Journal of Privacy and Confidentiality (2014) **6**, Number 2, 21–58

Top-Coding and Public Use Microdata Samples from the U.S. Census Bureau

Nicole Crimi* and William F. Eddy†

1 Introduction

The US Census Bureau regularly releases Public Use Microdata Samples (PUMS), data files which contain de-identified subsets of the data provided by respondents to some of its various surveys and to the Decennial Census itself. This allows data users to perform "micro" -analyses rather than the "macro" -tabulations which are regularly performed by the Bureau. These data users range from non-government (say, university) researchers to government policymakers. These micro-analyses typically depend on the joint distribution of two or more variables over individuals or households. As a very simple example, think of the relationship of wages of individuals to their individual ages by a linear regression equation. We will use this very simple example throughout this paper to illustrate the effects we are interested in. In order to protect the privacy of the data supplied by respondents, as required by Title 13 U.S.C., the Bureau uses a variety of methods to modify the data so that it is very difficult for data users to identify individual respondents. Although some kind of privacy protection measures are necessary by law, most of them (top-coding, in particular) have a detrimental effect on the micro-analyses because application of these privacy protection measures changes the interdependence of two or more variables and, in many cases, renders the analyses moot.

This paper is a very brief review of Census Bureau privacy protection methods and a small exploration of the effect that top-coding, in particular, has on some specific micro-analyses. Throughout this document: a) we have focussed on the American Community Survey (ACS) PUMS because it is one of the richest national datasets; and b) we have used Alaska and California as example states because they have, respectively, very small and very large populations and because the age distributions and wage distributions are quite different between the two states. We have performed each of our analyses for every state and the results for the other states are available in the supplementary materials. In Section 2 we discuss privacy protection methods in a little more detail and, in particular, focus on a detailed understanding of top-coding, as currently used by the Census Bureau. In Section 3 we give a brief description of the data sets that we used in our attempts to correct top-coding in Section 4. We introduce the Health and Retirement Study, a non-Census Bureau survey, as a potential tool for correcting the

*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA mailto:ncrimi@andrew.cmu.edu.

†Departments of Statistics, Machine Learning, Biological Sciences, and Center for Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA mailto:bill@cmu.edu.

Both authors were partially supported by the **N**ational Science Foundation— **C**ensus Bureau **R**esearch **N**etwork (**NCRN**) under grant NSF 11-30706.

effect of top-coding. In Section 4 we describe the various correction approaches we tried, why they failed, and why there appear to be no other viable approaches to restoring the distributional properties (e.g., the correlation) of pairs of variables, at least one of which has been top-coded. Section 6 discusses the errors in the Census PUMS discovered by [1] and the fix provided by the Census Bureau, and some additional errors we discovered in the Minnesota Population Center IPUMS. Finally, in Section 7 we briefly discuss the implications of our study for statistical and economic analyses based on PUMS data which have been top-coded.

2 Privacy Protection Methods

The Census Bureau applies a variety of privacy protection methods to its PUMS data files. These methods are not usually described by the Census Bureau in any detail so that it will be more difficult for users to "break" the methods and identify the data for a particular individual. We do not know for a fact whether the Bureau retains the original unaltered PUMS files or not. (A reviewer of an earlier version of this manuscript stated unequivocally: "The Census Bureau does maintain unaltered PUMS files.") If the Bureau does not it might actually be impossible to recover the original data without going to the original whole survey or Census, an approach only available to Census Bureau employees and those individuals with Special Sworn Status.

According to [4], methods that the Census Bureau uses to protect microdata files include

- Removal of direct identifiers;
- Setting geographic population thresholds;
- Data swapping;
- Global recoding;
- Rounding;
- Top-Coding; and
- Age detail.

The Census Bureau always removes direct identifiers such as names, addresses, and telephone numbers in the microdata which it releases. It also reduces the spatial resolution of identifiers of geographic location. Typically, these locations are only identified down to areas with a population greater than 100,000 people. By grouping values of continuous variables, some masking of individuals is attained. For some surveys, this is done for age (which is approximately continuous) and is referred to as "age detail." This can be thought of as a coarser grouping than that obtained by rounding. Additionally, noise may be added to age for a small subset of the data. The two main Census Bureau methods for altering micro data to protect privacy are data swapping and top-coding;

we discuss these in the next two subsections. The released records are usually referred to as "de-identified."

2.1 Data Swapping

In this method, parts of individual records are interchanged to introduce uncertainty to the data user as to whether or not certain data values correspond to the specific record. We know the swaps have always been between pairs of records because [16] begins a Census study of n-cycles in larger sets of records. According to the Federal Committee on Statistical Methodology FCSM Report 22 [7], "Although swapping does not change the marginal distribution of any variable in a file, it does distort joint distributions involving both swapped and unswapped variables." This could severely hinder the usefulness of the PUMS, since the ability to estimate joint distributions from the data is the major incentive to use the PUMS. A reviewer commented that swapping has a very minimal effect on the PUMS and is mainly used to protect tabular data that is published for small geographic areas. For Census 2000, the swapping was done based on households that match on several key variables. Those households are switched with geographically nearby households that have similar characteristics [2]. There are, however, major issues with transparency when it comes to data swapping in the ACS and Census PUMS. The percentage of data which is swapped, as well as which variables are matched to determine how the swapping occurs, are unknown outside the Census Bureau. Furthermore, which variables are swapped and which are not is also unknown. We have begun a separate activity to assess the effect of data-swapping on the joint distribution of a pair of variables, one of which is swapped and one of which is not swapped.

2.2 Top-coding

According to FCSM Report 22 (both the original [6] and the revised version [7]), top-coding is defined as "an upper limit on all published values of that variable. Any value greater than this upper limit is not published on the microdata file. In its place is some type of flag that tells the user what the top-code is and that this value exceeds it." [7] Top-coding of age, for example, would choose the top-code at a specific age (say, 90 years) and all records with ages above that age (90) would simply be recorded in the PUMS as that age (90) together with some flag indicating the fact that the particular value was top-coded. Bottom-coding is defined in a completely analogous way. We will not discuss bottom-coding further.

The type of top-coding that has been used in the ACS PUMS and Census PUMS is slightly different from the definition in [7]. The data documentation for the 2000 Census defines top-coding as "a method of disclosure limitation in which all cases in or above a certain percentage of the distribution are placed into a single category."[2]. The 2010 ACS file listing top-coded variables says that "Age, travel time to work, and all base dollar amounts are top-coded using the state mean of all cases greater than or equal to the top-code state minimum value" [14]. As a specific example, age (within

each state) has been truncated at some large value (say, 90 years) which varies by state. All of the truncated values are placed at a particular value (we will refer to it as the replacement value), chosen so that the (weighted) mean of the (age) distribution is approximately correct. This is not "top-coding" as defined in [7] so we will refer to it as "mean-corrected top-coding."

More specifically, the Census Bureau chooses truncation points for the mean-corrected top-codes by looking at either the top 3% for a "general universe" or the top .5% of each "specific universe" for a chosen variable. The choice is made for whichever one allows for the release of more data at a state level. In the ACS PUMS, the "general universe" is all non-zero values for the top-coded variable, while the "specific universe" includes all values in the calculations. In the 2010 ACS PUMS, for example, the age truncation point is different for each state, and the interval above each replacement value contains at least .5% of the data. This means that, for each state, the truncation at the .5% value (using all the data given, including zeros) allowed for more data to be released than using the value calculated by taking the top 3% of all the non-zero values.

An earlier version of this paper incorrectly interpreted the exact definition of specific and general universe. We were informed by a different Census Bureau employee during the review process that our initial understanding was incorrect. The main reason for our misunderstanding stemmed from the fact that the details for how the truncation points are chosen (the .5% or 3% rule) are not readily available in any public documentation. It was only through direct correspondence with a Census Bureau employee that we were informed of the methodology (in generalities), and this is what we based our conclusions on.

After the truncation point for a state is chosen, the replacement value is chosen by taking the mean of the data that needs to be top-coded. All of the values above the truncation point are then coded as the replacement value (the mean of all the top-coded data). Those values for the 2010 ACS PUMS are given in Tables 1 and 2 (at the end of the paper). This decision is based on the data collected in the ACS sample and included in the PUMS subset; it is not based on the entire population, or even on the entire sample but only that portion of the sample to be included in the PUMS.

See Figures 3 and 5 for two specific examples: Alaska and California. The replacement value is chosen to (roughly) preserve the mean of the specific (state) age distribution, but this results in a very strange tail behavior of the age distribution for the general (national) age distribution because the national data file is the union of the state data files; see Figure 2. It also results in strange tail behavior for the multi-year PUMS aggregations for individual states because the replacement value can change from year to year. See Figures 4 and 6 for two examples of this behavior.

The national file for both the ACS and the Census is simply an aggregate of all the individual state files. Although it would be possible to choose uniform top-codes for the whole nation while keeping the state file top-codes as-is, this introduces a privacy protection problem. If the national top-code was higher than all of the individual state codes, it would be possible to bypass some of the protection from each individual state top-code by comparing the data from the individual state file to that state's data within

the national file. In order to avoid this, the national PUMS are simply a union of all the state files with the top-codes that were already chosen for each state remaining the same.

2.3 The Simple Effect of Mean-corrected Top-Coding

Mean-corrected top-coding as a method of privacy protection creates, probably insoluble, problems for the analyst. The joint distribution of any variable that has been top-coded with any other variable, top-coded or not, is incorrect. It is apparently not possible to overcome this defect. We will show two failed attempts to overcome this and discuss a third possibility in Section 4. Consequently, (since age and wages are top-coded) it is not possible to study anything about the relationship between them or even separately to study the "oldest old" or the "richest rich" (measured by wages). This problem is common to all methods of top-coding. And because the top-code cut-off varies by state and the national PUMS file is the union of the state level files, any analysis of the national data is equally incorrect. More clearly, the tail of the top-coded age distribution in the national PUMS file is a mixture of the tails of all the top-coded state distributions and hence is very unusual. These problems are compounded if one were to perform a joint analysis of income and age. Both of these variables use mean-corrected top-coding (in the ACS PUMS) and, for example, a simple linear regression of wages on age is simply wrong and cannot be corrected. We will discuss this further in Section 5.

It is easy to see that the sample correlation between two variables is shrunk toward zero if one or both are top-coded. Suppose the n observations on the two variables are

$$(X_i, Y_i), i = 1, \ldots, n$$

and without loss of generality assume they each have sample mean zero and sample variance 1. Top-coding one or both of them is easily seen to shrink the correlation between them. The correlation in this case is defined as

$$\sum_{i=1}^{n} X_i Y_i.$$

For convenience, assume that the variable X is top-coded and that the pairs have been sorted so that $X_i \leq X_{i+1}$. If only the largest value of X, X_n, is reduced (to X^*) then the correlation will be closer to zero since

$$|\sum_{i=1}^{n-1} X_i Y_i + X^* Y_n| \leq |\sum_{i=1}^{n} X_i Y_i|.$$

3 Data Sources

This section is the beginning of an exploration of the possibility of correcting the PUMS data for the effects of top-coding. In principle, it might be possible to adjust the top-coded values in some way so that the original joint distribution could be recovered. The

approach we consider is to use other data sets to correct the tails of the distributions. Without access to the original data we will not be able to produce the exactly correct original data but we might be able to produce some adequate approximation to the tail by substituting an alternative set of data. Here we simply describe the data sets we considered; in the next section we detail our actual efforts.

3.1 The Census

As required by the US Constitution in Article 1, Clause 2, Section 3, a census of the population is conducted every ten years (in those years that end with a zero). In recent decades (between 1940 and 2000) there have been two versions of the Census, the "Short Form" and the "Long Form." The short form, which was delivered to approximately 83% of the households, was limited to a small number of demographic questions (8 in the year 2000), such as age, gender, and race. From 1970 through the year 2000, the other 17% of the population received the long form [15], which contains the questions from the short form, together with many additional questions (45 more in 2000) that go much further than the basic demographic questions. These include questions on occupation, presence of mental conditions, income, and several other topics. The long form of the Census was discontinued after the year 2000, and for the 2010 Census essentially everyone received the same 10 question short form. Before 2010, the Census Bureau released large numbers of Summary Tables based on the Census which generally were cross-tabulations of two or more variables that are of interest to the Bureau's data consumers. In aggregate these tables are generally referred to as the STF-3 (Summary Tables File 3).

