Disclosure Avoidance for Census 2010 and American Community Survey Five-year Tabular Data Products

Laura Zayatz, Jason Lucero, Paul Massell, Asoka Ramanayake

U.S. Census Bureau[1], Commerce/Census/SRD/5K011, 4600 Silver Hill Road, Washington, DC 20233-9100, Laura.zayatz@census.gov

Abstract: This paper describes the statistical disclosure avoidance techniques to be used for all U.S. Census 2010 and American Community Survey (ACS) five-year tabular data products. Many of these tables are published for very small geographic areas. The paper includes procedures for standard base tables, special tabulations, and a future online query system. Procedures include data swapping, rounding, collapsing categories, applying thresholds, table suppression, and generation of synthetic data.

Key Words: Disclosure Avoidance, Confidentiality, Public Use Data Products

1 Introduction

The U.S. Census Bureau collects its survey and census data under a pledge of confidentiality to our respondents. The agency also has the responsibility of releasing data for the purpose of statistical analysis. In common with most national statistical institutes, our goal is to release as much high quality data as possible without violating the pledge of confidentiality. We apply disclosure avoidance techniques prior to publicly releasing our data products to uphold our pledge.

We are nearing the end of the development stage of the disclosure avoidance procedures and software that will be used for all tabular data products from Census 2010 and American Community Survey five-year estimates. The techniques are designed to protect data confidentiality while preserving data quality. In Section 2 of this paper, we briefly describe the techniques that were used to protect tabular data from Census 2000. In Section 3, we explain why some changes in those techniques were necessary. In

---

1 This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Sections 4-7, we describe the techniques that will be used for Census 2010 base tables, ACS five-year base tables, special tabulations from both data collections, and a Microdata Analysis System under development. Section 8 offers a conclusion.

2 Disclosure Limitation for Census 2000 Tabular Data

2.1 Procedure for the 100% (Short Form) Data

Data swapping was used to protect short form tabular data (Zayatz, 2003). A small sample of households from the internal census data files was selected. Data from these households were swapped with data from other households that had identical characteristics on a certain set of variables but were from different geographic locations. Which households were swapped was not public information. The selection process was highly targeted to affect the records with the most disclosure risk. There was a threshold value for not swapping in blocks with a high imputation rate. Only records which were unique in their block based on a set of key demographic variables were swapped. The probability of being swapped had an inverse relationship with block size. In addition, records representing households containing members of a race category which appeared in no other household in that block had an additional p_1 probability of selection. All data products were created from the swapped file. For any tables that were iterated by race to be released, there had to be at least 100 people of a given race in a given geographic area.

2.2 Procedure for the Sample (Long Form) Data

Swapping was performed to protect the data. As with the 100% data, we used 2 different sets of key variables; one to identify the unique records and one to find the swapping partners. We held several variables fixed (unswapped). For example, travel time to work and place of work for a household would not make sense if swapped with a household geographically far away.

The procedure for producing the masked file was very similar to the procedure for the 100% data. Block group replaced block because block group is the lowest level of geography for publishing sample data. The threshold value for not swapping in block groups with a high imputation rate differed, and the probability that a unique record was selected for swapping had an inverse relationship with block group size The lower the sampling rate, the more likely the sample unique was not unique in the entire block group population. So a smaller sampling rate led to a lower chance of being swapped.

For any tables that were iterated by race to be released, there had to be at least 50 unweighted people of a given race in a given geographic area.

2.3 Procedures for the Special Tabulations

All special tabulations were generated from the swapped data files. All cell values were rounded according to the following scheme: 0 rounded to 0, 1-7 rounded to 4, 8 or greater rounded to the nearest multiple of 5. Totals were constructed before rounding, thus universes remained the same from table to table, but the tables were no longer additive. Quantiles (percentiles) were calculated in 2 ways. If they were calculated as an interpolation from a frequency distribution of unrounded data, no additional rounding was required. This was the technique used in the standard summary files. If they were point quantiles generated using SAS and Proc Univariate, they were rounded to 2 significant digits, and there need to be 5 nonoverlapping cases on either side of each quantile point. Thresholds on universes were often applied to avoid showing data for small geographic areas or small population groups. We often required 100 cases in universe for 100% data and 50 unweighted cases for sample data. Occasionally we required 3 unweighted cases in marginal totals for sample data for very small tables. Percents and rates were calculated after rounding. We allowed some exceptions when the numerator and/or denominator were not shown. The Disclosure Review Board also considered the mean cell size and the dimension of the tables when making decisions.

