Joint NSF-Census-IRS Workshop

on

Synthetic Data and Confidentiality Protection

July 31, 2009

Suitland, MD @ U.S. Census Bureau

Final Print Program

as of September 24, 2009

Any subsequent changes will be reflected in the online version of this program, at
http://www.vrdc.cornell.edu/news/nsf-census-irs-workshop2009/

Contact:

virtualrdc@cornell.edu

Organizing committee (DC)

* Ron Jarmin (U.S. Census Bureau)

* Arnold Reznek (U.S. Census Bureau)

Program committee

* John Abowd (Cornell University)

* Karen Masken (IRS)

* Jerry Reiter (Duke University)

* Lars Vilhuber (Cornell University)

Funding for the conference and its preparation were provided by National Science Foundation (NSF) through Grant SES-0922494, the U.S. Census Bureau's Center for Economic Studies, the Internal Revenue Service (IRS), and the School of Industrial and Labor Relations at Cornell University.

Program overview

:: 8:00 Breakfast and registration

Breakfast is self-catered in the Census cafeteria.

:: 9:00 Opening remarks

..1.. **Tom Mesenbourg**, Deputy Director of the U.S. Census Bureau

..2.. **John Abowd** (Cornell University)

:: 9:30 Session 1: Imputation methods and synthesizers

Talks are 20 minutes each, with 20 minutes available for discussion and comments.

..1.. **Jerry Reiter** (Duke University): "Easily implemented, nonparametric synthesizers based on algorithmic methods in computer science."

Abstract

Many users of synthetic data, or any data altered to protect confidentiality, are understandably skeptical that analyses done on synthetic data will yield reasonable results. In this talk, I present recent research on methods to improve the analytic validity of synthetic data. Specifically, I talk about nonparametric methods of data synthesis that have the potential to capture complex distributional relationships.

..2.. Gary Benedetto (U.S. Census Bureau) and **Simon Woodcock** (Simon Fraser University), "Partially-synthetic linked employer-employee data"

Abstract

We describe recent and ongoing work to limit disclosure risk in linked employer-employee data via partial synthesis. We focus primarily on applications involving to the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program database. This includes applications of the Woodcock and Benedetto (forthcoming) synthesizer, which combines a simple imputation model (e.g., regression) with density-based transformations that preserve the distribution of the confidential data on specified subdomains. We also discuss the problem of synthesizing links between employers and employees.

..3.. **Robert Creecy** (U.S. Census Bureau): "The Feasibility of Creating a Fully Synthetic Decennial Census Microdata File"

Abstract

The data products the Census Bureau publishes from the Decennial Census are primarily in the form of pre-specified tables. These tables are cross classifications of the data collected on the short form questionnaire consists of relatively few variables: sex, age, race, ethnicity, relationship to householder and housing tenure. The most detailed set of data products known as Summary File 1 (SF1), consists of 286 tables. Some of the SF1 tables are available at the block level of geography, while others are available only at the tract level. Confidentiality of respondents data is protected by this restricted publication of tabular data and by data recoding and data swapping procedures. This paper explores the feasibility of creating a fully synthetic version of the Decennial Census microdata file that could be used for a wide variety of analytical purposes, including tabulations of custom geography, while still maintaining confidentiality.

..4.. **Michael Larsen** (George Washington University) and Jennifer Huckett (Battelle Institute), "Synthetic data methods using quantile regression and hot deck with rank swapping"

Abstract

Government agencies face demands to release accurate, timely data while upholding promises of privacy and confidentiality. We consider combining quantile regression with hot deck imputation to produce releasable, usable data. Conditional quantile regression models can capture complex relationships characteristic of demographic and economic data. Predicted values at several quantiles are computed to simulate values for several confidential variables. Values for other variables are imputed from the original data using hot deck imputation and rank swapping. These values are combined to form a data set for release that has low disclosure risk and high data utility.

..5.. John Abowd, **Fredrik Andersson**, Matthew Graham, Lars Vilhuber and Jeremy Wu, "Formal Privacy Guarantees and Analytical Validity of OnTheMap Public-use Data"

Fredrik Andersson (Office of the Comptroller of the Currency)

John Abowd, Lars Vilhuber (Cornell University and U.S. Census Bureau)

Matthew Graham, Jeremy Wu (U.S. Census Bureau)

Abstract

OnTheMap is an interactive mapping application developed by the U.S. Census Bureau's LEHD Program that displays where people live and work together with other employer and employee characteristics. The graphical user interface allows the construction of sophisticated queries to a geographically integrated employer-employee database whose core table is an origin-destination table resolved to the Census block level. OnTheMap is the first official data application ever released by a statistical agency that relies on synthetic data as its primary method of confidentiality protection. This paper documents the data development procedures, the disclosure avoidance protocol, and the analytical validity of the public-use data files incorporated into OnTheMap Version 3 (released September 2008). The statistical disclosure limitation procedure offers formal privacy guarantees based on the concept of differential privacy. This paper documents the level of privacy protection and the analytical validity of the released data.

