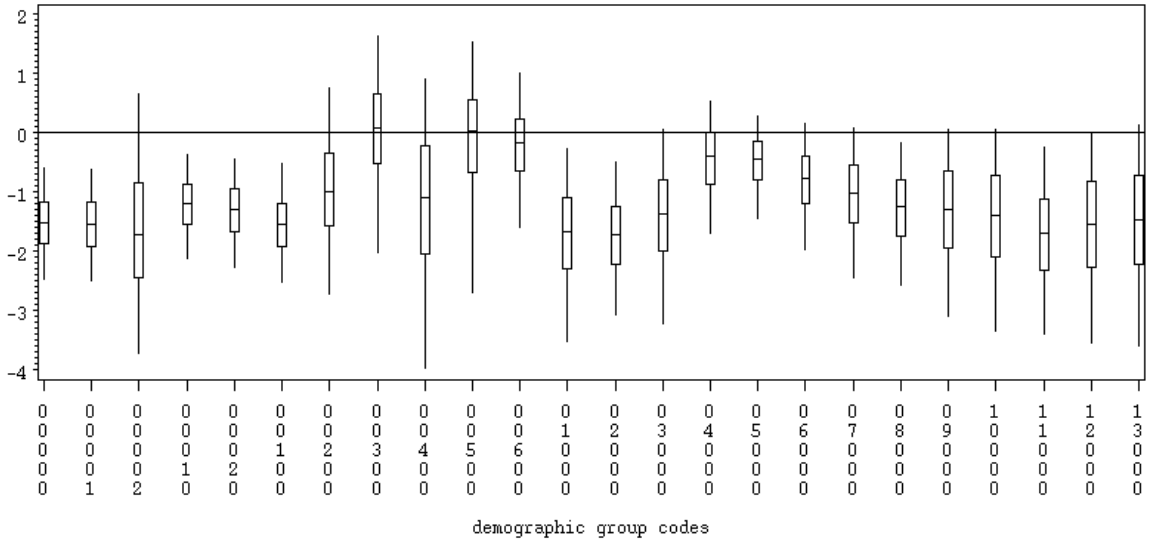


Figure 12: *Natural logarithm of MSE of tract-level administrative control estimates minus natural logarithm of MSE of county-level census short-form control estimates for all tracts for selected demographic groups (Note: distribution of tract differences summarized by boxplots) - see note at appendix for definition of demographic groups*

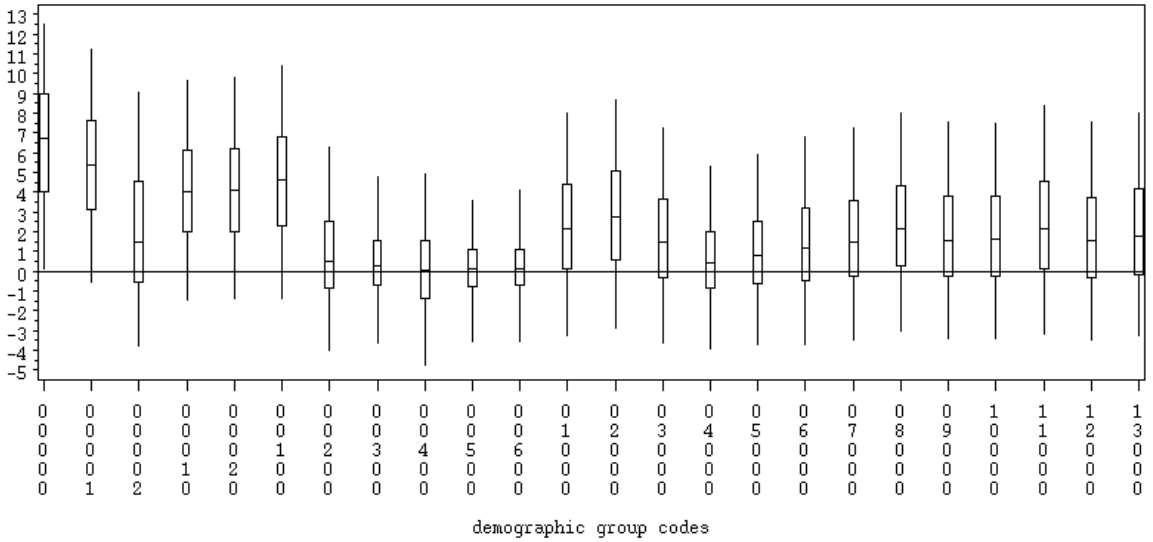


Figure 13: *Natural logarithm of Variance of tract-level administrative control estimates minus natural logarithm of Variance of county-level administrative control estimates for all tracts for selected demographic groups (Note: distribution of tract differences summarized by boxplots) - see note at appendix for definition of demographic groups*