2.4 The Advanced Query System (AQS) of American Factfinder (AFF)

AFF was developed to allow for broader and easier access to the standard tables and to allow users to create their own data products. One part of AFF was the AQS. The goal of the AQS was to allow users to submit requests electronically for user-defined tabular data. A request would pass through a firewall to an internal Census Bureau server with a previously swapped, recoded, and topcoded microdata file. The table would be created and electronically reviewed for disclosure problems. If it was judged to have none, the table would be sent back electronically (Rowland and Zayatz, 2001).

3 Why Will the Techniques Used for Census 2000 be Changed?

3.1 The American Community Survey

For Census 2000, there was a short form questionnaire and a long form questionnaire. The Census Bureau attempted to get information on characteristics such as sex, age,

Hispanic/NonHispanic, race, relationship to householder, and tenure (owner or renter) from 100% of the population through what we call "short form" questions. The 100% data were published in the form of tables. Many of the tables were published at the block level. The average block contained 34 people. Some of the more detailed tables were published at the block group level. The average block group contained 1,348 people. Approximately a one in six sample of the population received the long census form which, in addition to the short form, collected information on characteristics such as marital status, education, ancestry, language, place of birth, citizenship, income, industry, and occupation. The sample data were also published in the form of tables. Some of the tables were published at the block group level. Some of the more detailed tables were published at the tract level. The average tract contained 4,300 people.

Census 2010 will consist only of the short form questionnaire with virtually the same questions that were used in Census 2000. The long form has been replaced with the American Community Survey (Torrieri, 2007). This survey collects data on approximately 2.5% of the population each year. Tables are published from a given year for geographic areas with a population of at least 65,000. They are published from a combination of 3 years of data for areas with a population of at least 20,000. They will be published from a combination of 5 years of data for block groups and tracts as well as other larger areas. Changes in data collection and publication often require changes to disclosure avoidance for data products.

3.2 Group Quarters (GQ) Data

It is very difficult to use data swapping to protect GQ data, which are data about people living in nursing homes, college dormitories, prisons, military barracks, homeless shelters, etc. Many of the GQ facilities have populations that typically hold people of a given age or sex. The main problem occurs when pairing people to be swapped because many pairs would not make sense, and these would be obvious in any released data products. For example, swapping an elderly person in a nursing home with a young person in a college dormitory does not make sense. Another technique is desired for these data.

3.3 Increase in the Number of Tables Containing a Given Variable

There has been a considerable increase in the number of tables that the Census Bureau plans to publish from the Census 2000 long form to the ACS 5 year table package. There are also several variables that appear in a large number of these mainly 2 and 3 dimensional tables. Such a variable may be cross tabulated with a large number of

other variables. This can cause a disclosure risk if there is a category for such a variable that has an unweighted count of 1 (representing only 1 household or 1 person) in a small geographic area. Tables could be linked together to form a microdata record for a person or household in the small geographic area.

As an example, consider the variable Means of Transportation to Work. There may be a person who is the only person in sample who rides a bicycle to work in some geographic area. A univariate table shows a weighted count of the number of people who rode a bicycle to work in that (say) tract, and let's say his weight is 30. A data user would see a 30 in the table cell of people who rode a bicycle to work in that tract. There is also a table that includes the weighted number of people who rode a bicycle to work by occupation. There would be one occupation category for people who rode a bicycle that shows a weighted value of 30 and the rest of the occupation categories are zeros. There might be about 20 tables that give cross tabs with people who rode to a bicycle to work like these. Income and race are two examples. The person with the weight of 30 is clearly one person. His 20 characteristics could easily be linked and form a microdata record for a small geographic area (tract in this example). This is a new problem and requires special attention in the disclosure avoidance procedures.

3.4 The AQS vs. a Generalized Microdata Analysis System (MAS)

The AQS was built for Census 2000. It will not be rebuilt for Census 2010 and ACS. Also, it was limited to table generation. We are developing a similar system that will work for Census 2010 data, ACS data, and data from other surveys. The system will perform various types of statistical analyses on the data, as well as table generation. The system is called the "Microdata Analysis System" (Steel and Zayatz, 2006).

4 Procedures for the Census 2010 Standard Tables

The swapping procedures for household data are essentially the same as those used for Census 2000. For GQ data, see Section 5.2 below on the ACS. The technique is the same for Census 2010, but there are fewer variables to synthesize.

5 Procedures for the ACS Five-Year Standard Tables

5.1 Household Data

The swapping procedures are similar to those used for Census 2000. Due to the increase in the number of tables and the number of variables appearing in many tables, we have increased the number of variables in the key used to determine which records have a disclosure risk. We have also increased the percent of records that will be swapped.

