2013

Expanding the Role of Synthetic Data at the U.S. Census Bureau

Ron Jarmin
U.S. Census Bureau, ron.s.jarmin@census.gov

Thomas A. Louis
U.S. Census Bureau, thomas.a.louis@census.gov

Javier Miranda
Center for Economic Studies, US Census Bureau, javier.miranda@census.gov

Expanding the Role of Synthetic Data at the U.S. Census Bureau

Abstract
National Statistical offices (NSOs) create official statistics from data collected directly from survey respondents, from government administrative records and from other third party sources. The raw source data, regardless of origin, is usually considered to be confidential. In the case of the U.S. Census Bureau, confidentiality of survey and administrative records microdata is mandated by statute, and this mandate to protect confidentiality is often at odds with the needs of data users to extract as much information as possible from rich microdata. Traditional disclosure protection techniques applied to resolve this tension have resulted in official data products that come no where close to fully utilizing the information content of the underlying microdata. Typically, these products take for the form of basic, aggregate tabulations. In a few cases anonymized public-use micro samples are made available, but these are increasingly under risk of re-identification by the ever larger amounts of information about individuals and firms that is available in the public domain. One potential approach for overcoming these risks is to release products based on synthetic or partially synthetic data where values are simulated from statistical models designed to mimic the (joint) distributions of the underlying microdata rather than making the actual underlying microdata available. We discuss recent Census Bureau work to develop and deploy such products. We also discuss the benefits and challenges involved with extending the scope of synthetic data products in official statistics.

Keywords
cofidentiality synthetic data, official statistics

Comments
Presented at World Statistical Congress 2013

Expanding the Role of Synthetic Data at the U.S. Census Bureau*

Ron Jarmin
U.S. Census Bureau, Washington, DC, USA, ron.s.jarmin@census.gov

Thomas A. Louis
U.S. Census Bureau, Washington, DC, USA and Johns Hopkins University, Baltimore, MD, USA, thomas.a.louis@census.gov

Javier Miranda
U.S. Census Bureau, Washington, D.C., USA, javier.miranda@census.gov

National Statistical offices (NSOs) create official statistics from data collected directly from survey respondents, from government administrative records and from other third party sources. The raw source data, regardless of origin, is usually considered to be confidential. In the case of the U.S. Census Bureau, confidentiality of survey and administrative records microdata is mandated by statute, and this mandate to protect confidentiality is often at odds with the needs of data users to extract as much information as possible from rich microdata. Traditional disclosure protection techniques applied to resolve this tension have resulted in official data products that come no where close to fully utilizing the information content of the underlying microdata. Typically, these products take for the form of basic, aggregate tabulations. In a few cases anonymized public-use micro samples are made available, but these are increasingly under risk of re-identification by the ever larger amounts of information about individuals and firms that is available in the public domain. One potential approach for overcoming these risks is to release products based on synthetic or partially synthetic data where values are simulated from statistical models designed to mimic the (joint) distributions of the underlying microdata rather than making the actual underlying microdata available. We discuss recent Census Bureau work to develop and deploy such products. We also discuss the benefits and challenges involved with extending the scope of synthetic data products in official statistics.

Keywords: confidentiality, synthetic data, official statistics

Introduction

National statistical offices (NSOs) face a constant tension between providing data users detailed data on the population and the economy and maintaining the confidentiality of the underlying information they use to construct these data products. This tension is particularly acute in cases where users require estimates for small domains. In this paper we discuss issues surrounding one promising approach – creation of synthetic data using multiple imputation techniques as in Rubin [1993] and Raghunathan et al. [2003]. We use recent experience at the U.S. Census Bureau to discuss both the benefits and challenges to NSOs and their data users from expanded use of synthetic data under a variety of settings. These include supporting the release of public-use microdata [see Kennickell, 1998, Kinney et al., 2011, Benedetto et al., 2013], and protections applied to data that underlie online tools

*Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

and applications (apps) [see Machanavajjhala et al., 2008]. We also discuss some practical issues and limitations surrounding further development of synthetic data by NSOs. The reality is that there is no single method of protecting confidentiality available to NSOs that would satisfy the entire range of user requirements. Thus, a variety of disclosure avoidance methods should be applied to products that are differentiated to meet user needs [Foster et al., 2010].

