Developing and Testing the Microdata Analysis System at the U.S. Census Bureau

Bryan Schar, Michael Freiman, Amy Lauger

U.S. Census Bureau
4600 Silver Hill Road, Washington, DC 20233

Abstract

Data users in government, private industry, non-profits and academia have substantial demand for data from the Census Bureau's censuses and surveys. Hence, the Census Bureau aims to disseminate data widely and with as much detail as possible while keeping the pledge of confidentiality given to all respondents. Although the Census Bureau produces many estimates, tables, and public use microdata , existing data products don't always meet users' needs. This paper describes the development and capabilities of the Microdata Analysis System (MAS), an online remote access system in development at the Census Bureau. The MAS will allow any user to request custom tables and analyses from the underlying microdata. Tables and estimates will be produced quickly and interactively for the user's specified geography and analysis variables. The system will also dynamically generate measures of variance for all estimates. Data will be protected to ensure respondent confidentiality. We describe the capabilities of the system and evaluate the planned disclosure avoidance methodology to protect the data and the system's data utility. We also outline challenges with the planned methodology for future research.

Introduction

"The Census Bureau's mission is to serve as the leading source of quality data about the nation's people and economy. We honor privacy, protect confidentiality, share our expertise globally, and conduct our work openly."

Effective data dissemination is critical to the Census Bureau's mission of being the leading source of quality data about the nation's people and economy. While the Census Bureau's data products are highly trusted and valued, users have sometimes expressed frustration with the difficulty of discovering Census Bureau content and making it useful for their needs. To address these concerns, the Census Bureau recently established the Center for Dissemination Services and Consumer Innovation (CEDSCI) with the mandate to create and implement enterprise-wide solutions to make the Census Bureau's data dissemination environment adaptive, customer-centric, open, and accessible (Blash, 2014).

As part of its initiative to transform the way the Census Bureau disseminates data, CEDSCI has identified the need for a data dissemination tool that

- Is readily available, free, and easy to use via the Internet
- Creates custom tabulations and estimates "on-the-fly"
- Allows users to define custom "universes" of interest
- Allows users to define custom geographies and categories
- Creates accurate variance measures for all estimates
- Produces data that meet confidentiality standards (Blash, 2015)

The Census Bureau currently has many programs and tools to disseminate and provide access to the data it collects from the decennial census, the American Community Survey (ACS), and the more than 130 other surveys it conducts (U.S. Census Bureau, 2016). These include, but are not limited to, standard tabulations and profiles, custom tabulations, public-use microdata, and the Federal Statistical Research Data Centers (RDCs). All of these programs are designed to provide data users with the highest quality and most relevant data possible while upholding the Census Bureau's legal and ethical obligations to protect the confidentiality of respondents. However, none of these programs fully meets the needs identified by CEDSCI. Many of the capabilities desired by the Census Bureau's user community would be included in the Microdata Analysis System (MAS), which is currently under development at the Census Bureau.

Current Data Products and Remote Access

Standard Tabulations and Profiles

The Census Bureau releases a multitude of standard tabulations and profiles. American Fact Finder (AFF) currently is the primary system that the Census Bureau uses to disseminate these products. It is free to access via the Internet, and contains numerous tables available at various levels of geography. However, AFF does not allow users to create custom tabulations based on Census Bureau data. Instead, users are only able to obtain pre-defined tables at pre-defined geographies from AFF.

For example, suppose a user wished to obtain an estimate from the ACS 5-year data about the number of people in a certain age group in an area comprising three tracts. Using AFF, the user may be able to acquire the individual estimates for each tract and combine the data manually. However, without access to the underlying microdata, the user would not be able to produce an accurate margin of error for this combined estimate. In order to estimate it, the user would have to make assumptions, which would likely be incorrect, about the relationships between the individual estimates.

Custom Tabulations

Data users who are unable to obtain what they need through AFF or other online tools can request a custom tabulation from the Census Bureau. However, these tabulations are infeasible for many data users due to cost and timing. For example, the minimum cost for an ACS special tabulation is $3,000, and many cost much more. Preparing and releasing a custom tabulation for the ACS usually takes at least eight weeks. In addition, the Census Bureau's Disclosure Review Board must review and approve all custom tabulation requests before work on the tabulation is started (Census, 2016a).