3.2 The American Community Survey

The American Community Survey is a continuous survey (that is, data is collected every month) with a complex sampling plan that officially began in 2005 after a number of years of development and testing. Approximately 1 in 38 US households are sampled each year. A main goal behind the introduction of the ACS was to use it as a replacement for the "Long Form" of the Census [12]. This was done in order to have a constant flow of data each month or year that would provide more temporal accuracy than having a large amount of data from the one Census year each decade. The hope was that the cost of a significant loss in national geographic resolution was worth the gain in temporal resolution; also the increased temporal resolution would allow for time-series analyses. The ACS is used to allocate more than $400 billion in government funds every year and is also used by many non-government institutions [13].

Along with single-year tabulations (summary files) and associated PUMS files, each year the Census Bureau releases 3-year and 5-year summary files and their associated PUMS files. The multi-year summary files are simply the single year data aggregated over those three or five previous years. The level of geographic detail for the release of the summary files depends on which aggregate tabulations are being used. For the 1-year aggregations, the summaries are released for areas with population greater than

65,000. For the 3-year aggregations, those areas with population greater than 20,000 have summaries released. Summaries are released for areas down to the census block-group level in the 5-year aggregates [9]. Some of these summary tables are suppressed to protect respondent privacy. The 3 and 5-year PUMS contain the same sample units as the aggregated 1-year PUMS files for their respective years. For these aggregate PUMS files, several variables are changed from the 1-year files, including income adjustments, housing and person weights, and replicate weights [10].

3.3 The Health and Retirement Study

The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging and is conducted by the Institute for Survey Research at the University of Michigan. It surveys more than 20,000 people over the age of 50 every two years (with smaller surveys in the off-years). The entire household is studied, meaning that a spouse of someone who is selected for the HRS (older than age 50) is automatically included in the study even if they aren't over the age of 50. Once a person is in the study, they are there for the rest of their life, allowing for research to be done over several years for the same person. The study collects information on an extremely wide variety of topics, from questions about income and wealth to questions about health; see [11]. PUMS are released for the survey, but unlike the ACS and Census, macro-tabulations are unavailable [5]. We studied the HRS thinking we might be able to impute values in the ACS that had been altered by Census Bureau privacy protection measures based on information about the distribution of some variables from the HRS.

3.4 Census Public Use Microdata Samples (PUMS)

Census PUMS are a data file of "de-identified" responses for a subsample of the respondents to a survey. For a respondent in the PUMS, the answers to individual questions from the survey will be available in that respondent's record, with specific identifiers such as names and addresses removed so as to protect the privacy of the individual. The PUMS are important because they allow users to perform their own micro-analyses, making them a valuable resource for researchers who wish to study the relationships between and among variables at a finer level of detail than that given by the Census Bureau summary tables. PUMS are not available for every survey. However, the Census Bureau has made them available for the Census ([3], [2]) and the ACS, and the University of Michigan has released them for the HRS.

The Census PUMS has typically been released about two years after the official Census date and from 1980 to 2000 has contained both a 1% sample and a 5% sample of the long form records (1960 and 1970 only contained a 1% sample)[2]. The 1% and 5% samples are independently drawn, and a household may be included in only one of the two samples, making the total amount of PUMS data available equivalent to 6% of the total population.

There were two differences between the 1% and 5% file in 2000. The first is in the

level of geographic detail available. The Census Bureau has introduced even higher levels of geographic areas for this purpose. A PUMA is a Public Use Microdata Area defined by the Census Bureau; it is an aggregation of lower level Census geographic areas within a state and is constructed in consultation with each state government and has a population of more than 100,000. A super-PUMA is the aggregation of two or more PUMAs within a state that totals more than 400,000 population. The smallest geographic area that is defined for the 1% Census PUMS is the "super-PUMA", which has a minimum population of 400,000. The 5% sample contains a variable for the super-PUMA, but also goes a step further down to the PUMA (minimum population of 100,000). The geographic differences between the 1% and 5% file have changed each decade for the Census PUMS from 1980–2000. For example, the 1990 Census had PUMAs in both files, but the 5% PUMAs were different from the 1% PUMAs that year.

The second difference between the two files deals with the level of detail revealed for categorical variables. In the 1% file, the only restriction for variables is an 8,000 national minimum population for race and Hispanic origin. The 5% file is much more restricted, with a 10,000 national minimum population for all categorical variables and categories within these variables. The variable distinctions were not used in 1980 or 1990. [2]. They were not different in the 1% and 5% files.

Prior to 2010, the Bureau released PUMS files corresponding to the long form data. The 2010 Census PUMS has not been released as of this writing and we have been told there are no current plans to release it, primarily because of budgetary reasons (we note that it only has responses to the short form questions). The ACS PUMS is typically released early in the year following the nominal year of the survey, and it contains data on nearly 1% of the total US population. The HRS releases several different PUMS files, and in recent years these have generally been released within 3 years of the completion of the field data collection.

3.5 Integrated Public Use Microdata Series (IPUMS)

While PUMS are released by the individual institutions that run the studies (the US Census Bureau for the ACS and the Census, and the ISR at the University of Michigan for the HRS), the IPUMS is a project at the University of Minnesota Population Center intended, in part, to make research using the PUMS from the Decennial Census, ACS, and various other surveys easier to work with by collecting and freely distributing the data [8]. One of the major ways this is accomplished is through a uniform system of coding. Over the years, questions in both the Census and ACS have changed as well as the possible answers and the coding used in the PUMS data has followed. The uniform system in IPUMS makes it easier to study data across different years, samples, and surveys. Another unique feature of the IPUMS is the use of family interrelationship variables, in which the record for an individual also contains information about the mother, father, or other family member. Also, the online system allows users to pick specific variables they would like to study, allowing for smaller data sets instead of requiring the user to first download an entire data set and then pick through the variables

individually. The data behind the IPUMS for Census and ACS is the same as the PUMS data released through the Census Bureau (except for any errors introduced during the transfer into the IPUMS system. See Section 6 below.)

4 Attempts to Reconstruct the ACS Data

We, naively, thought it might be possible to "correct" the bias introduced into the ACS PUMS data by top-coding; we thought that by using other information we might be able to estimate the "missing" tail of the distribution. In this section we describe two attempts we made and one that we couldn't make.

4.1 Using the Census Summary Tables as the True Distribution

In an attempt to repair the damage done to the marginal distribution of age by top-coding the ACS PUMS data for 2010, we looked to the published data counts for the 2010 Census. These published counts were used to estimate the tails of the histograms of the top-coded variable. For all ages below the top-coded value, we kept the original counts (from the 2010 ACS PUMS). Above the top-coded value, the distribution of age was taken from the published Census counts. We simply used a ratio estimator to smooth out the top-coded section of the tail of the distribution. See Figures 3 and 5. Although this seemed to work well for age, it was not possible for us to attempt this for wage or any other top-coded variable because the Census Bureau does not release Census Summary Tables using these variables. Our intention was to attempt to repair a two-way distribution of age and wage. However, because the Census Bureau doesn't release any Census tables on wage, they also don't release two-way tables of age and wage. Correcting the bivariate distribution was not possible.

In the case of the 2010 Census, although the data might be a good match for the 2010 ACS, this would become less accurate as the years get further in the future from 2010 since there is no way to account for changes within specific ACS PUMS records collected since the 2010 Census. The issue with using the Census for the base data, although it doesn't apply to the 2010 ACS, is one of timeliness and accuracy. For both data sets, the largest problem with attempting to reconstruct the PUMS is that they are simply superficial repairs. When we reconstruct the distribution, the actual PUMS records aren't changed because we do not know which "corrected" value goes with which record. Therefore there is no useful further analysis that can be done. Repairing the distribution is only valuable to see how the tail should look if it weren't top-coded. If we could impute the ages and other top-codes within the individual PUMS records (using, e.g., a missing-at-random model), it might be possible to perform analyses. This, however, ignores the major problem that we are trying to correct by replacing the top-codes: we can't match the characteristics associated with a specific age (or other top-coded variable); the missing-at-random model is almost certainly not correct.

4.2 Using the HRS as the True Distribution

Using the Census tables we were unable to repair more than the simple age distribution, thus we look to a different source for wages and other top-coded variables. The HRS PUMS are not top-coded on wage or age, making them a possibly good choice to study those variables, especially because we were attempting to repair a joint distribution. This was only partially successful. The subjects of the HRS are older than 50 years, and although people younger than 50 are included in the data (spouses, etc.), the focus is on those older than 50. It would therefore be inaccurate for us to attempt to fix all the top-coded wages, since not all the people in the ACS with top-coded wages are above the age of 50. Therefore, we were only able to repair the wage distribution for those greater than age 50. For the bivariate distribution of age and wage, our main goal was to focus on the intersection of the two variables that had been top-coded. Since the two top-coded variables we were looking at were age and wage, the fact that the HRS focused on those over age 50 was irrelevant because all of the age top-codes are higher than age 50.

4.3 Using the Census PUMS as the True Distribution

The ACS was specifically designed to be a replacement for the Decennial Census "Long Form." Therefore, the joint distribution between variables which are common to the ACS and the Census should be the same, allowing us to reconstruct the tail of the ACS joint distribution from the tail of the Census joint distribution. We did not attempt this for the simple reason that there is not a common collection time (e.g., year) for the two data sets. Specifically, the 2000 Census predates the release of the first (2005) ACS PUMS by five years. Alternatively, the 2010 Census does not include the Long Form variables and there are currently no plans to release a Census PUMS for the 2010 data. One potentially interesting aspect is that there are many additional variables that might be common to the ACS and the Census; this would allow us to condition on one or more of these other variables and produce different estimates for the top-coded variables depending on that conditional distribution.

5 Wages Regressed on Age

Our intended goal in this study was to estimate the relationship between Wages and Age in the 2010 ACS using the PUMS data; we recognized at the outset there would be problems and we simply hoped to discover a method to correct them.

We began by looking at marginal histograms of the two variables. Those for (raw) age are given in Figures 3 and 5 for our two specific examples, Alaska and California. Those for (raw) wages are given in Figures 9 and 10 for our two specific examples, Alaska and California. We note that because of the large number of zero wage values, we removed the zeroes to make the remainder of the distributions visible.

The scatterplots of Wage versus Age for the same two states are given in Figure

11 and 12. In both plots the effect of top-coding is easily seen. Those values which were top-coded on age are seen near the right margin and those values which top-coded on wage are seen near the top margin. In the California scatterplot (Figure 12) the single point which is rightmost and uppermost represents all those individuals who were top-coded on both variables. In both scatterplots it should be noted that the sort of ovoid point cloud one often sees in such a plot is not visible because the presence of a point in the scatterplot indicates the presence of a data point at a particular age/wage combination but not how many such data points were present. There are many fewer age/wage pairs than there are individuals. Consequently, each plotted point represents many individuals. To visualize this fact would require a three-dimensional plot, the use of color, or some similar device.

For each state we performed a number of regressions to relate wage to age. Specifically, we regressed Wage on Age for all individuals 16 years of age and older, yielding an intercept coefficient and a slope coefficient. We then repeated the regression for all individuals 17 years of age and older; and then repeated for all 18 years and older ... We stopped at the maximum possible age for each state; remember, the top code truncation point varies by state. The idea was to learn a little about how truncation affects the estimated regression coefficients.

Figures 15 (intercept) and 16 (slope) show a bar plot of the estimated coefficients for California. The first thing we note is that they vary quite smoothly as we truncate more and more data (as one might expect). We note that the slope coefficient is always negative, becoming most negative when we have eliminated those with an age below (roughly) 50. The intercept is always positive and when appropriately scaled is almost exactly the negative of the slope, particularly for the estimates where we have eliminated those with an age below (roughly) 50. This suggests that the relationship of wages to age is not linear; the conditional mean looks similar to the intercept.

Figures 13 (intercept) and 14 (slope) show the same bar plots for the estimated coefficients from Alaska. We note they do not vary as smoothly as the California plots, presumably because of a much smaller population. Second, the bar plot of slopes is not unimodal and not consistently negative. We conjecture that this is also due to the small population and an unusual tail distribution for age (not like California). We also note that the bar plot of estimated intercepts is also approximately the negative of a scaled version of the estimated slopes.

A rough summary for the other states is that the behavior of the equivalent plots for many other states was generally very similar to California. We did note a number of exceptions and those were mainly in states we think of as having small populations. A number of these exceptions appeared similar to the plots for Alaska. Very few plots had the upper tail of the slope distribution go positive as though the upper tail changed sign as it crossed zero. And one, the District of Columbia, had a total of three modes. See the Supplementary Materials.