Synthetic Micro Data products at the U.S. Census Bureau

We begin by discussing two cases where the Census Bureau has utilized the disclosure avoidance offered by synthetic data techniques to release detailed public-use micro data products. The first is the SIPP-Synthetic Beta [see Benedetto et al., 2013] that combines survey data from the Survey of Income and Program Participation with administrative records from the Internal Revenue Services and the Social Security Administration. This product is the result of a collaboration between Cornell and Duke Universities and staff from the Census Bureau's Social, Economic and Housing Statistics Division. The second is the Synthetic Longitudinal Business Database (SynLBD) [see Kinney et al., 2011] which is the first business establishment level public-use micro dataset ever released by a U.S. statistical agency. This product is the result of a collaboration between Cornell and Duke Universities, the National Institute for Statistical Sciences and staff from the Census Bureau's Center for Economic Studies.[1]

These products were developed to provide broader access to rich data whose characteristics are similar to what was previously accessible only by Census Bureau employees or special sworn researchers at the Census Bureau's network of Research Data Centers (RDCs). In both cases, the product is a synthetic microdata set intended for use by researchers with advanced analytical skills and whose research requires access to microdata. In the case of the SIPP, public-use microdata were already available, however, a subset of users required SIPP records augmented with IRS earnings histories and SSA beneficiary information. In the case of the SynLBD, there were no public-use establishment level microdata with which to examine business dynamics topics such as job creation/destruction, business formation and growth and business exit.

These products are still considered experimental and, while publicly available, users must access both products through the Cornell Virtual RDC. This facilitates a research protocol where results obtained using the synthetic data can be replicated by Census Bureau staff using the gold standard confidential microdata. The protocol provides the users with feedback on the validity of inferences made with the synthetic data, and also provides the Census Bureau with valuable information that can be used to improve future versions of the synthetic data products.[2]

The SIPP-Synthetic Beta and the SynLBD allow the Census Bureau to meet the needs of a class of data users that falls between traditional public use products and restricted access micro data. In the case of data on business dynamics, the Census Bureau followed a strategy of providing multiple access

[1]NSF Grants #0427889 and #0339181 supported the development. The Synthetic Data Server at the Cornell VRDC is supported by grant #1042181

[2]Researchers interested in using these products can utilize a streamlined application procedure available on both the Census Bureau and Cornell VRDC websites. Application decisions are based solely on feasibility and generally occur within 10 business days. Detailed access protocols can be found for the SynLBD at http://www.census.gov/ces/dataproducts/synlbd/index.html, and for the SIPP Synthetic Beta at http://www.census.gov/sipp/synth_data.html.

modes to better meet the needs of data users. This strategy included the development of public-use tabulations in the new Business Dynamics Statistics program [see Haltiwanger et al., 2008] as well as restricted access to the gold standard Longitudinal Business Database (LBD).[3]

Synthetic Data to Support Tools and Apps

Synthetic data serves two purposes that can't be met with either public tabulations or the gold standard microdata. First, they provide much easier access to microdata while providing an avenue to obtain analytically valid results.[4] Second, synthetic data allows those sophisticated researchers whose research question requires access to the gold standard a way to explore the data and develop and test code outside the RDC environment, thereby increasing their productivity once in the RDC lab.

An important and potentially major area of expansion for NSOs is the use of synthetic data in support of online tools and apps. With the explosion of the Internet and mobile devices users increasingly want instant access to data that is accurate, timely and geographically specific. Moreover, they want the ability to mash it up with data from an increasing number of private data providers. Failure by NSOs to adapt their data products to meet users demands will result in official statistics becoming less relevant and opens the possibility that decision makers will rely on data products that do not meet the rigorous quality standards of official statistics, but possess other appealing characteristics.

The value of synthetic data for aiding NSOs like the Census Bureau in providing data to support online tools and apps stems from its ability to support small domain estimates. In particular, synthetic data products can provide estimates for much smaller geographic areas than traditional data products without risking respondent confidentiality. This clearly has advantages for creating mobile apps that provide users information about their current location or other small areas of interest.

The Census Bureau currently utilizes partially synthetic data to support its *OnTheMap* and *OnTheMap for Emergency Management* online mapping tools.[5] *OnTheMap* allows users to specify custom geographic areas to display data on worker residence and work locations. *OnTheMap for Emergency Management* mashes up *OnTheMap* data with information on disaster events from various Federal agencies to allow users to examine the impact of such events on workers and their employers. The disclosure protections that make these products possible are described in Machanavajjhala et al. [2008].