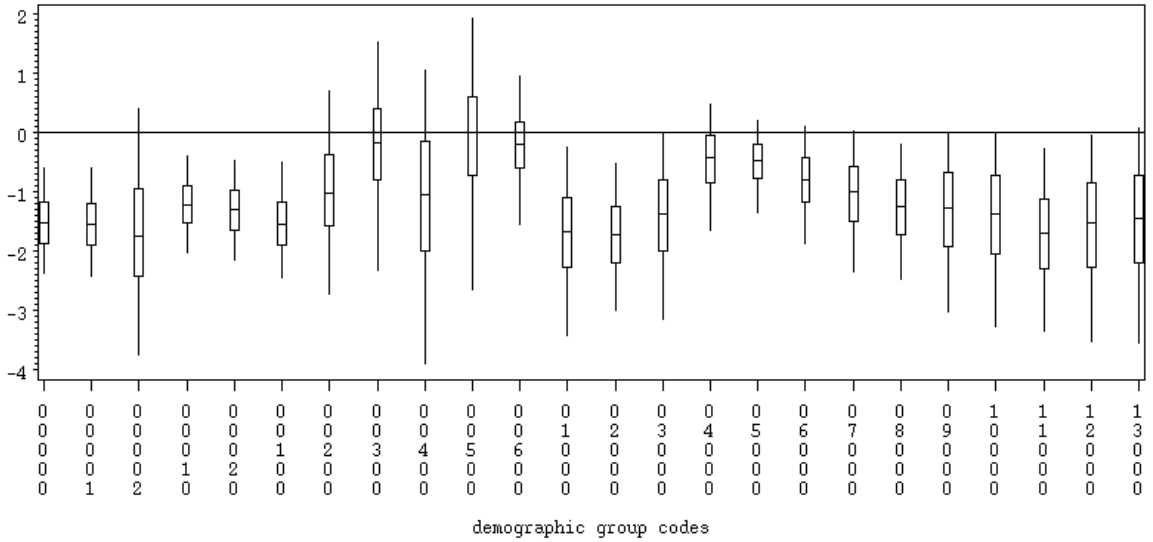


Figure 14: *Natural logarithm of MSE of tract-level administrative control estimates minus natural logarithm of MSE of county-level administrative control estimates for all tracts for selected demographic groups (Note: distribution of tract differences summarized by boxplots) - see note at appendix for definition of demographic groups*

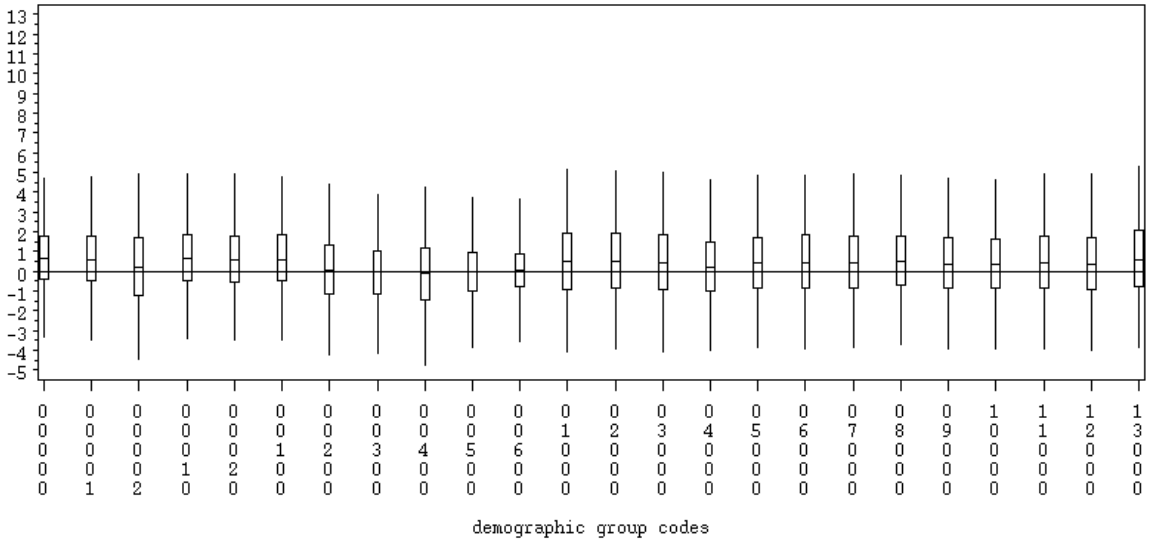


Figure 15: *Natural logarithm of Variance of county-level administrative control estimates minus natural logarithm of Variance of county-level census short-form control estimates for all tracts for selected demographic groups (Note: distribution of tract differences summarized by boxplots) - see the appendix for definition of demographic groups*