Public Use Microdata

Public-use microdata files contain individual record-level data from surveys and censuses. The Census Bureau makes these files freely available via the Internet to data users so that they can create their own tabulations and analyses. Given the high potential for the disclosure of confidential information about respondents, these files contain many confidentiality protections including de-identification, limiting geographic detail, coarsening, noise infusion, top coding, and in some instances subsampling. However, these additional protections reduce the utility of the files. The limit on geographic detail is often a particularly large hindrance for data users. For instance, the ACS public use data include no geographic identifiers below pre-defined areas of at least 100,000 people. Some other surveys have much higher population thresholds.

Federal Statistical Research Data Centers

If the data products as disseminated by the Census Bureau are not suitable for particular data users, they can consider accessing restricted microdata within Federal Statistical Research Data Centers (RDCs). Over 20 sites are located across the nation at various universities, non-profit research institutions, and government agencies. The RDCs offer access to data from the Census Bureau, the Agency for Healthcare Research and Quality, and the National Center for Health Statistics. Other agencies' data should be available in the future. The RDCs offer secure facilities managed by the Census Bureau to provide access to restricted use microdata. This means all research conducted using the RDCs must be done on-site (Census, 2016b).

To use the RDCs, researchers must first submit a research proposal that shows a clear benefit to the Census Bureau. In addition, each researcher must be a U.S. citizen, pass a background investigation, and swear to protect the data from disclosure under penalty of law. The proposal and approval process may take months and researchers must pay substantial fees for the time within the RDCs.

The Census Bureau must review and approve all output the researcher creates and wants to release from the RDCs. The output must fall within the scope of the researcher's proposal and must not reveal any confidential information about respondents. This review process can take several weeks and sometimes researchers are not able to release the output they would like.

While the RDCs are a valuable source of microdata access for many researchers, their use is limited by physical proximity to the researcher, the need to create a well-developed research proposal with clear benefits to the Census Bureau, time, and monetary cost.

Microdata Analysis System Design

The Microdata Analysis System (MAS) is a free online remote data analysis in development at the Census Bureau as part of the CEDSCI initiative. It will allow the public to conduct various analyses interactively using Census Bureau microdata.

The MAS is meant to be a successor to the Advanced Query System (AQS), which the Census Bureau released following Census 2000. The system was available to the public for only a short time before the system's contract expired. The AQS allowed users to produce tabulations from a query of internal use Census 2000 microdata. Users could select a population universe, geographic areas, and variables for tabulation that were not part of the standard tables released through AFF (Hawala, 2004).

The MAS has been designed to share many features of the AQS. Users will be able to restrict their analyses to particular universes of interest, request analyses at various levels of geography and for custom geographies built from collections of smaller geographies, and conduct analyses on combinations of variables that are not part of standard tabulations and profiles.

Advanced Query System (AQS) Disclosure Avoidance Methods

A major concern of any remote data access system is the disclosure of confidential information about respondents. This risk is somewhat mitigated because the source dataset used by the MAS has already been subject to some pre-tabulation disclosure protections. For example, the internal use ACS microdata used to create official tabulations are subject to data swapping and topcoding. However, while these protections are adequate for the creation of standard tabulations and profiles, a remote data analysis system with the planned functionality of the MAS requires more protections to avoid disclosing confidential information. Given the MAS is meant to be a successor to the AQS, the logical first step was to examine if the MAS could use protections similar to those used by the AQS.

AQS users could only create tables using a predetermined set of recodes. For example, users may only have been able to request a table of age, which had 14 predetermined categories. Users were able to create custom combinations based on those recodes, but they could not modify or further partition them. In addition, the detail of the recodes would change based on other table factors. For example, users requesting a table of age at a lower level of geography, such as the tract level, may have been limited to seven broad age categories rather than 14 detailed ones.