The Census Bureau has recently released apps that deliver traditional products in new ways. The *Americas's Economy* mobile app allows users to get data on key economic indicators on their mobile devices in a user friendly way. Perhaps more important is the recent release of the Census Bureau API (Application Programming Interface) that makes several popular datasets such as the American Community Survey available to external application developers.[6]

Currently app developers are limited to pre-defined tabulation versions of the data pushed out through the API. Thus, developers can't, for instance, create estimates for custom geographic areas that are based on the user's current location or other user-defined areas. This type of functionality is common in many popular apps and is clearly something users value. This constraint also limits how

[3]These products are aimed at users with different requirements and skill sets. For details on obtaining access to restricted-access "gold-standard" micro data via Census Bureau RDCs see http://www.census.gov/ces/rdcresearch/

[4]Researchers accessing the synthetic data can request validation of their research results.

[5]These tools are available at http://onthemap.ces.census.gov/ and http://onthemap.ces.census.gov/em.html

[6]See http://www.census.gov/developers/

developers can mash up Census Bureau data with other data sources. From a disclosure avoidance perspective, this is probably a good thing since such mash-ups increase the risk of revealing respondent information. Synthetic estimates could be useful here; synthetic data could be pushed through APIs to allow developers to create useful apps that use accurate Census Bureau estimates to inform users without risking respondent confidentiality.[7]

Issues and Challenges of Expanding the Use of Synthetic Data

It is perhaps too early to determine how data users will respond to official statistical products that utilize synthetic data. As with other disclosure avoidance techniques, some valuable information must be lost in the construction of synthetic data. Further, the characteristics of a synthetic data product will depend on the models used to generate the data. These will limit the range of valid applications in ways that may not be obvious to users. Sophisticated users are rightly concerned that inferences drawn from synthetic data may not always be valid; a problem exacerbated in the case of small area applications. Thus, a mechanism like that currently used for the SynLBD and SIPP Synthetic Beta will need to be in place to enable novel uses of synthetic data. This mechanism places a burden on the scarce resources of NSOs, but demands could be minimized by creating automated tools to validate user results in real time and thus provide information on the classes of uses where synthetic estimates yield erroneous inferences. In all cases NSOs should make abundantly clear the potential limitations of synthetic data as they strive to make data more easily accessible and useful.

Using synthetic data requires users to combine multiple implicates to obtain valid estimates. While this may not be a burden for sophisticated users, recent experience with multi-year estimates from the Census Bureau's American Community Survey suggest that a large segment of the user community will find this challenging. Thus, tools that execute combining rules should be developed.

In cases where users employ synthetic data directly through apps and online tools such as *On-TheMap*, they may be unaware of what goes on behind the scene and believe the data are more accurate than is the case (many data users already do this with more traditional products). Thus, enhanced communication of statistical and other uncertainties without adversely impacting the functionality of apps and tools need to be developed.

These issues highlight what is perhaps the largest impediment to increased use of synthetic data at NSOs and other statistical organizations – insufficient resources and staff with the skills required to develop, deploy and support new synthetic data products. In the case of the Census Bureau, all the products and tools described above were developed with heavy collaboration from academic researchers and with financial support from external foundations and institutions. While the collaborations that led to the products and tools described above greatly expanded the number of staff with such skills, they are still limited to a small number of individuals in the Research and Methodology Directorate and a few staff in the Economic and Housing Statistics Division. A combination of hiring staff with the required skills and training existing staff will be required in order to significantly expand the use of synthetic data.

[7] A research collaboration between Data Science for Social Good, the University of Chicago Computation Institute, and the Census Bureau's Center for Economic Studies is currently exploring ways to deploy synthetic data through APIs and to make large size data custom accessible through apps on mobile devices.

Next Steps

Modern societies require ever larger amounts of information to adjust and respond to the challenges posed by a dynamic and increasingly competitive integrated world economy. Businesses, households, and policymakers need rich, timely and accurate data to make informed decisions. In the absence of reliable official statistics, they will make the best of what is available which may, or may not, be constructed with the same level of rigor and quality control. NSOs have a wealth of information from survey and administrative sources to meet the needs of data users but are constrained in what they can release by the same confidentiality pledge that allows them to collect such high quality data in the first place. Synthetic data products offer a way to expand the amount of information NSOs can release to data users while maintaining the confidentiality of respondents.