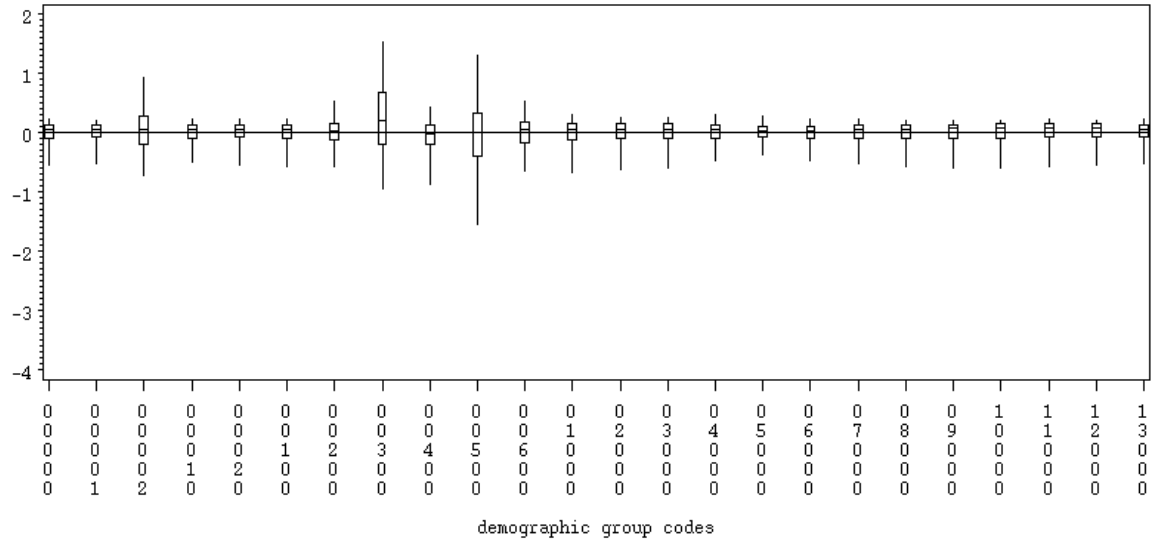
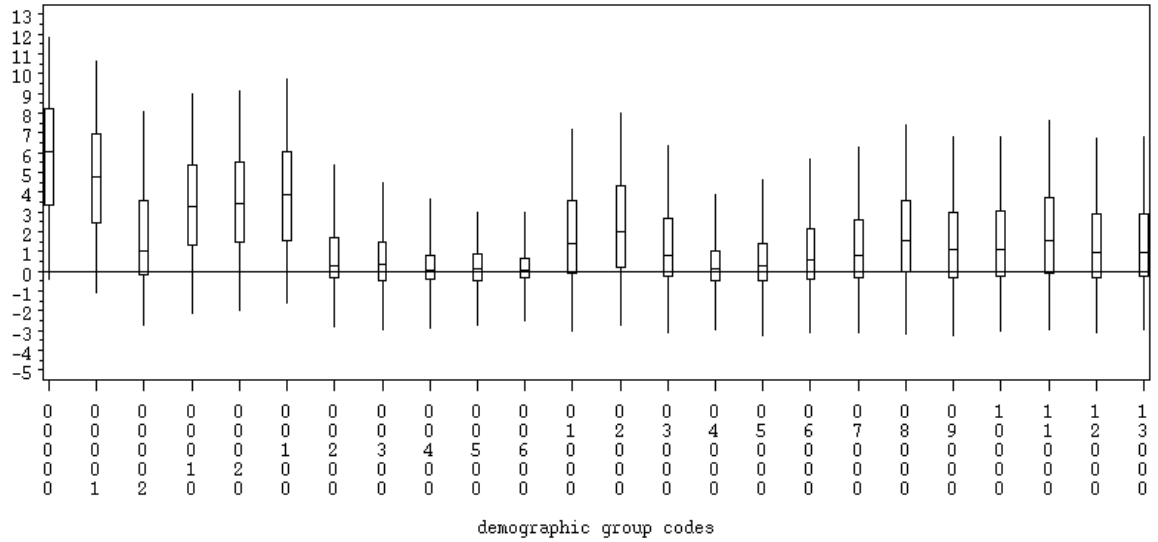