Another protection was to apply real-time checks to prevent the release of those tables considered too sparse, and thus a disclosure risk. For a table to be released by the AQS, the following had to be true for every geography requested:

- The mean number of unweighted observations in each interior cell of the table must be greater than a confidential fixed value m.
- The median number of unweighted observations in each interior cell of the table must be greater than a confidential fixed value n.
- Among non-zero cells, the proportion of interior cells with exactly one unweighted observation must be no greater than a confidential fixed value p (Hawala, 2004).

If all of these conditions were true, the AQS released the requested table to the user. Otherwise, the AQS would inform the user that the requested table could not be released. If the request was for a composite geography, such as a collection of tracts, every component of that composite geography would have to pass the above checks separately or the table would not be released. This rule prevented information in sensitive tables from being revealed through a differencing attack.

As an example, suppose an intruder requested a table for a tract that failed one of the three checks above. The system would inform the intruder that the table couldn't be released. Without the above restriction on composite geographies, the intruder could then request this same table twice: once for a composite geography that includes the sensitive tract in addition to a geography that would pass, and again for same composite geography that excludes the sensitive tract. By taking the difference between these two tables, the intruder could obtain the information about the sensitive tract that the system would not provide directly.

Testing AQS Disclosure Avoidance Methods

Staff in the Census Bureau's Center for Disclosure Avoidance Research recently tested the disclosure risk and data utility of using a set of disclosure rules for the MAS similar to those in the AQS. A brief summary of the method and results of this testing is below. A more detailed discussion is in Freiman, *et al.* (2015).

The 2009 – 2013 ACS 5-year dataset was used to create one-way tables of 23 variables, two-way tables based on combinations of 16 variables, and three-way tables based on combinations of 15 variables. These one-, two-, and three-way tables were produced for each state, county and tract in the country using sets of predefined variable recodes. The sparseness checks developed for the AQS were applied to these tables to determine whether they would pass. The measure of utility was the proportion of tables that passed. Figures 1, 2, and 3 provide the results of this evaluation. In these figures, the x-axis is the number of cells in the table and the y-axis is the pass rate for the relevant geography. The points represent the pass rates for individual variables or combinations of variables, while the red lines are curves of best fit created using a kernel smoother.

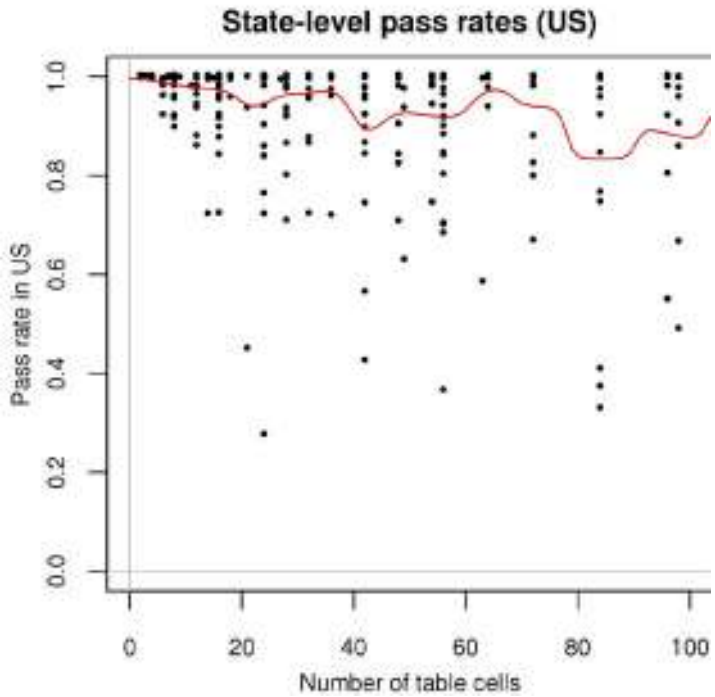

Figure 1: Proportion of states for which a table passes the proposed disclosure rules, plotted against the number of cells in the table.