5.2 Group Quarters Data

We will use partially synthetic data to protect ACS GQ data (Hawala, 2008). Some values will be synthesized for at-risk respondent records. We use synthetic data techniques to impute these values. Predicted values are obtained for all respondents for this variable based on a model that uses the other non-identifying variables as predictors. We use predictive mean matching to find synthetic values. For each value to be synthesized, the absolute distances between its predicted and the other predicted values of the non-synthesized data values are computed. Then the synthetic value will be the observed value of the respondent that has the closest predicted value. Once a value is synthesized, it will be used as a predictor for synthesizing other variables. When synthesizing variables, it is also necessary to impose constraints on the model in order to prevent illogical response combinations in the data. For example, it is not possible to allow a synthetically generated age to be 10 years for a mother of three children.

5.3 Publication Rules

Medians or other quantiles will be calculated as an interpolation from a frequency distribution of unrounded data. They are not subject to rounding. Estimates in the form of means or aggregates, defined here as a sum of the values for each of the elements in the universe (e.g., the sum of the income of all households in a given geographic area), will be based on at least 3 sample cases. For the Selected Population Profiles, there must be at least 50 unweighted sample cases over the 5-year period in the universe (specific population subgroup) in a given geographic area for the profile to be released. Tables involving a geographic area other than current place of residence (such as workplace tables, place of birth, residence 1 year ago) crossed with characteristics other than current place of residence must have at least 50 unweighted cases in sample in the universe of the table over the 5-year period. Tables of unweighted counts of people and housing units may only be shown for areas where there are at least 3 occupied housing units in sample.

Tables with more than 100 cells cannot be released for block groups. If a table is iterated by a variable such as race/ethnicity or gender, the set of iterated tables should be considered as a single table. The iterated variable should be considered a dimension

when counting the number of cells in the table to determine if the set of iterated tables can be released for block groups. Certain other tables will not be published for block groups. They include: tables where the universe is restricted to the foreign born or a subset of the foreign born; tables containing estimates of or characteristics of non-citizens; tables containing characteristics of unmarried partners; tables containing estimates of or characteristics of people that were married, widowed, divorced, or became mothers within the last 12 months; tables containing characteristics of people living in Group Quarters; tables containing detailed type of Group Quarters (categories that can be shown at the block group level are Institutional and Non-Institutional); tables containing detailed language categories (categories that can be shown at the block group level are English, Spanish, Other Indo-European, Asian/Pacific Islander, and All Else); and tables containing specific type of disability, disabled by race, or number of disabilities other than "0," "1," or "2 or more."

6 Procedures for Special Tabulations

6.1 Census 2010

The procedures will remain virtually the same as those used for Census 2000 100% short form data.

6.2 The American Community Survey

The procedures will remain virtually the same as those used for Census 2000 sample long form data. Means and aggregates must be based on at least 3 values.

The universes allowed for Group Quarters data are as follows: Non-Institutional groupings including College Dormitory Facilities, Military Facilities, Other Facilities, and Institutional groupings including Nursing Facilities and Skilled Nursing Facilities, Adult Correctional Facilities, Juvenile Correctional Facilities, Other Facilities. For a given geographic area and a given data product (1, 3, or 5 year), there must be at least 50 unweighted cases in any given type of facility (as well as 50 in an Other category) and those 50 cases must come from at least 3 different facilities. Categories may be combined to reach these thresholds. Previously released requests will be considered to ensure that there are no complementary disclosure problems.

For Demographic Profiles from user-defined geographic areas (neighborhoods), all areas must have at least 300 (weighted) people in them. Using a computer program, the user-

defined areas will be compared with standard Census Bureau areas to make sure users cannot obtain data from very small geographic areas by subtraction. If such small areas are found, the boundaries of the user-defined areas must be changed.

If we receive requests for special tabulations where a given variable is cross-tabulated with many other variables, we may perform additional disclosure avoidance techniques such as more data swapping, more data synthesis, and thresholds and suppression prior to creating those tabulations.