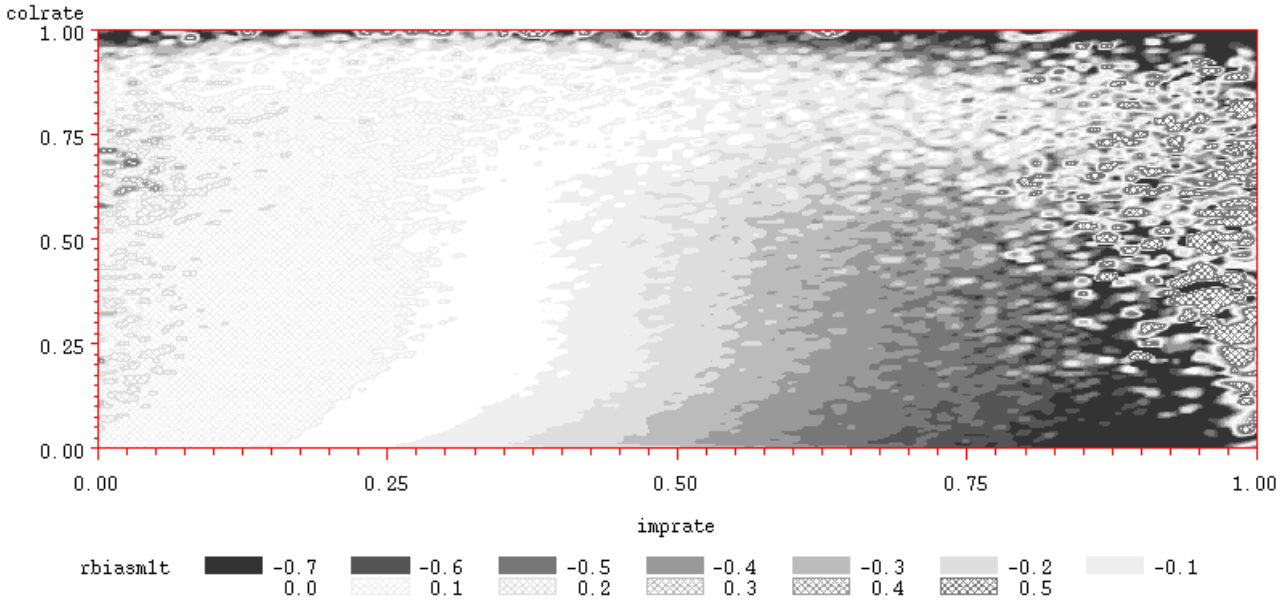