6 Errors in the Data

6.1 Census Bureau Errors

We successfully replicated the ACS and Census findings of Alexander et al. [1]. We visually compared Figures 17–18 with figures in their paper and they appear to be identical. The discoveries that resulted from their research were apparent errors in the PUMS for the ACS, for the 2000 Census, and for the Current Population Survey (CPS). In the 2000 Census 5% PUMS and the 2005 and 2006 ACS PUMS, the discrepancies were found by summing up the person weights at each age (to find the resulting population estimate), separately for each gender, and then dividing by the released estimates from all of the ACS or Census data for that year. For the ACS, this was done by age groups as opposed to individual ages because the published estimates do not provide the same level of detail as the Census estimates and are instead broken down into age groups. Above age 65, the estimates from the PUMS differ by up to 15% from the published estimates. The main cause for these problems was determined to be the Census privacy protection measures. Because the truncation ages were changed following 2005 in the ACS, we can assume that this was due to some error in the top-coding of age as applied to the PUMS. This discovery is an excellent example of the repercussions when privacy protection measures are not carried out properly; this effect is over and above the obvious effects on joint distributions which we study here. Not only were problems discovered relating to age and gender, there were also issues relating to labor estimates for the 2000 5% Census PUMS and the 2006 ACS marriage estimates.

To ensure that the problem discovered in [1] did not occur in the 2010 data, we performed the same analysis on the 2010 ACS PUMS as we had done for the 2006 ACS PUMS. From this analysis it appears that the problem did not occur in the 2010 ACS PUMS. The PUMS estimates never differ by more than 1% from the published counts.

6.2 IPUMS Errors

Replicate weights are available for the ACS data for each year starting in 2005. There are 80 separate replicate weights at the household and person levels that allow users to generate empirically derived standard error estimates. These standard errors can then be used in hypothesis testing and in the construction of confidence intervals around the sample estimate of interest. The method for generating these standard errors for any analysis is to run the analysis using the full sample weights and then run 80 (!) additional analyses using each set of replicate weights in turn. Then simply estimate the standard error for the variable of interest

$$SE(X) = \{\frac{4}{80}\sum_{i=1}^{80}(X_r - X)^2\}^{\frac{1}{2}}$$

where X is the result from the analysis using the full-sample weight and X_r is the result from the analysis using the r^{th} set of replicate weights.

When looking at the replicate weights in IPUMS, we discovered a very small number

of discrepancies between the IPUMS and ACS PUMS for 2010. All of the weights and replicate weights in IPUMS should be the same as those in the ACS PUMS because the IPUMS is based on the ACS PUMS. After comparing both, we found that there were 13 discrepancies in the replicate weights. Most of the errors seem to be those of insertion/deletion of characters; for example, where a replicate weight should have been -1856, it became 856 in the IPUMS. There was also one instance of switching, in which two neighboring records had one of their replicate weights switched. Although this number of errors is small relative to the number of records contained in the files, it brings to question the accuracy of the rest of the variables in the IPUMS. See Table 3.

7 Summary and Conclusion

Privacy protection measures are necessary if the PUMS are to be available at all. However, whatever methods are used need to ensure that the probability distribution of the data, particularly the joint probability distributions (which are almost always the target of study), are not affected. As the ACS is used in a large amount of non-government research as well as to allocate over $400 billion in government funds, it is extremely important that its PUMS data accurately reflects the actual data that have not been subject to privacy protection measures. Our findings indicate that this method of mean-corrected top-coding, only one of many privacy protection measures, irreparably damages the data.

Currently, researchers who are interested in studying the "oldest-old" for the nation as a whole face the dilemma that mean-corrected top-coding obscures the age of exactly their population of interest. Furthermore, because the top-codes are chosen for each state separately, there are automatically state-level effects which cannot be corrected. The same issue is also found in all other top-coded variables. The problem is compounded when looking at a joint distribution of two top-coded variables (such as age and wage). There is no way to correct even a simple linear regression, making it impossible to accurately perform certain types of analyses.

The critical issue in privacy protection for the PUMS concerns the geographical dependence of the data. If geography is not revealed in detail then in almost every case discovering the individual that corresponds to a particular record is impossible. For example, the oldest individual in any sample is very unlikely to be the unique oldest person in the country; if the geography is sufficiently blurred (e.g., only revealed at the level of a PUMA) it will be impossible to recover the identity of that sampled individual. Similarly, the wealthiest individual in any sample is very unlikely to be Warren Buffet; if the geography is sufficiently blurred it will be impossible to recover the identity of that sampled individual.

We believe that using geographic population thresholds together with data swapping provides sufficient privacy protection while preserving the probability distributions of the data.

Top-coding as a method of privacy protection has been used by the Census Bureau

for many decades. Age does not improve the quality of the method. The Census Bureau, perhaps rightly, is very conservative with respect to disclosure avoidance. However, as a consequence, almost all analyses of PUMS data are incorrect and, more generally, may even be misleading. Perhaps the Census Bureau could "solve" this problem by abandoning PUMS completely and rather, substitute synthetic data which, at least in principle, would not be directly affected by privacy concerns.

Acknowledgments

We'd like to thank several people who helped us along the way. Jaime (JT) Trujillo discovered the IPUMS errors. Beka Steorts and Steve Fienberg provided specific helpful comments on an early draft manuscript. The entire CMU NCRN team provided a valuable sounding board as we developed these ideas. Four anonymous reviewers of the original version made a large number of helpful comments which helped us remove many errors from the original manuscript and improve the presentation.

State Truncation Value by Year

State	2005	2006	2007	2008	2009	2010	2011
AL	89	**90**	90	90	90	90	90
AK	85	85	**86**	**84**	**86**	**85**	**86**
AZ	88	**89**	**90**	90	90	90	90
AR	89	**91**	**90**	90	**91**	**90**	90
CA	88	**89**	89	**90**	90	90	90
CO	88	88	**89**	**90**	**89**	**90**	**89**
CT	89	**91**	91	**92**	**91**	**92**	92
DE	87	**89**	**90**	**89**	**91**	91	**90**
DC	89	**90**	90	**91**	91	91	91
FL	90	**91**	91	91	91	91	91
GA	87	**88**	**89**	**88**	**89**	89	**88**
HI	91	91	**90**	**91**	**90**	**91**	91
ID	89	**90**	90	90	**91**	**90**	**89**
IL	89	**90**	90	90	90	**91**	91
IN	89	**90**	90	90	90	90	90
IA	90	**92**	92	92	92	92	92
KS	89	**91**	91	91	91	91	91
KY	88	**90**	90	90	90	90	90
LA	87	**89**	89	**90**	90	90	90
ME	89	**91**	**90**	**91**	91	**90**	**92**
MD	88	**90**	**89**	89	**90**	**91**	**90**
MA	90	**91**	91	91	91	91	91
MI	89	**90**	90	90	90	**91**	91
MN	89	**91**	91	91	91	91	91
MS	88	**90**	90	90	90	**89**	89
MO	89	**90**	**91**	91	**90**	**91**	91
MT	89	**90**	90	90	**91**	91	**92**
NE	90	**91**	91	**92**	**91**	91	91
NV	87	**88**	88	88	88	88	88
NH	88	**90**	90	90	90	90	**91**
NJ	89	**90**	90	**91**	91	91	91
NM	89	**90**	90	90	90	90	90
NY	89	**91**	91	91	91	91	91
NC	88	**89**	89	89	89	89	**90**
ND	90	**92**	**93**	93	93	**92**	**93**
OH	89	**90**	90	**91**	**90**	90	**91**
OK	88	**91**	**90**	90	90	90	90
OR	90	**91**	**90**	**91**	91	91	91
PA	89	**91**	91	91	91	91	91
RI	89	**91**	91	91	**92**	92	92
SC	88	**89**	89	**90**	90	90	90
SD	90	**91**	**92**	92	92	92	**93**
TN	88	**89**	89	**90**	90	90	90
TX	87	**89**	89	89	89	89	89
UT	87	**88**	**89**	**88**	88	88	**89**
VT	88	**90**	90	90	**91**	91	**90**
VA	88	**89**	89	89	89	89	**90**
WA	89	**90**	90	90	90	90	**91**
WV	90	**91**	**90**	90	90	90	90
WI	89	**91**	91	91	91	91	91
WY	88	**89**	89	**90**	90	**91**	89

Table 1: Top-coded truncation ages for the ACS broken down by State and Year (2005-2011). Whenever there is a year-to-year change in the truncation age for a specific state we have indicated that with a bold face value. Nearly all states had a change in the truncation age in 2006. This is discussed in Section 6.1.

State Replacement Value by Year							
State	2005	2006	2007	2008	2009	2010	2011
AL	92	**93**	93	93	93	93	93
AK	88	**89**	89	**87**	**89**	**88**	**89**
AZ	91	**92**	**93**	**92**	**93**	93	93
AR	92	**94**	**93**	93	**94**	**93**	93
CA	91	**92**	92	**93**	93	93	93
CO	91	91	**92**	**93**	**92**	**93**	**92**
CT	92	**94**	94	**95**	**93**	**95**	**94**
DE	90	**92**	**93**	**91**	**94**	**93**	93
DC	92	**93**	93	93	**95**	**94**	94
FL	92	**94**	94	94	94	94	**93**
GA	90	**91**	**92**	**91**	**92**	92	**91**
HI	95	**94**	**93**	**94**	**93**	**94**	94
ID	91	**93**	93	93	93	**92**	92
IL	92	**93**	93	93	93	**94**	94
IN	92	**93**	93	93	93	93	93
IA	92	**95**	95	95	95	95	**94**
KS	92	**94**	94	94	94	94	94
KY	91	**93**	93	93	93	93	93
LA	90	**92**	92	**93**	93	93	**92**
ME	92	**94**	**93**	**94**	94	**92**	**95**
MD	91	**93**	**92**	92	**93**	**94**	**93**
MA	93	**94**	94	94	94	94	94
MI	92	**93**	93	93	93	**94**	94
MN	92	**94**	94	94	94	94	94
MS	91	**93**	93	93	93	**92**	92
MO	92	**93**	**94**	94	**93**	**94**	94
MT	91	**93**	93	**92**	**94**	94	**95**
NE	93	**94**	94	**95**	**94**	94	94
NV	90	**91**	91	**90**	**91**	91	91
NH	91	**94**	**93**	93	**92**	92	**93**
NJ	92	**93**	93	**94**	**93**	**94**	**93**
NM	91	**93**	93	93	93	**92**	93
NY	92	**94**	94	94	94	94	94
NC	91	**92**	92	92	92	92	**93**
ND	92	**94**	**95**	95	95	95	95
OH	92	**93**	93	**94**	**93**	93	93
OK	91	**94**	**93**	93	93	93	93
OR	92	**94**	**93**	**94**	**93**	**94**	94
PA	92	**94**	94	94	94	**93**	**94**
RI	91	**94**	94	94	94	**95**	**94**
SC	91	**92**	92	**93**	93	93	**92**
SD	92	**94**	94	94	**95**	**94**	**95**
TN	91	**92**	92	**93**	93	93	93
TX	90	**92**	92	92	92	92	92
UT	90	**91**	**92**	**91**	91	91	**92**
VT	92	**93**	**92**	**93**	**94**	**93**	93
VA	91	**92**	92	92	92	92	**93**
WA	92	**93**	93	93	93	93	**94**
WV	93	**94**	**93**	93	93	93	93
WI	91	**94**	94	94	94	94	94
WY	91	91	**92**	**93**	93	**94**	**92**

Table 2: Top-coded replacement ages for the ACS broken down by State and Year (2005-2011). Mean-corrected top coding of age uses a replacement age which is different than the truncation age. The replacement age is chosen to make the (weighted) average of the age distribution after top coding the same as it was before top-coding. We have indicated those values that change from year-to-year within each state by bold face value. Nearly all states had a change in the truncation age in 2006. This is discussed in Section 6.1.