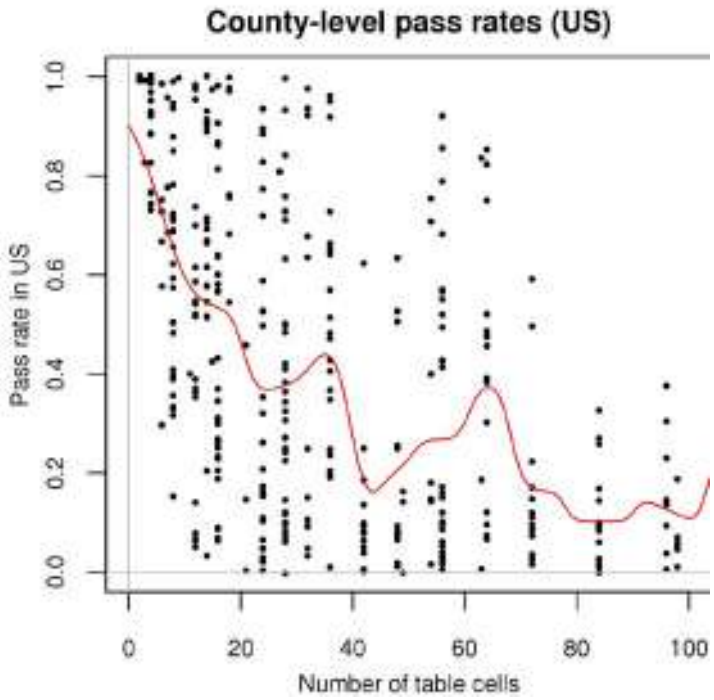

Figure 2: Proportion of counties for which a table passes the proposed disclosure rules, plotted against the number of cells in the table.

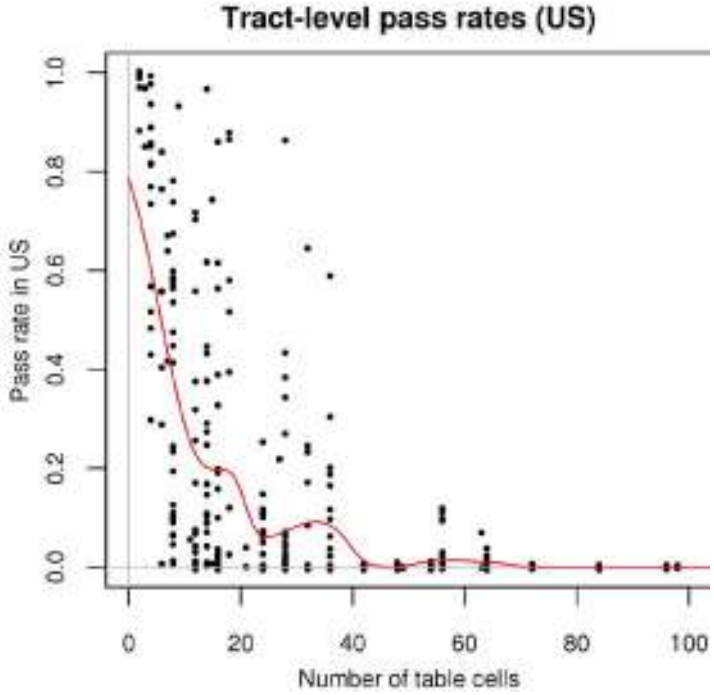

Figure 3: Proportion of tracts for which a table passes the proposed disclosure rules, plotted against the number of cells in the table.

As can be seen in these figures, the pass rates for tables are substantially high for state level tables but dramatically decrease at the county and tract level. For example, tables with about 15 cells had approximately a 53 percent chance of passing at the county level and an 18 percent chance of passing at the Census tract level. Pass rates are also substantially lower for composite geographies since each component must pass on its own.

The group also researched alternative thresholds for the sparseness checks and their effect on data utility. The goal was to find cutoffs that would maximize the utility of the MAS while maintaining confidentiality standards. The results of this research suggested that changing the cutoffs could not substantially increase the utility of the MAS without substantially increasing the disclosure risk.

This testing showed that relying primarily on table suppression to mitigate disclosure risk for the MAS would require the denial of a quite high proportion of sub-state tables.

Pre-Tabulation Disclosure Avoidance

Given the limitations of relying primarily on table suppression for disclosure avoidance, the team is now researching the use of pre-tabulation disclosure avoidance methods. These methods would apply additional protections to the source microdata. Such methods include noise infusion and synthetic data.