In this paper, we described recent efforts at the Census Bureau to develop and deploy synthetic data products and the challenges to expanding their use. First, producing synthetic datasets requires specialized knowledge that is currently not widely available inside statistical agencies. Second, the use of synthetic datasets poses challenges to users. They may not fully understand limitations and might be challenged by the additional computations required to generate statistical inferences. NSOs will need to devote additional resources to educate the public regarding the responsible use of these data. Third, despite the potential benefits that synthetic datasets offer to the public, the resources available to statistical agencies to expand their offering of these products are limited.

These challenges are not going to disappear, and NSOs risk becoming increasingly irrelevant if they fail to produce the timely and accurate data users demand. The Census Bureau has taken on these challenges by partnering with academe and funding institutions to advance the research and implementation necessary to create, improve and make available these data products. This partnership has been fruitful, new data products been successfully developed and deployed, and Census Bureau staff have gained valuable knowledge, skills and experience by working with leading academic researchers. Partner researchers have tapped the knowledge and experience of economists and statisticians inside statistical agencies who are familiar with these data and their uses, and so have benefited from the insights and depth of knowledge inside statistical agencies. This "two-way street" has empowered creation and access to synthetic datasets.

Synthetic data applications are now being developed with an eye to making additional information available to a wider group of users. These new applications are being developed through entrepreneurial efforts by individuals, inside and outside the Census Bureau, who understand the needs of data users. Looking ahead, NSOs need to find ways to scale these efforts to enable them to expand the use of synthetic data beyond the small number of datasets currently available. We believe this is the best way to provide external developers the ability to more fully incorporate rich, accurate and reliable official statistics into apps and online tools that give users the flexibility to generate custom small area estimates.

To accomplish this, NSO's need to facilitate and foster collaboration between internal and external researchers to develop and deploy new synthetic data products. This collaboration relies on access by external researchers to confidential data inside secure environments. Feedback loops, like those currently employed at the Census Bureau, are needed to provide users information about the reliability of synthetic data products and to give development teams information to improve subsequent products. We need to build up the skill sets of staff at NSOs to oversee and undertake this kind of work. In this regard, academic institutions could partner with NSOs to develop relevant courses

and conduct training.

NSO staff are best positioned to fully understand the breadth and depth of their data resources. They are sworn to protect respondent confidentiality, but within that constraint want to make information broadly available in as much detail as feasible. We believe resolving this tension and making data available to users on a multitude of platforms as needed requires increased use of synthetic data along with clear communication of its strengths and weaknesses. NSOs have the responsibility to take the lead in these and other activities to increase the relevance and accessibility of high quality and reliable official statistics.

References

Gary Benedetto, Marth Stinson, and John Abowd. The creation and use of the SIPP Synthetic Beta. mimeo, U.S. Census Bureau, 2013. URL `http://www.census.gov/sipp/SSBdescribe_nontechnical.pdf`.

Lucia Foster, Ron Jarmin, and Lynn Riggs. Resolving the tension between access and confidentiality: Past experience and future plans at the U.S. Census Bureau. *Statistical Journal of the IAOS*, 26:113–122, 2010.

John Haltiwanger, Ron Jarmin, and Javier Miranda. Business Dynamic Statistics: An overview. Business Dynamic Statistics Briefing, Kauffman Foundation, 2008. URL `http://www.census.gov/ces/pdf/BDS_Overview_2009.pdf`.

Arthur Kennickell. Multiple imputation in the Survey of Consumer Finances. Working paper, Board of Governors of the Federal Reserve System, 1998. URL `http://www.federalreserve.gov/econresdata/scf/files/impute98.pdf`.

Satkartar K. Kinney, Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd. Towards unrestricted public use business microdata: The Synthetic Longitudinal Business Database. *International Statistical Review*, 79(3):362–384, December 2011. URL `http://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html`.

Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. *International Conference on Data Engineering (ICDE)*, 2008. URL `http://lehd.ces.census.gov/doc/help/ICDE08_conference_0768.pdf`.

T.E. Raghunathan, J.P. Reiter, and D.B. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19:1–16, 2003.

Donald B. Rubin. Discussion of statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468, 1993.