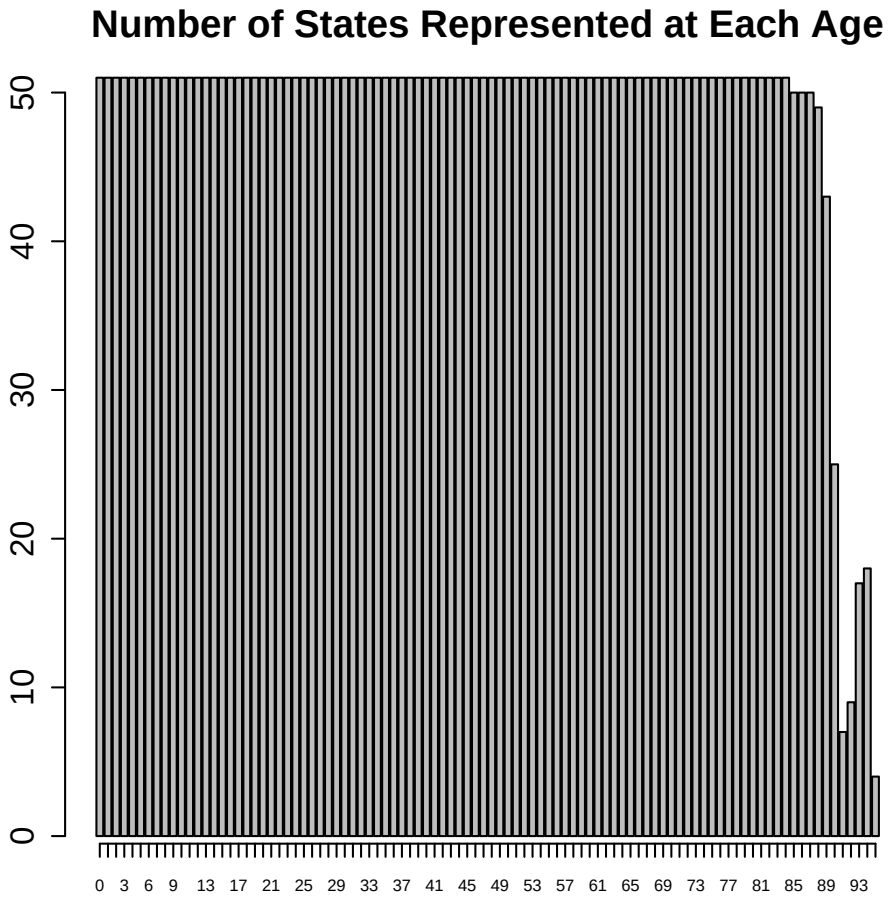

Figure 1: Barplot of the number of states represented by actual data at each age in the 2010 ACS PUMS files. Note that for each age before age 85, 50 states along with the District of Columbia are represented. After age 84, the number of states decreases (but not monotonically) and finally ends with only four states represented at age 95. This lack of representation by actual data at the oldest ages is due to top-coding.

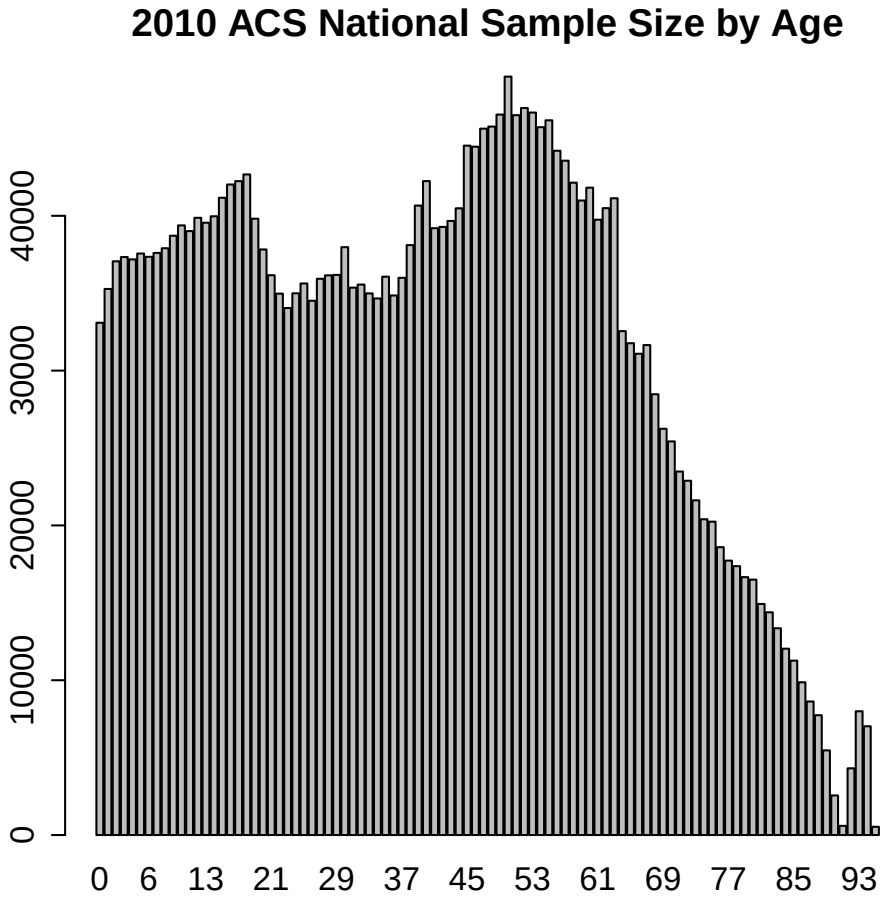

Figure 2: Histogram of the age distribution for the 2010 ACS 1-year national file. The upper tail of the distribution is particularly interesting. The national file is simply an aggregate of all the state files for that year. Because the mean-corrected top-coding is done separately by state, the resulting national age distribution has a very unusual tail.

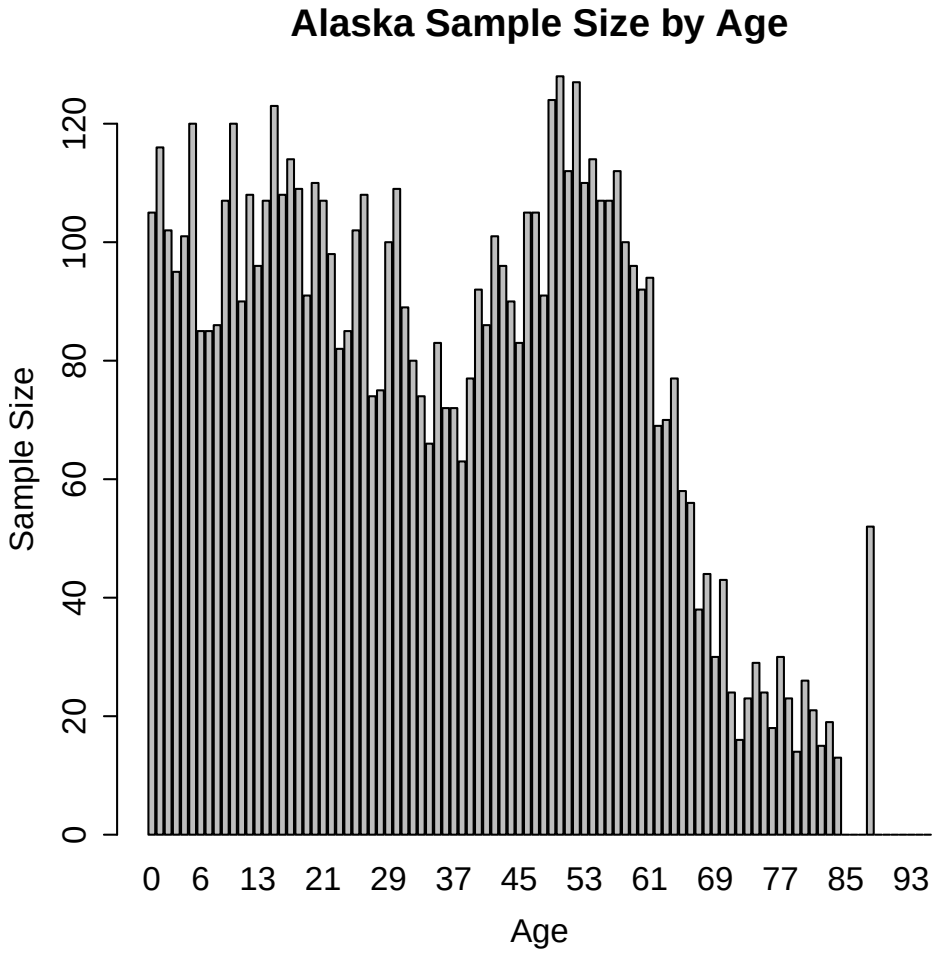

Figure 3: Histogram of the age distribution for the Alaska 2010 1-year ACS PUMS data. The data is the mean-corrected top-coded data. Note that using the mean-corrected top-coding method, the ages were truncated at 85 and each truncated observation was placed at 88 so that weighted mean of the adjusted data matched the weighted mean of the unadjusted data. Refer to Tables 1 and 2 to see the various truncation values and replacement values for 2010 (and other years) for each state.

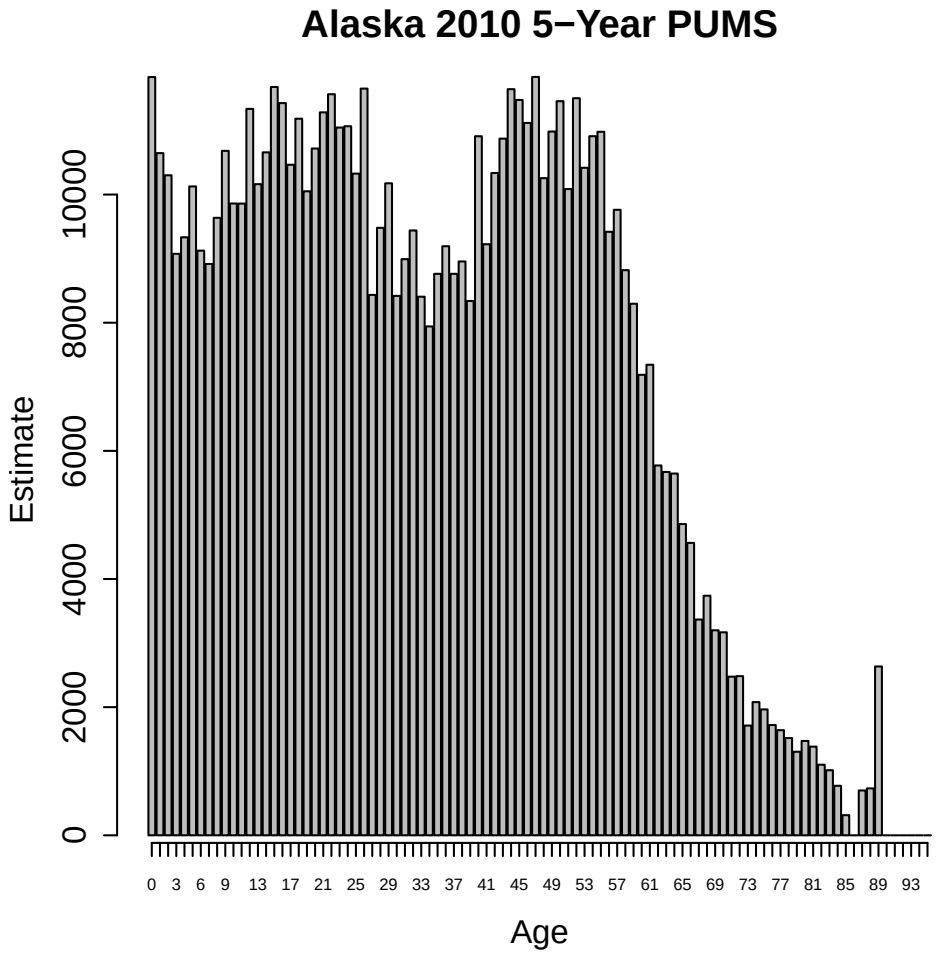

Figure 4: Histogram of the age distribution for the Alaska 2010 5-year ACS PUMS data. The 5-year state file is an aggregate of the yearly state files. Since the mean-corrected top-coding may change each year, the resulting 5-year age distribution has an unusual tail. The tail behavior is the result of the different replacement values for each year.