Noise Infusion
Noise infusion involves modifying continuous microdata values using a multiplicative or additive noise factor. For example, one way of applying noise to the MAS microdata would be to multiply each continuous value Z by a noise multiplier M drawn from a random noise distribution to create a perturbed value Y. That is, $Y = Z * M$. A limitation of noise infusion is that it can only be used to protect continuous variables.

Synthetic Data
Synthetic data involves creating models that capture key characteristics of the underlying dataset and then using those models to create a "new dataset with those same characteristics. With fully synthetic data, all of the records and variables in a dataset are replaced with simulated values. With partially synthetic data, only selected variables or selected records are replaced with simulated values.

Pre-tabulation disclosure avoidance methods decrease data utility in some ways. Perturbing the source microdata would lead to estimates similar to, but slightly different from, those used internally by the Census Bureau. However, in other ways, these methods may enhance data utility. Some of the rules that had previously assumed necessary in the MAS could be relaxed. For example, perhaps the rules regarding sparseness thresholds, variable recodes, or composite geographies could be changed.

The decision between post- and pre-tabulation disclosure avoidance methods comes down to tradeoffs in different types of utility. With an emphasis on post-tabulation disclosure avoidance methods, fewer estimates are released but they are more consistent with other published estimates. With pre-tabulation disclosure avoidance, more estimates are released but they are less consistent with the published estimates. The Census Bureau needs to consider what mixture of post- and pre-tabulation protections benefits users the most and best meets the data dissemination needs that CEDSCI has identified.

Conclusion

The Census Bureau is designing and testing the MAS as part of an initiative to transform the way it disseminates data. This online remote data analysis tool is being designed to allow the Census Bureau to meet the data dissemination needs identified by CEDSCI while upholding its legal and ethical obligations to protect the confidentiality of respondents.

Ongoing testing will determine the best ways to maximize the utility of the MAS while maintaining confidentiality standards. The preliminary results of this testing suggest that using pre-tabulation data perturbation to protect MAS data would increase its overall utility as a data dissemination tool. Methods such as increased data swapping, synthetic data, and noise infusion will be considered. These techniques will be compared to each other and to post-

tabulation protections based on measures of impact on the disclosure risk and utility of the MAS and feasibility of implementation.

References

Blash, Rebecca. (2014). "Center for Enterprise Dissemination Services and Customer Innovation: The Transition Team to Implement the Findings of the Data Dissemination Task Force: Charter." Internally distributed memorandum. December 10, 2014. Washington D.C.: U.S. Census Bureau. https://www.census.gov/sdc/615blash.pdf. Accessed January 22, 2016.

Blash, Rebecca. (2015). "Center for Enterprise Dissemination Services and Consumer Innovation: The Future of Data Dissemination." Presentation to the 2015 State Data Center and Census Information Center Annual Training Conference. June 2, 2015. Washington D.C.: U.S. Census Bureau.

Freiman, Michael H., Lauger, Amy, Lemons, Marlow, Schar, Bryan, and Hasenstab, Kyle. (2015). "Developing and Testing the Microdata Analysis System." In JSM Proceedings, Government Statistics Section, American Statistical Association.1016-1029. Alexandria, VA. http://www.eventscribe.com/2015/ASA-JSM/assets/pdf/233956.pdf. Accessed January 22, 2016.

Hawala, Sam, Zayatz, Laura, and Rowland, Sandra, (2004). "American FactFinder: Disclosure Limitation for the Advanced Query System." Journal of Official Statistics, Vol. 20, No. 1, 2004. Stockholm, Sweden. https://www.census.gov/srd/sdc/AdvancedQuerySystem.pdf. Accessed January 22, 2016.

U.S. Census Bureau, (2016a). "American Community Survey (ACS): Custom Tables." https://www.census.gov/programs-surveys/acs/data/custom-tables.html. Accessed January 22, 2016.

———, (2016b). "Federal Statistical Researcher Data Centers." http://www.census.gov/fsrdc. Accessed January 22, 2016.

_____, (2016c). "Our Surveys and Programs." http://www.census.gov/programs-surveys/surveys-programs.html. Accessed January 22, 2016.