:: 11:45-12:45 Lunch

Census cafeteria, self-catered. Vegetarian meals are available in the Census cafeteria. An average meal is between $5 and $10.

:: 12:45 Session 2: Synthetic data in public use micro-data products

Talks are 20 minutes each, with 20 minutes available for discussion and comments.

..1.. Martha Stinson, **Gary Benedetto**, and Melissa Bjelland, “Summary of Methods and Preliminary Assessment of the SIPP Synthetic Beta, version 5.0”

Martha Stinson, Gary Benedetto (U.S. Census Bureau)

Melissa Bjelland (Cornell University)

Abstract

This paper summarizes the methodology and quality assessment of the most recent version of the SIPP Synthetic Beta (SSB v5.0), a public use dataset that combines variables from the Census Bureau’s Survey of Income and Program Participation (SIPP), the Internal Revenue Service’s (IRS) individual lifetime earnings data, and the Social Security Administration’s (SSA) individual benefit data. The combination of long earnings histories and benefits information from administrative data sources with the detailed demographic data collected in the SIPP (we refer to this merged database as the Gold Standard) poses two major problems to the data provider. For one thing, there exists a large amount of missing data due to the combination of survey non-response with failure to link some survey respondents to administrative records. Moreover, when the link is successful, the administrative data potentially compromise the confidentiality protection of the existing SIPP public use files. To address the first problem, we use multiple imputation to fill in the missing values of variables by sampling from the posterior predictive distribution (PPD) conditional on all of the confidential data. The result of this process is multiple micro-datasets (completed data implicates) that have the same structure and same values for non-missing items as the Gold Standard but differ where the item was originally missing. For each completed data implicate, we sample again from the PPD for variables deemed to be sensitive, but this time we replace every record for each sensitive variable with multiple draws from the PPD. The result is multiple micro-datasets (synthetic data implicates) for each completed data implicate that share the same structure and same values for non-sensitive variables as the Gold Standard but differ for every record of every sensitive variable. This technique is called partial data synthesis since some variables contain actual responses while others have been replaced by values sampled from their PPD. The benefits of this data completion and confidentiality protection method are that the data users can run

their analyses on each synthetic implicate exactly as they would have if they had access to the single, confidential data set. After getting results for each synthetic implicate, relatively simple formulae exist to combine these results to get proper point estimates and measures of variance that take into account the uncertainty introduced by the modeling. Moreover, since every value of the vast majority of variables on the file have been replaced by random draws from a probability distribution, the partially synthetic data offer a very high level of confidentiality protection. In this paper, we also attempt to assess the analytic validity of the partially synthetic data and determine the disclosure risk of making such data available to the public.

..2.. **Saki Kinney** (NISS), “Synthetic Longitudinal Business Database”

Abstract

The U.S. Census Bureau’s Center for Economic Studies (CES) and partner institutions have developed a set of synthetic microdata use files from the Longitudinal Business Database (LBD). The LBD links selected variables for business establishments from the Census Bureau’s Business Register annual files. An initial ”synthetic LBD” file has been proposed for release that contains one synthetic implicate of establishments’ birth year, death year, employment and payroll, and an unaltered industry code (three-digit Standard Industrial Classification). We will discuss the generation and evaluation of the Synthetic LBD.

..3.. **Jörg Drechsler** (IAB), “New Data Dissemination Approaches in Old Europe Synthetic Datasets for a German Establishment Survey”

Abstract

In 2006 the German Institute for Employment Research (IAB) launched a research project to investigate possibilities for releasing synthetic datasets for its longitudinal establishment survey, the IAB Establishment Panel. After several successful pre-tests with fully and partially synthetic datasets based on a subset of the original data, the IAB decided to develop partially synthetic datasets for the latest wave of the establishment survey. The actual release of these data is planned for summer 2009.

In this talk we will describe the evolution of the project from the first small simulation studies over the imputation of the missing values and the discussions which variables to synthesize to the final synthesis. We will also discuss our disclosure risk evaluations and present some first results on the data utility of the generated datasets.

..4.. Sam Hawala and **Rolando Rodriguez** (U.S. Census Bureau), "Disclosure avoidance for group quarters in the American Community Survey: Details of the synthetic data method"

Abstract

For the third consecutive year, the U.S. Census Bureau will release public American Community Survey (ACS) microdata and tables containing information from respondents in group quarters. The Bureau utilizes synthetic data methods to protect these respondents from identity disclosure. We discuss the details of implementation of the synthetic data method. In particular, we focus on the modeling strategies used and the algorithms implemented to ensure consistency with ACS edit requirements.

..5.. **Trivellore Raghunathan** (University of Michigan), "Diagnostic Tools for Assessing Validity of Synthetic Data Inferences"

Abstract

Demands for micro data, especially if collected using public funds, is increasing. Disseminating such data also raises concerns about protecting confidentiality of respondents, a pledge given by the data collector to the data provider. An attractive proposal is to create synthetic data sets with some or all variables generated that emulate key statistical properties in the observed data set. The success of this approach in a public-use data context critically depends upon demonstrating whether the synthetic data inferences are "valid". In this talk, we will discuss the notion of validity from the frequentist and Bayesian perspectives, tools for assessing the validity and diagnostics to measure the similarity of synthetic and actual data inferences. These tools could be used to refine the models or reject synthetic data sets that are too dissimilar to the actual data sets.

:: 14:45-14:55 Coffee break

Coffee is available in the Census cafeteria. Beverages and cookies will be available in the back of the room.

:: 14:55 Session 3: Synthetic data and disclosure avoidance

Talks are 20 minutes each. A 5 minute discussion follows each individual paper.

..1.. **Arnold Reznek** (U.S. Census Bureau), "Disclosure avoidance issues at the Census Bureau"

Abstract

The U.S. Census Bureau has released a number of synthetic data sets over the last several years and is developing others. Many of these data sets have been described at this workshop. In developing and deciding whether to release synthetic data sets, the Census Bureau has had to decide where the disclosure risk lies, balance confidentiality and data utility, and develop disclosure analysis methods. This talk summarizes these issues, mostly in the context of the Longitudinal Business Database synthetic "beta version" that has recently been approved for release by the Census Bureau and the IRS.

..2.. **Stefan Bender** (IAB Germany), "Pulling the wool over users' eyes - Why is a German Research Data Center interested in synthetic data?"

Abstract

The last decade saw an ever-growing demand for micro data access in Germany. Because of existing rules for the anonymization of individual and household micro data, these data are available for research purposes since the 90s. Establishment/firm data on the other hand are mostly accessible only via guest stays and/or remote execution in the German research data centres (RDC). Beside data descriptions, only randomly generated test data are available for the researchers to develop their analysis code outside the RDC.

Some years ago, a joint research project initiated by the German Federal Statistical Office started to generate micro data on the firm level. Their perturbation techniques consisted mostly of micro-aggregation or adding noise to the data. Because of well know lacks of these techniques the RDC of the BA in the IAB started a project to generate synthetic datasets of one wave of the IAB Establishment Panel.

The motivation behind the project was twofold: (a). If the data quality of the generated synthetic datasets would be considered too low, the datasets could still be used to improve the quality of the test data for external researchers. (b). If the synthetic datasets would provide high data quality, the data could be released as scientific use files providing an additional way of data access for firm data. Still, the RDC will have to face the challenge of increasing acceptance in the research community for results stemming from synthetic data.

..3.. **Nick Greenia** (IRS), "Confidentiality Issues with Tax Data"

Abstract

Tax records provide a great deal of financial data at the firm, worker, and organizational level. Consequently, they provide an important resource for both the study of businesses,

as evidenced by their use in the Census Business Register, and for the study of workers, as evidenced by their use in the linked SIPP/SSA/IRS data file. This presentation summarizes the confidentiality challenges presented by the release of even ”phony” or synthetic data in the context of tax data.

The confidentiality concerns that have made it more difficult to produce traditional public use files have made synthetic data a potentially promising addition in the portfolio of options for the production of information to users. Other approaches include aggregate tables, remote access or execution, and access to the gold standard confidential microdata themselves at RDCs or via other approved sites and processes. In realizing this promise, a number of challenges with respect to confidentiality must be addressed, as tax data have a set of legal protections that differ from Census data. These include the lack of a statute of limitations for protecting taxpayer confidentiality and the fact that the Tax Code does not distinguish among tax items according to their sensitivity. In addition, the tax system is built on voluntary compliance, and the protection of confidentiality is a cornerstone of this compliance system. It is thus imperative that any data dissemination avoid unauthorized disclosures in perpetuity.

..4.. **Jennifer Madans** (NCHS). “Disclosure avoidance issues at NCHS”

Abstract

The National Center for Health Statistics (NCHS) uses a variety of mechanisms to maximize access to data while protecting confidentiality. The vast majority of our data is made available in public use data sets using standard disclosure control methods. Perturbed data sets have also been released on a limited scale and access to restricted use data to approved users is provided through Research Data Centers. NCHS has not yet developed and released synthetic public use data sets. There are several reasons for this including the cost of developing the methodology and the data sets and the skepticism of this approach expressed by some segments of our user community. Data production is a costly process particularly for an agency of the size of NCHS and resources for experimentation into alternative public release formats are limited especially if the new formats will not be accepted by considerable portion of our users. However, NCHS would be interested in further exploring adding synthetic files to our existing access methods if the creation and maintenance of the files was relatively cost efficient, if the technical assistance needed by users was not resource intensive and if demonstrating the integrity of the files so that they would be acceptable to our users was not overly burdensome. Since NCHS is not in a position to conduct independent research in this area, we would welcome collaborating with other agencies to meet these goals.

:: 16:15 Closing remarks

..1.. **Donald Rubin** (Harvard University)

:: 17:00 End of workshop

`$Id: workshop_program_details.tex 977 2009-08-14 02:59:22Z vilhu001 $`