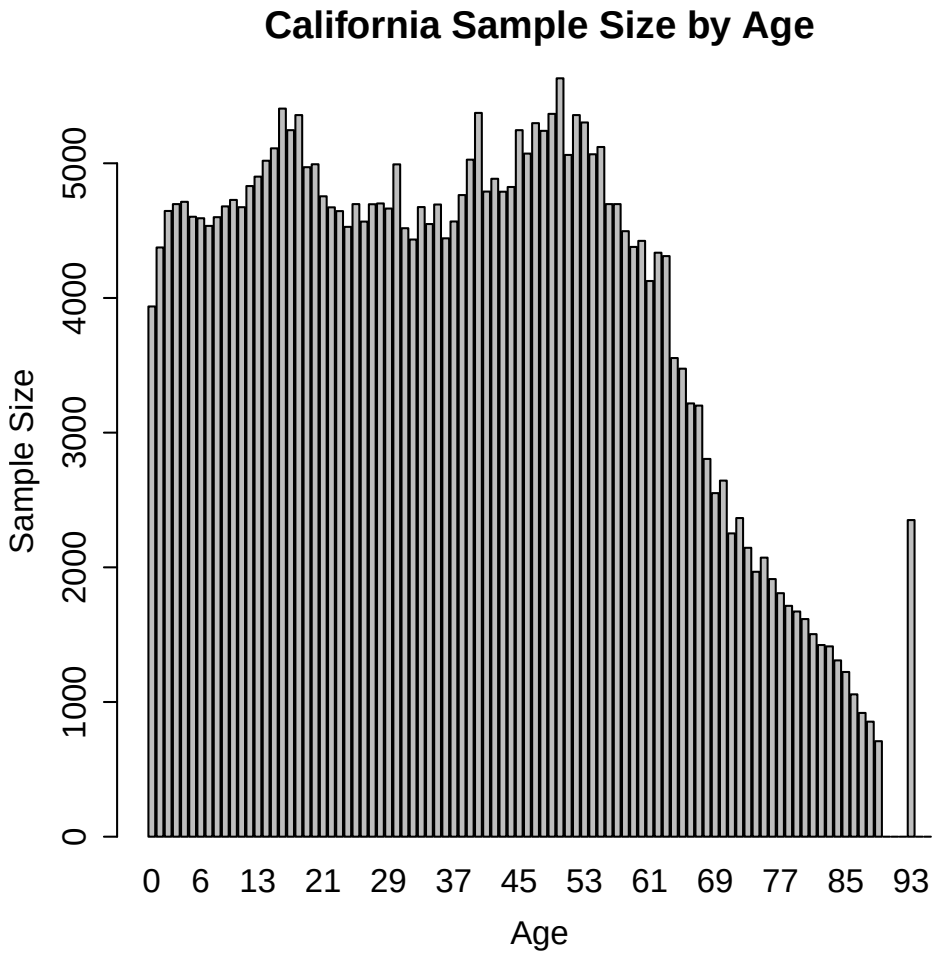

Figure 5: Histogram of the age distribution for California 2010 1-year ACS PUMS data. The data is the mean-corrected top-coded data. Note that using the mean-corrected top-coding method, the ages were truncated at 90 and each truncated observation was placed at 93 so that weighted mean of the adjusted data matched the weighted mean of the unadjusted data. Refer to Tables 1 and 2 to see the various truncation values and replacement values for 2010 (and other years) for each state.

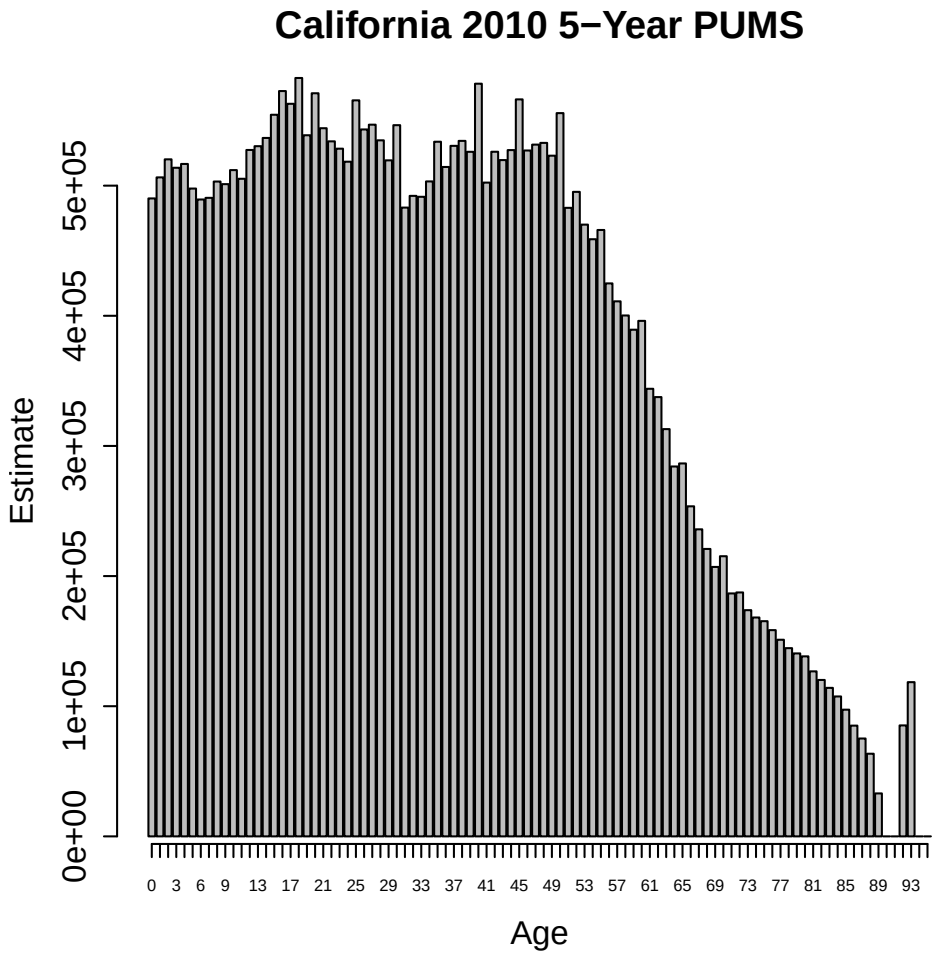

Figure 6: Histogram of the age distribution for the 2010 5-year ACS state file for California. The 5-year state file is an aggregate of the yearly state files. Since the mean-corrected top-coding may change each year, the resulting 5-year age distribution has an unusual tail. The tail behavior is the result of the different replacement values for each year.

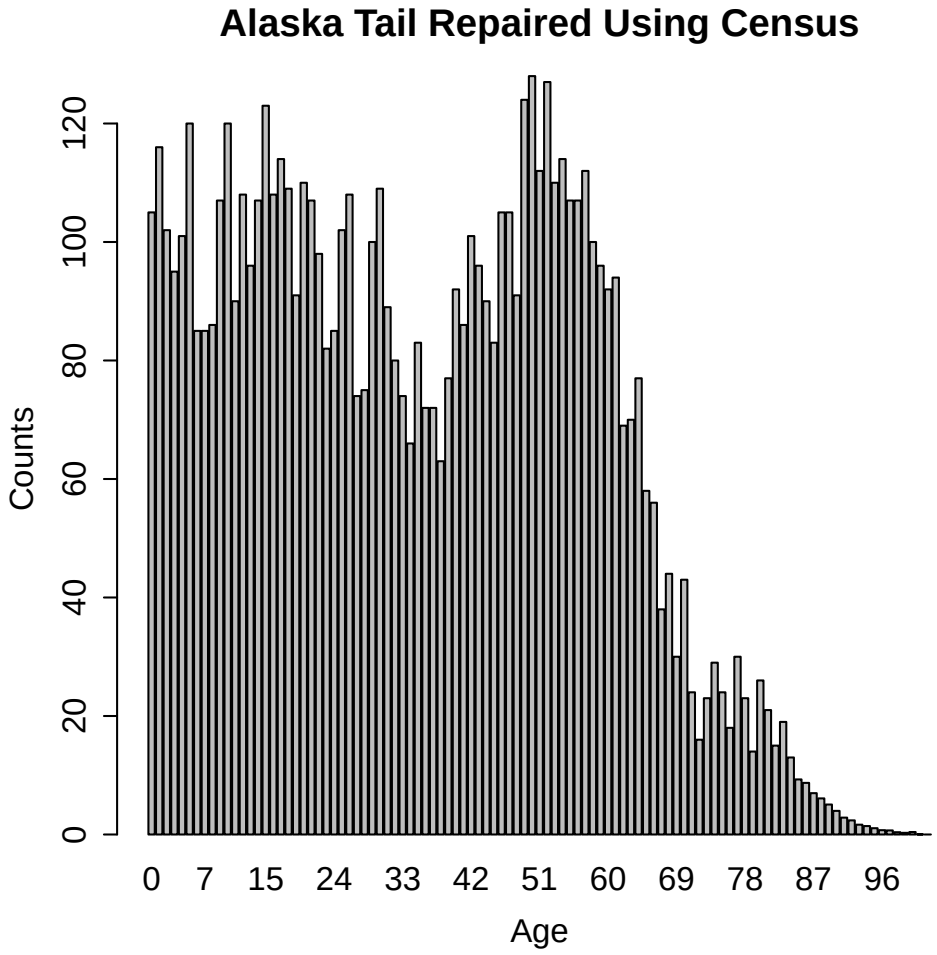

Figure 7: Histogram of the age distribution for the Alaska 2010 1-year ACS PUMS data. The portion of the data above the top-coded truncation value has been ratio-corrected to match the 2010 Census. Compare this histogram to Figure 3. This is a quite reasonable approach for the 2010 ACS PUMS, although there is probably less year-to-year variation in the reconstructed tail than in the actual population. It becomes less and less reasonable with each passing year; e.g., correcting the 2015 ACS PUMS to the 2010 Census is probably not especially useful.

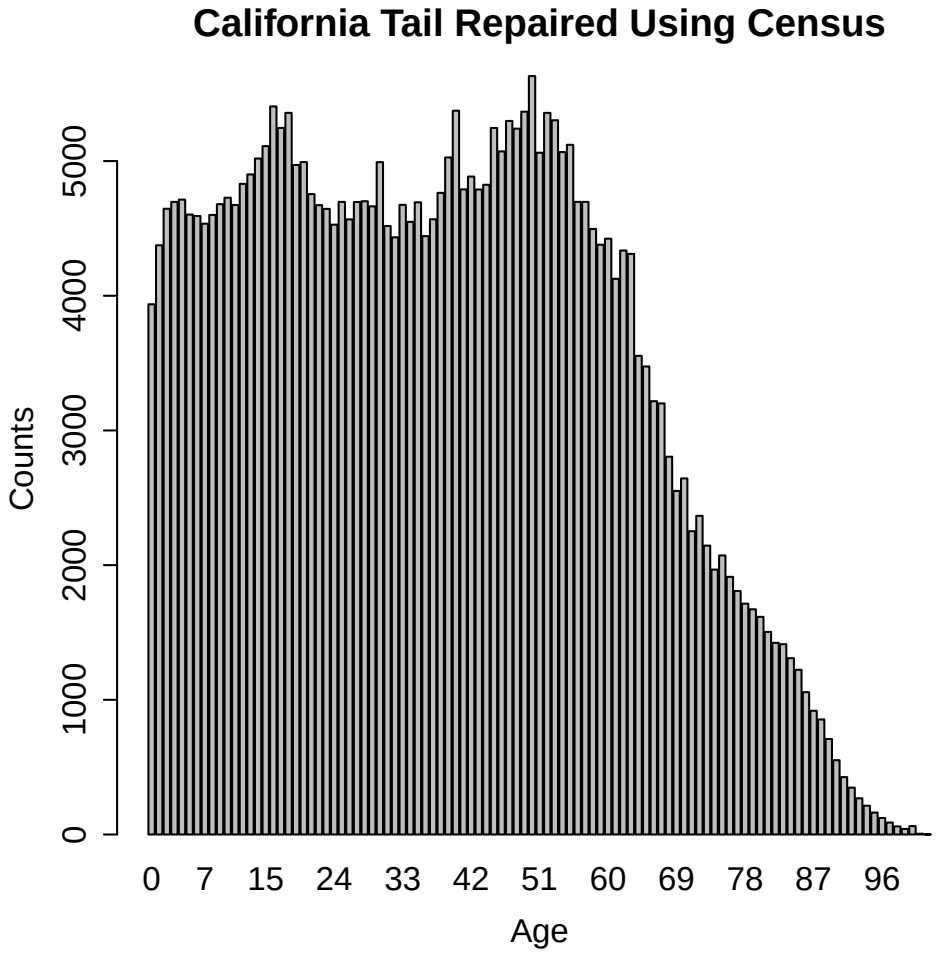

Figure 8: Histogram of the age distribution for the California 2010 1-year ACS PUMS data. The portion of the data above the top-coded truncation value has been ratio-corrected to match the 2010 Census. Compare this histogram to Figure 5. This is a quite reasonable approach for the 2010 ACS PUMS; because of the large population size there is probably much less age-to-age variation than in Alaska. See Figure 7 and its caption. It becomes less and less reasonable with each passing year; e.g., correcting the 2015 ACS PUMS to the 2010 Census is probably not especially useful.

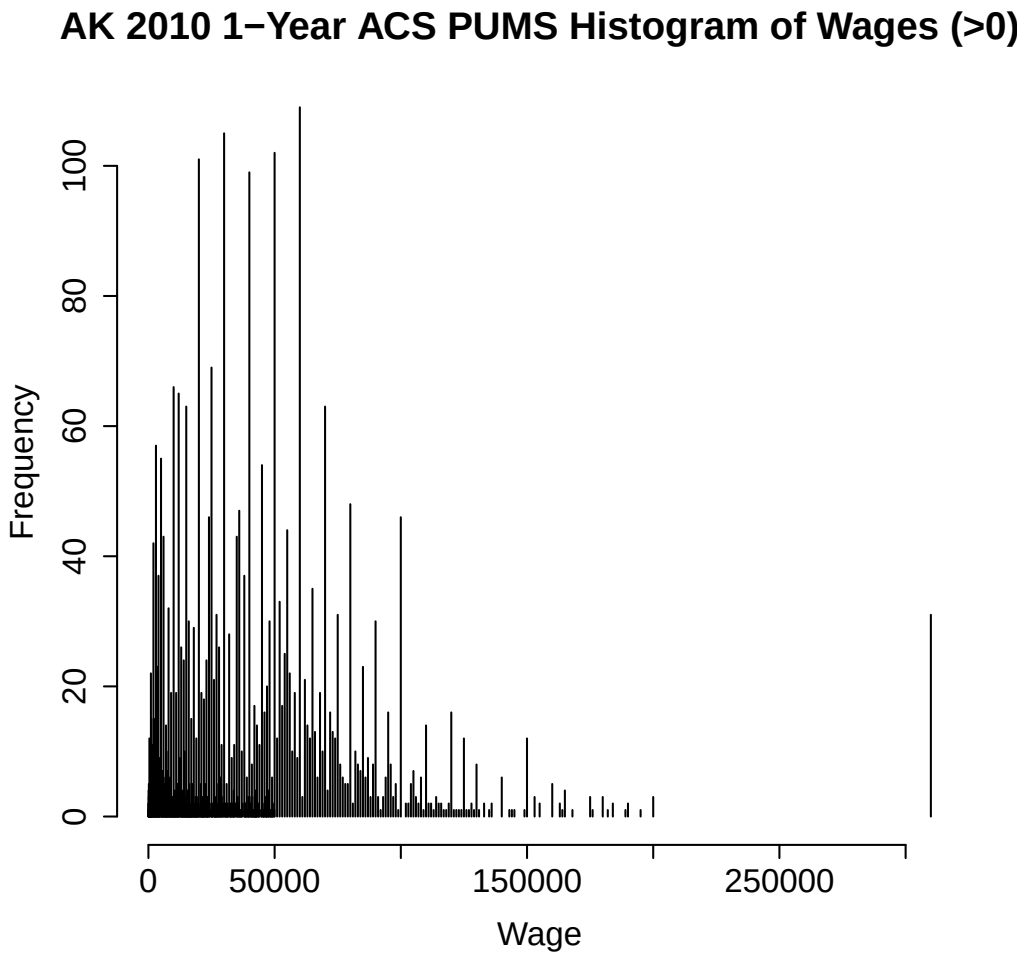

Figure 9: Histogram of wages for the 2010 1-year ACS PUMS for Alaska. Note that by using the mean-corrected top-coding method Alaska wages were truncated at $200,000 and each truncated observation was placed at $310,000. The large spikes at rounded values is typical of wage distributions.

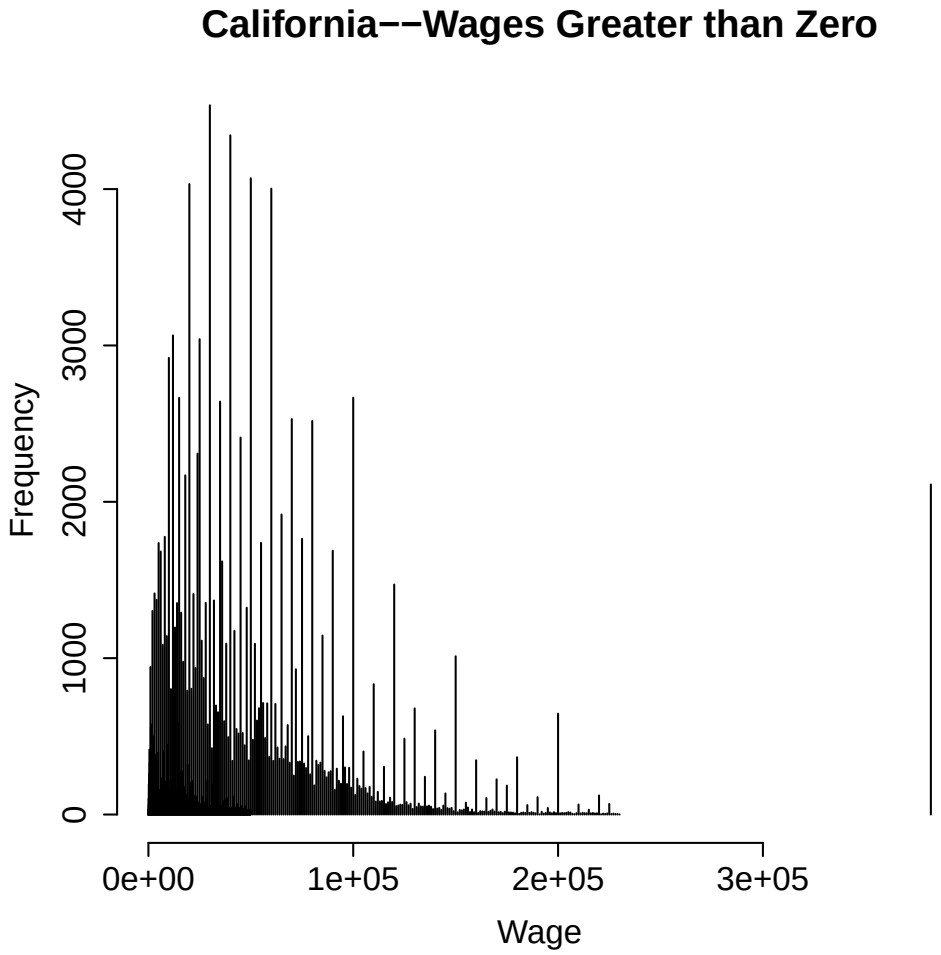

Figure 10: Histogram of wages for the 2010 ACS PUMS for California. Note that by using the mean-corrected top-coding method California wages were truncated at $230,000 and each truncated observation was placed at $382,000. The large spikes at rounded values is typical of wage distributions.

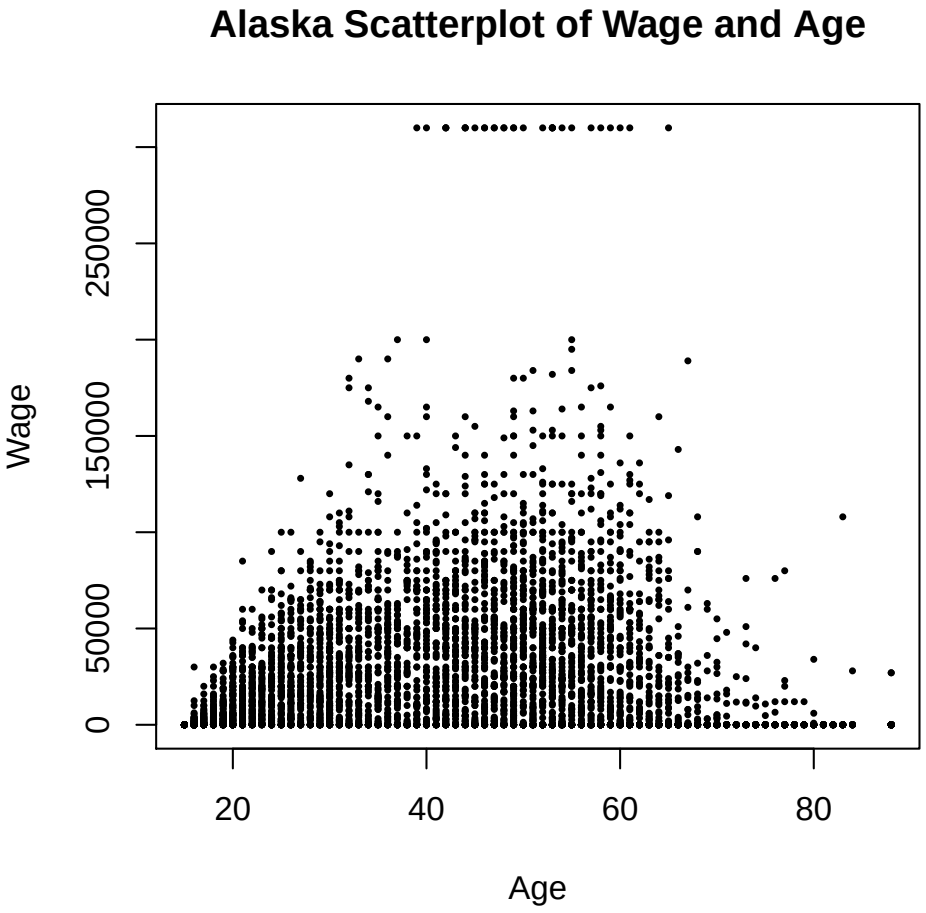

Figure 11: Scatterplot of wages by age for the Alaska 2010 1-year ACS PUMS. Note that both the top-codes of age and wage are visible, with truncation age of 85 and corresponding replacement age of 88, and truncation wage of $200,000 and corresponding replacement wage of $310,000, respectively.

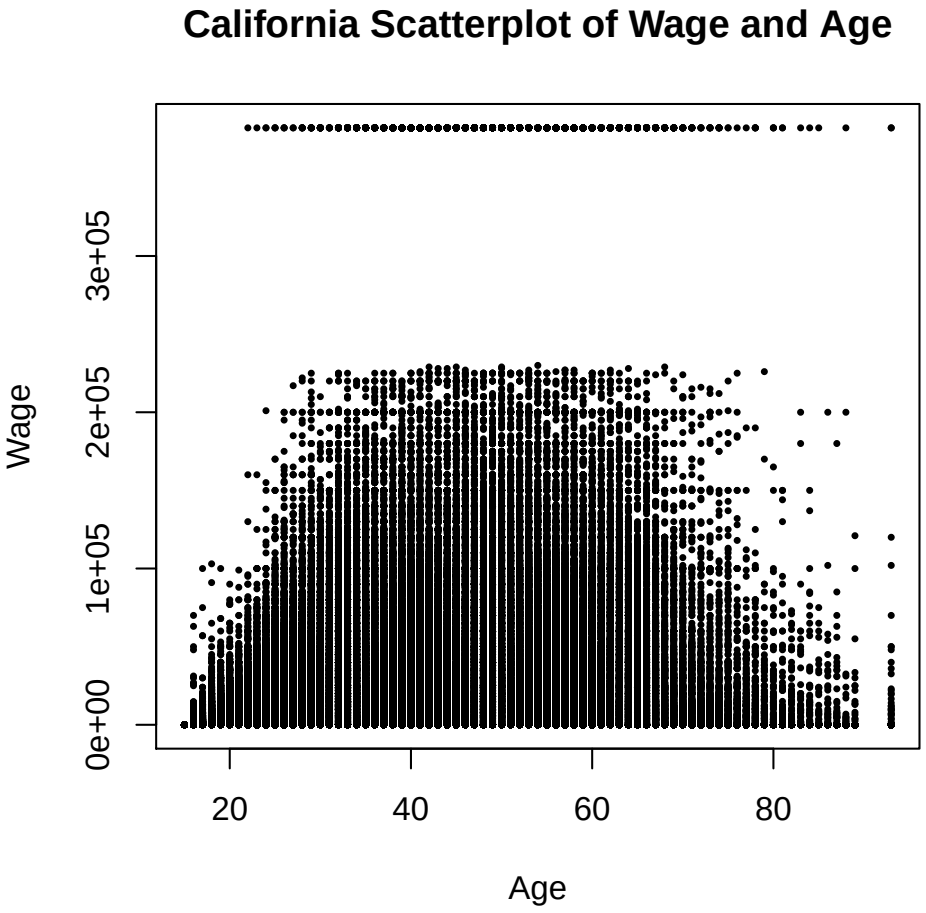

Figure 12: Scatterplot of wages by age for the California 2010 1-year ACS PUMS. The main part of the plot looks completely usual but the portion corresponding to large wages or large ages shows the clear effect of mean-corrected top-coding. Note that top-coding creates an intersection between the two top-codes in the top right corner.

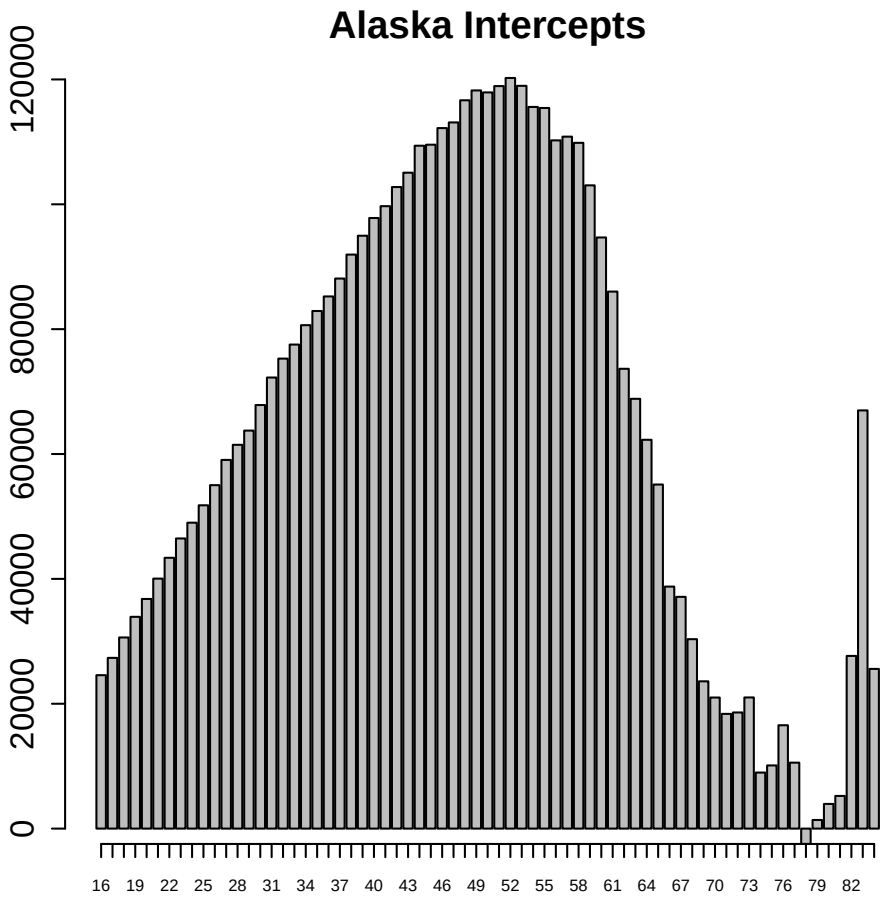

Figure 13: Barchart of the intercept coefficients of a simple regression of wages on age for the 2010 ACS 1-year PUMS data for Alaska. We performed successive regressions after removing the age categories one at a time, starting with age 15. The leftmost bar is the intercept estimated from all the data; the next bar is the intercept using all the data but that for age 16, etc.

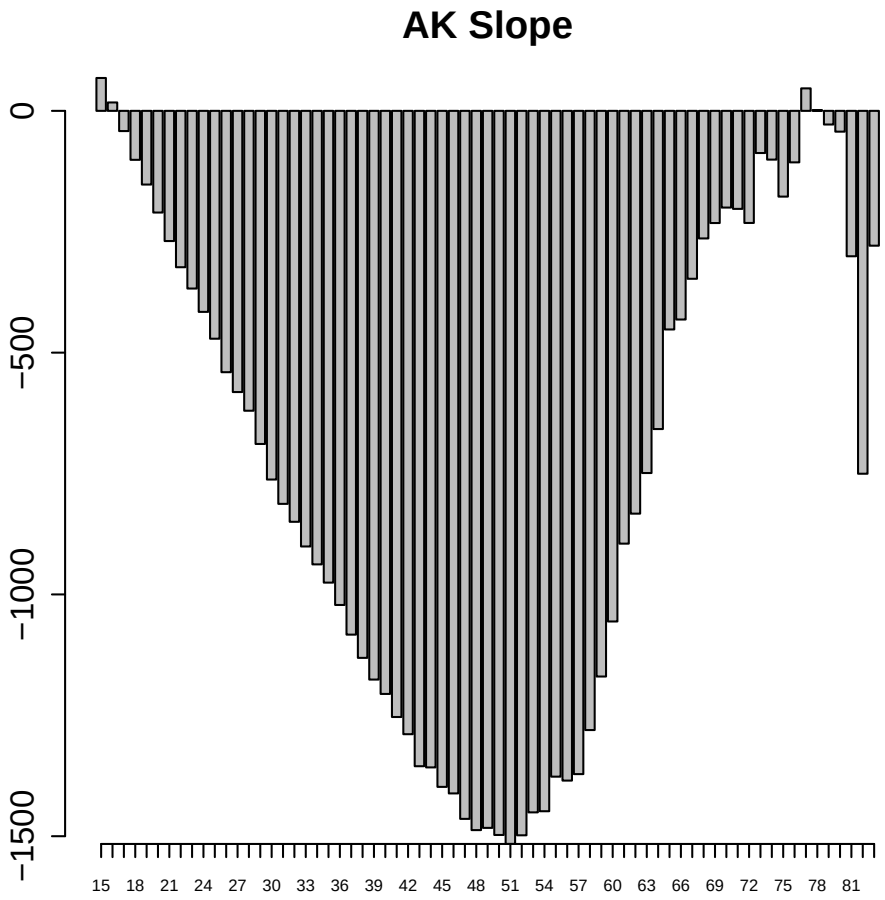

Figure 14: Barchart for the slope coefficients of a simple regression of wages on age for the 2010 ACS 1-year PUMS data for Alaska. We performed successive regressions after removing the age categories one at a time, starting with age 15. The leftmost bar is the slope estimated from all the data; the next bar is the slope using all the data but that for age 16, etc.

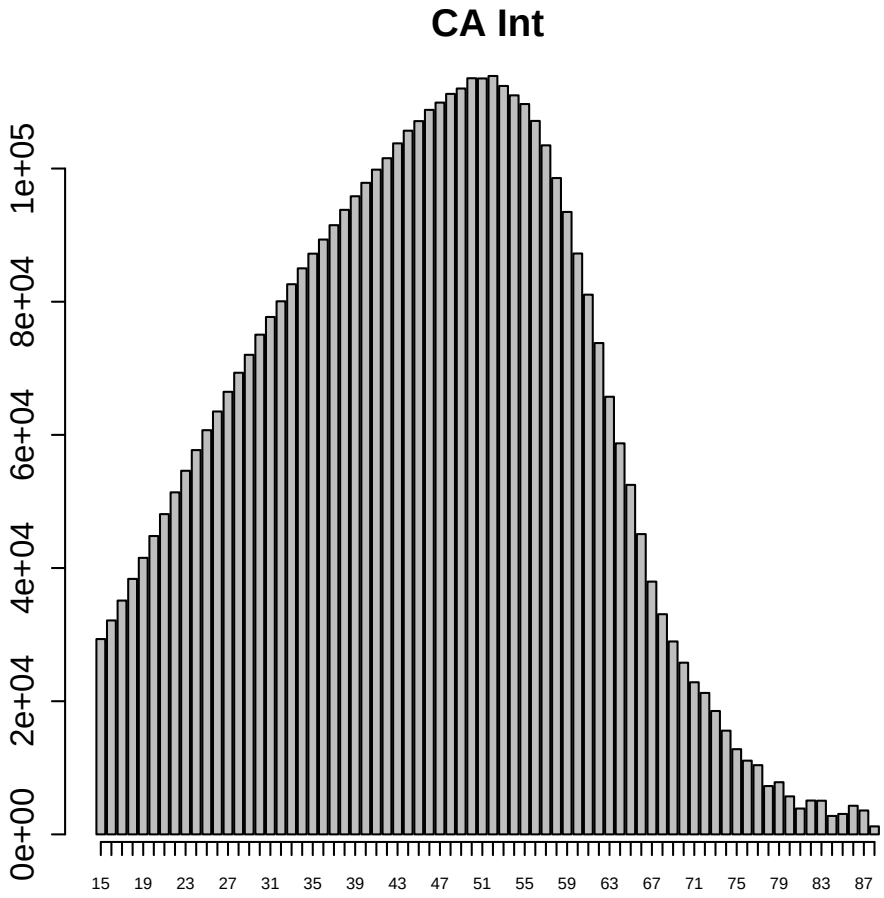

Figure 15: Barchart for the intercept coefficients of a simple regression of wages on age for the 2010 ACS 1-year PUMS data for California. We performed successive regressions after removing the age categories one at a time, starting with age 15. The leftmost bar is the intercept estimated from all the data; the next bar is the intercept using all the data but that for age 16, etc.

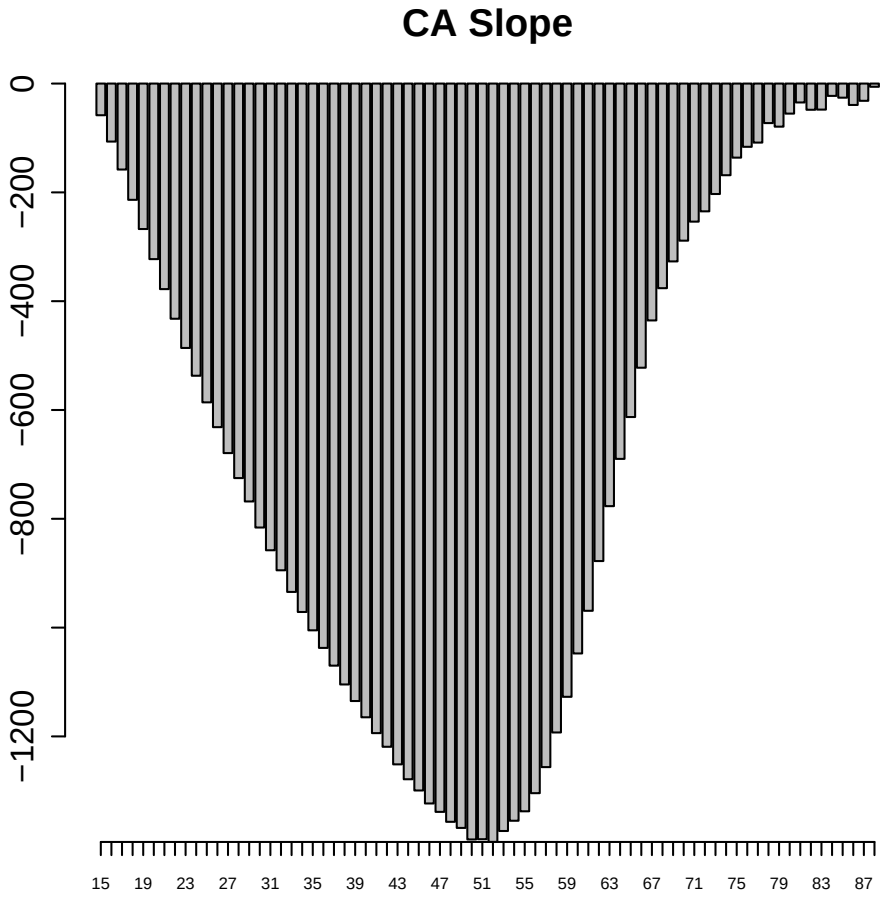

Figure 16: Barchart for the slope coefficients of a simple regression of wages on age for the 2010 ACS 1-year PUMS data for California. We performed successive regressions after removing the age categories one at a time, starting with age 15. The leftmost bar is the slope estimated from all the data; the next bar is the slope using all the data but that for age 16, etc.

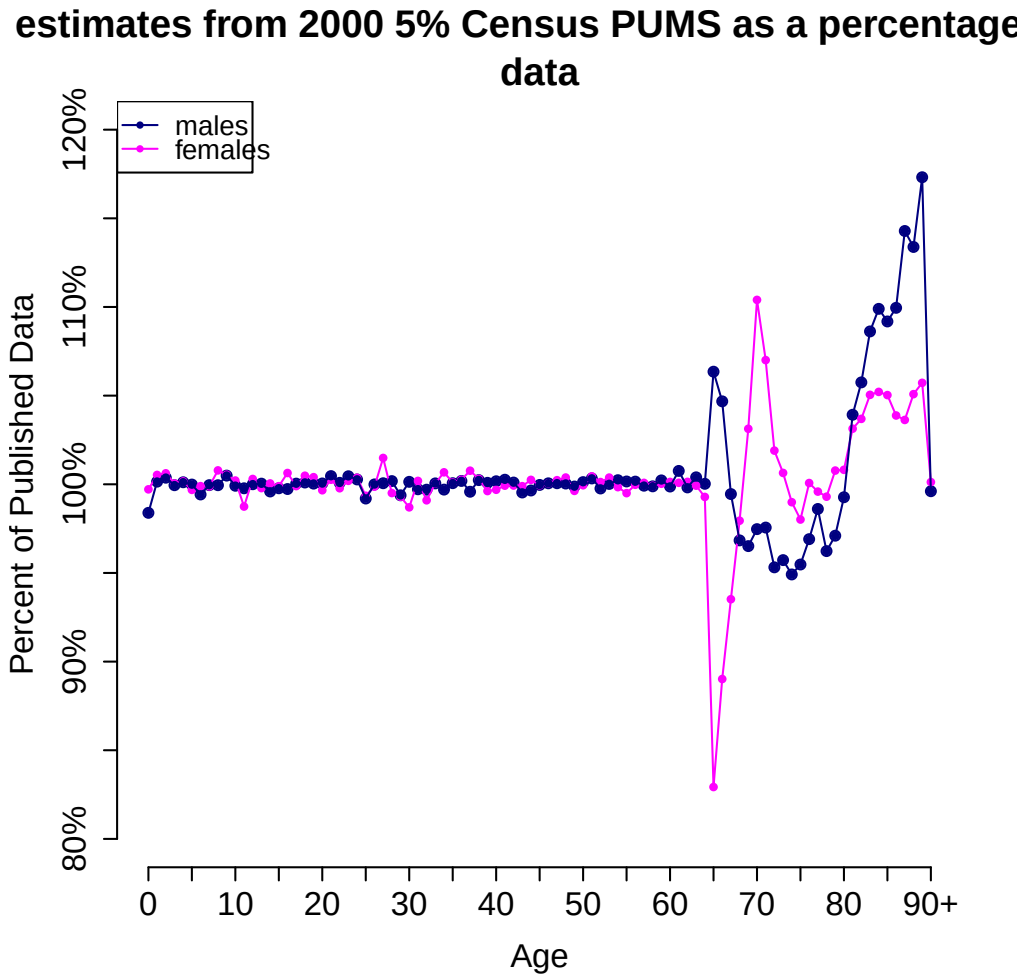

Figure 17: Recreation of Figure 1 in [1]. The 2000 Census 5% PUMS is split into males and females and then the sum of the person weights is divided by the published 2000 Census estimates. This graph shows that after age 65 the PUMS estimates differ dramatically from the published estimates, while below age 65 the fraction is very close to 100% for every age. This is simply a confirmation that our calculations are identical to those of Alexander et al. [1].

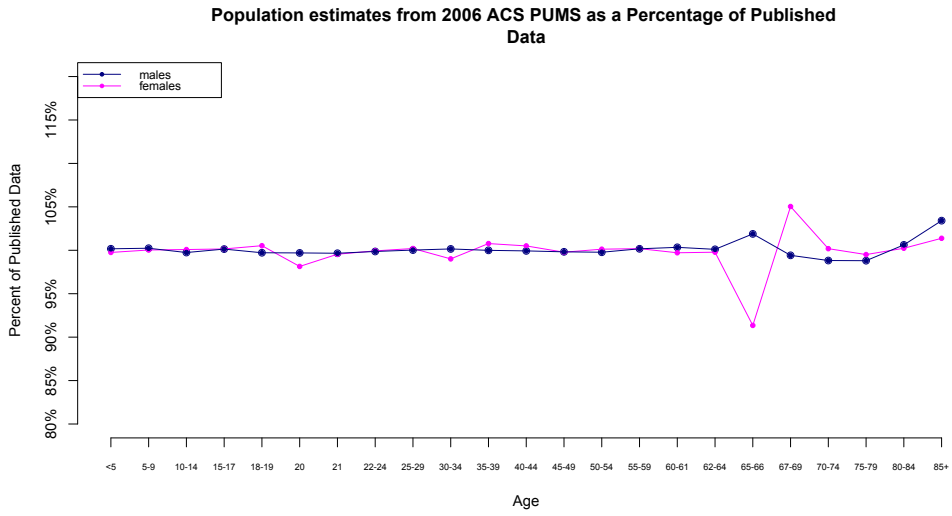

Figure 18: A close recreation of Figure 2 in [1]. The 2006 ACS PUMS is split into males and females and then the sum of the person weights is divided by the published 2000 Census estimates. This differs from their Figure 2 because the ACS only publishes age group estimates, not single age-year estimates. Once again, the estimates differ dramatically beginning at age 65, while all age groups below age 65 are close to 100%.

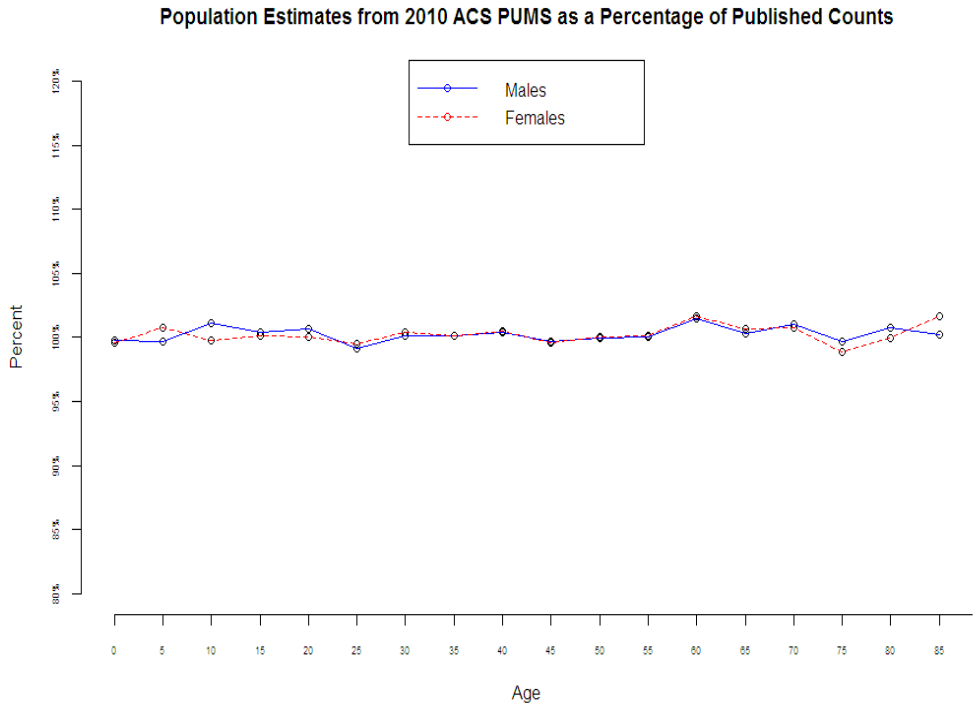

Figure 19: This graph was produced from the 2010 1-year ACS PUMS data using the same calculation as for Figure 18. Unlike Figure 18, none of the estimates are off by more than 1% for any age group.

Changed IPUMS Values

Serial Number	Person Order	Variable Changed	ACS PUMS Value	IPUMS Value	Type of Change
499584	2	pwgtp36	-1096	-96	switched with Serial Number 499584 Person Order 3
499584	3	pwgtp36	-96	-1096	switched with Serial Number 499584 Person Order 2
130235	2	pwgtp38	-1856	-856	deletion
309262	1	pwgtp62	-1505	-505	deletion
415031	1	pwgtp62	-1381	-381	deletion
546561	4	pwgtp38	-1083	-83	deletion
568705	3	pwgtp38	-3455	-455	deletion
1047767	1	pwgtp62	-1214	-214	deletion
1122204	1	pwgtp62	-1319	-319	deletion
1153692	2	pwgtp62	-1684	-684	deletion
1219000	2	pwgtp38	-1274	-274	deletion
621748	3	pwgtp2	-3160	-160	deletion
984734	3	pwgtp42	-1316	-316	deletion

Table 3: A list of the replicate weights from the 2010 ACS PUMS whose value changed from the Census Bureau data files to the IPUMS data base.

References

[1] Alexander, J.T, Davern, M., Stevenson, B. (2010). Inaccurate age and sex data in the Census PUMS files: Evidence and implications. *Public Opinion Quarterly*, 74:551–569.

[2] U.S. Census Bureau (2003). Census 2000, Public Use Microdata Sample, (PUMS), United States, Technical Documentation. Available at `http://www.census.gov/prod/cen2000/doc/pum.pdf`

[3] Bureau of the Census. (1992). Census of Population and Housing 1990: Public Use Microdata Sample U.S. Technical Documentation. Washington, DC: The Bureau.

[4] Hawala, S. (2003). Microdata disclosure protection research and experiences at the U.S. Census Bureau. *Workshop on Microdata, Stockholm, Sweden*. Available at `http://www.census.gov/srd/sdc/microdataprotection.pdf`

[5] Health and Retirement Study, (2010 HRS Core (Final) (v.3.0)) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2013).

[6] Office of Management and Budget, Federal Committee on Statistical Methodology. (2005). Statistical Policy Working Paper 22 (First version, 1994): Report on Statistical Disclosure Limitation Methodology. Confidentiality and Data Access Committee of the Office of Information and Regulatory Affairs.

[7] Office of Management and Budget, Federal Committee on Statistical Methodology. (2005). Statistical Policy Working Paper 22 (Second version, 2005): Report on Statistical Disclosure Limitation Methodology. Confidentiality and Data Access Committee of the Office of Information and Regulatory Affairs.

[8] Ruggles, S., Alexander, J. T., Genadek, K., Goeken, R., Schroeder, M. B., and Sobek, M. (2010). Integrated Public Use Microdata Series: Version 5.0. [Machine-readable database]. Minneapolis, MN: Minnesota Population Center [producer and distributor].

[9] U.S. Census Bureau. (2008). A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know. Washington, DC: U.S. Government Printing Office. Available at `http://www.census.gov/acs/www/Downloads/handbooks/ACSGeneralHandbook.pdf`.

[10] U.S. Census Bureau, American Community Survey Office. (2013). 2007-2011 PUMS Accuracy of the Data. Available at `http://www.census.gov/acs/www/Downloads/data_documentation/pums/Accuracy/2007_2011AccuracyPUMS.pdf`.

[11] National Institutes of Health. (2007). Growing Older in America: The Health and Retirement Study. National Institute on Aging. Available at http://hrsonline.isr.umich.edu/index.php?p=dbook.

[12] U.S. Census Bureau. (2009). Design and Methodology. American Community Survey. Washington, DC: U.S. Government Printing Office. Available at https://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf.

[13] U.S. Census Bureau. (2013). American Community Survey Information Guide. Available at http://www.census.gov/acs/www/Downloads/ACS_Information_Guide.pdf.

[14] U.S. Census Bureau. (2010). Public Use Microdata Sample (PUMS) Files: 2010 PUMS Top Coded and Bottom Coded Values.

[15] U.S. Census Bureau, Census History Staff. (2012). Questionnaires. Available at http://www.census.gov/history/www/through_the_decades/questionnaires/.

[16] DePersio, M., Lemons, M., Ramanayake, K. A., Tsay, J., and Zayatz, L. (2012). n-Cycle Swapping for the American Community Survey. In *Privacy in Statistical Databases,* vol. 7556 of *LNCS*. New York: Springer-Verlag. 143–164.