7 The Microdata Analysis System

7.1 Goals of the System

The Microdata Analysis System (MAS) is currently under development at the Census Bureau (Lucero, Zayatz, and Singh, 2009). It is designed to allow data users to perform various statistical analyses of survey and census data (for example, regressions, table generations, generation of correlation coefficients, etc.) without actually accessing the underlying confidential microdata. The underlying confidential microdata will contain more detail than currently published data products. Census 2010 and ACS data will be available for analysis in the system. We would like to expand the system to include any and all data sets (including establishment data) and any and all types of statistical analyses. The MAS will probably be used by people with needs for fairly simple statistical analyses (news media, policy makers, teachers, and students). Users that have the need to use the underlying data for more exploratory data analysis will have to continue to use the public use microdata files (though they may not offer as much detail) or the Research Data Centers (though they are not as cheap or easy as using the MAS). At this point in time, a goal is to offer the MAS as a free service. Also at this point in time, we are not certain if we will keep track of all queries to the MAS. The system will be designed to prevent users and automated robot programs from bombarding the system with large numbers of queries. There will be no modification of data on the fly.

7.2 Confidentiality Rules

Here we give a brief overview of the confidentiality rules and procedures within the MAS. Much more detail can be found in (Lucero [1], 2009 and Lucero [2], 2009). Categorical recodes of actual raw variables will be used for universe formation. To form a universe, users would first select *m* recoded variables, then select up to *j* observed bin levels for each of the *m* recoded variables. A universe query on the MAS can be thought

of as a request for a set of cell counts from the *m*-way table of unweighted counts.

Each universe must pass a series of confidentiality rules and procedures. The MAS does not allow any universes to be derived from an *m*-way table that contains (*m*-1) dimensional marginal totals equal to 1 or 2. In addition, *n*, the total number of observations within a universe must meet a minimum size threshold. Given that the universe has passed these two thresholds, it is then subsampled by removing a very small subset of q observations at random. If the exact same universe is selected again by any user, then the exact same *q* observations are removed to yield the same subsampled universe as before. On the MAS, all statistical analyses are then performed on the new subsampled universe with *n-q* total observations.

In addition, the MAS implements other confidentiality rules depending on the type of analyses. For example, no more than 20 independent variables may be selected for any regression model. Only two-way and three-way interaction terms may be included in the regression model, however, no regression model with more than four variables may be fully interacted. All optional transformations are limited to a predetermined list. Each dummy variable must pass a minimum size threshold, or else it will be absorbed into the intercept along with the dummy variable that represents the reference category level. All diagnostic residual plots are based on synthetic residuals, which mimic the actual residuals vs. fitted values (Reiter, 2003). Estimated regression coefficients are passed back to the user without restriction.

8 Conclusion

While data swapping will still be the primary way of protecting Census 2010 and ACS five year tabular data products, some changes will be made to the Census 2000 disclosure avoidance procedures. These include the generation of partially synthetic data for Group Quarters respondents, an increase in the amount of data swapping and number of variables in the key used to find households with a disclosure risk, and the development of a Microdata Analysis System. Additional disclosure rules will be used to identify certain tables that will not be published for the smallest of geographic areas. These are the current disclosure avoidance plans, but they could be changed as they are still being evaluated.

References

American Community Survey Design and Methodology Report, available online at <http://www.census.gov/acs/www/Downloads/dm1.pdf>.

Hawala, S. (2008), "Producing Partially Synthetic Data to Avoid Disclosure," *Proceedings of the Section on Government Statistics*, American Statistical Association.

Lucero [1], J. (2009), "Confidentiality Rules for Universe Formation and Geographies for the Microdata Analysis System," Statistical Research Division Confidential Research Report Series, Census Bureau, to appear.

Lucero, [2] J. (2009), "Confidentiality Rule Specifications for Performing Regression Analysis on the Microdata Analysis System," *Statistical Research Division Confidential Research Report Series*, Census Bureau, to appear.

Lucero, J., Zayatz, L., and Singh, L., (2009), "The Current State of the Microdata Analysis System at the Census Bureau," *Proceedings of the American Statistical Association, Government Statistics Section*, [CD-ROM] (to appear), Alexandria, VA, American Statistical Association.

Reiter, J.P., (2003), "Model Diagnostics for Remote-Access Regression Servers," *Statistics and Computing*, 13, pp. 371-380.

Rowland, S. and Zayatz, L. (2001), "Automating Access with Confidentiality Protection: The American FactFinder," *Proceedings of the Section on Government Statistics*, American Statistical Association.

Steel, P. and Zayatz, L. (2006), "Description of a Microdata Access System" for Presentation to the Census Advisory Committee of Professional Associations, US Census Bureau, October 27, 2006.

Torrieri, Nancy K. (2007), "America is Changing, and So is the Census: The American Community Survey," *The American Statistician*, 61.1.

Zayatz, L., (2003) "Disclosure Limitation for Census 2000 Tabular Data" Working Paper #15, ECE/Eurostat workshop on statistical data confidentiality, <http://www.unece.org/stats/documents/2003/04/confidentiality/wp.15.e.pdf>.