Figure 16: *Natural logarithm of MSE of county-level administrative control estimates minus natural logarithm of MSE of county-level census short-form control estimates for all tracts for selected demographic groups (Note: distribution of tract differences summarized by boxplots) - see the appendix for definition of demographic groups*



contribute to bias. In general, high collapse rates or high imputation rates lead to biased results. However, with reasonable amounts of collapsing and administrative tract imputations, the biases are relatively small.

Figure 17: *Evaluation of Tract-Level Bias of Tract-Level Administrative Record Controls for all demographic groups listed in the Appendix : Estimated bias versus proportion cross-classification of proportion collapsed and proportion imputed (Note: scatter plot grouped by local median values and smoothed)*



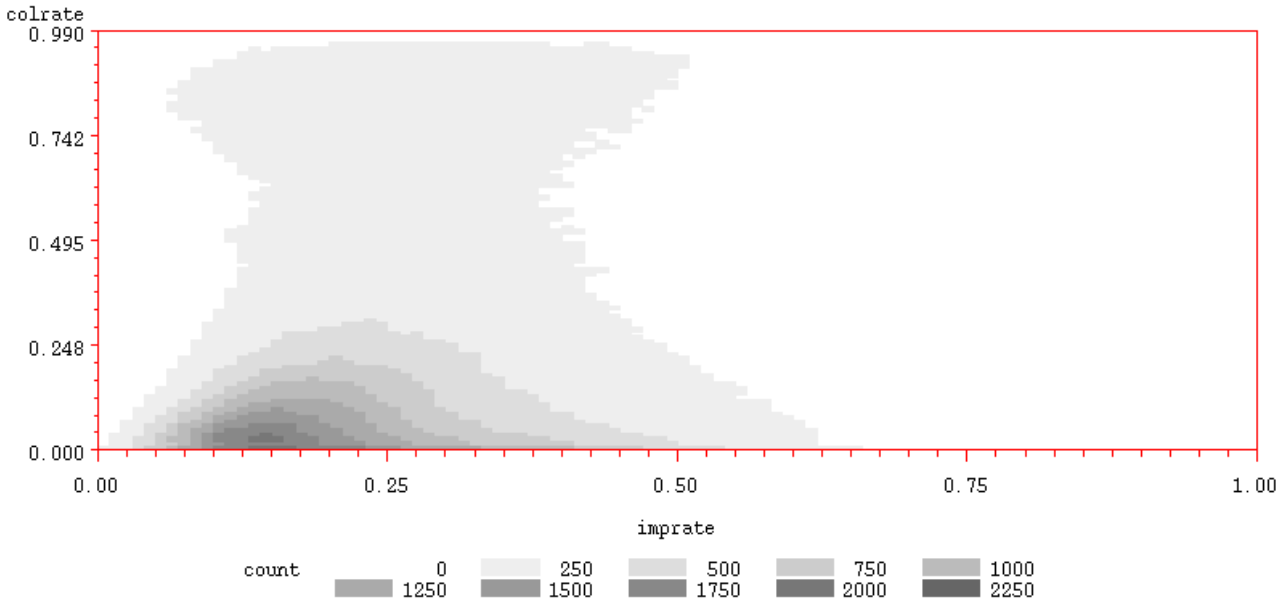
5 Discussion

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. There is some variance reduction due to control to the tract-level versus county level, but there appears to be a similar increase in bias negating the benefits of smaller variance when considering total error.

Should an interest in tract-level adjustment of coverage that preserves current residence arise, this project may help point to some ways to achieve this.

First, matching of administrative records in terms of reducing the number of unmatchables needs to be improved. Second, as compared to the Census short-form, the coverage of administrative records is still a problem. This current work could help serve as a check of what kind of matching errors are detrimental. Third, collapsing of cells needs to be revisited with an aim to achieving collapsed cells that are less bias prone. The collapsing scheme was created to be used at the county level for the ACS sample design. Attempting to apply these collapsing rules to smaller geographical levels make the underlying assumptions more important and may lead to a less accurate estimator. An alternative way may be to collapse across neighboring tracts

Figure 18: *Number of tract-level demographic components represented in Figure 17*



(or within an entire county) before collapsing across certain domains. Determining these collapsing rules would require more work.

Our method of imputing administrative record tract address for sampled cases that do not match to an administrative record is associated with biased controls. Seeking out alternative methods of imputation may be beneficial. One way would be to correct our biased estimates by forcing them to calibrate with the MAF/administrative record housing unit calibration used by Fay (2006). That is, use the calibration to remove some of the bias but still retain the links to administrative records at the person level.

6 References

Asiala, M. (2004), Specifications for Weighting the ACS 2003 HU Sample (ACS-W-6B), Second draft, Internal U.S. Census Bureau Memorandum to L. McGinn through R. Singh, Washington, D.C. 2004.

Farber, J. and Miller, E. (2003). Matching Census 2000 to Administrative Records. 2003 Proceedings of the American Statistical Association, Section on Government Statistics [CD-ROM]. American Statistical Association, Alexandria, VA. 1387Q1 1390.

Fay, Robert E. (2006) Using Administrative Records with Model-Assisted Estimation for the American Community Survey ASA Proceedings of the Joint Statistical Meetings, 2995-3001 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 Design and Methodology American Community Survey U.S. Government Printing Office, Washington, DC, 2006.
U.S. Census Bureau (2007). Population Estimates, http://www.census.gov/popest/estimates.php.

A Appendix: Codes for Demographic Groups

The five digit codes represent the following groups used in Figures 11-16.

First two digits:	00	All Ages
	01	0-4
	02	5-14
	03	15-17
	04	18-19
	05	20-24
	06	25-29
	07	30-34
	08	35-44
	09	45-49
	10	50-54
	11	55-64
	12	65-74
	13	75+
Third digit:	0	All Races
	1	White
	2	Black
	3	Indian
	4	Asian
	5	Pacific Islander
	6	Multiple Race
Fourth digit:	0	All Sexes
	1	Male
	2	Female
Fifth digit:	0	All Ethnicities
	1	non-Hispanic
	2	Hispanic

Table 1: