LONGITUDINAL EMPLOYER - HOUSEHOLD DYNAMICS

TECHNICAL PAPER NO. TP-2003-10

Synthetic Data and Confidentiality Protection

Date	:	September 2003
Prepared by	:	John M. Abowd and Julia I. Lane
Contact	:	U.S. Census Bureau, LEHD Program FB 2138-3 4700 Silver Hill Rd. Suitland, MD 20233 USA

This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. [This document is released to inform interested parties of ongoing research and to encourage discussion of work in progress.] This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grant SES-9978093 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging Grant 5 R01 AG018854-02, and the Alfred P. Sloan Foundation. The views expressed herein are attributable only to the author(s) and do not represent the views of the U.S. Census Bureau, its program sponsors or data providers. Some or all of the data used in this paper are confidential data from the LEHD Program. The U.S. Census Bureau is preparing to support external researchers' use of these data; please contact U.S. Census Bureau, LEHD Program, Demographic Surveys Division, FOB 3, Room 2138, 4700 Silver Hill Rd., Suitland, MD 20233, USA.

Synthetic Data and Confidentiality Protection

John M. Abowd and Julia Lane[1]
Cornell University and The Urban Institute

September 2003

[1] This paper was presented at the Workshop on Microdata held on August 21-22, 2003 in Stockholm, Sweden. Much is drawn from joint work and discussions with John Haltiwanger and Martha Stinson. We thank Fredrik Andersson and Simon Woodcock for helpful comments. All errors are our own.

I. Introduction

The creation of demographic public use micro-data files fuelled a scientific and policy revolution. Restricted access to business micro-data in the early 1990's sparked a similar sea-change in the quality of analytical and policy work. In both cases a wide range of social scientists developed new empirical insights into social behavior, policy makers were able to base decisions on high quality statistical analysis, and entire generations of students grew up with the statistical and analytical tools that only access to micro-data can bestow. Even more compellingly, this massive social benefit was largely achieved simply by disclosure protecting already existing datasets – with little additional respondent or taxpayer burden.

It is now apparent, however, that new challenges threaten the ability of national statistical institutes (NSIs) to release high quality public use data files. Technological advances in computer capacity and matching technology combined with the explosion of online access to federal, state and local administrative records mean that NSIs must either severely degrade the quality of public use data files or refuse to release them in order to protect respondent confidentiality.

The response to this threat by the statistical community has been to develop new technical and non-technical approaches to preserve confidentiality but maintain the same quality of statistical analysis than was possible using old techniques. The NSI community is also responding to the issue – the Conference of European Statisticians recently established a working group to recommend approaches to micro-data access. This paper discusses the promise of a combination of several approaches: the development of inference-valid synthetic micro data, the establishment of a restricted

access site for remote analysis of the data and access to the "gold standard" analytical data set through a Research Data Center network. It also describes the actual approach currently being used to develop micro data files that can be used for the analysis of retirement decisions as well as the impact of reforms to state-specific needs-tested programs. It also describes the promise of the development of other datasets - particularly multiple public use files that can be created from the same underlying data that can be targeted at different audiences.

II. Background

A major issue with developing new data products of this type is that it is necessary to build a common body of understanding of the set of applications in which the product yields high quality analytical results. This requires a comparison of the results based on the synthetic data product with the results on the "gold standard" confidential source file. This poses serious constraints precisely because access to the "gold standard" file is, by definition, highly restricted. An obvious solution is to develop a two-part access protocol. The first part is to provide access to the full metadata repository of information, together with the synthetic data at a restricted access data center – a "virtual" Research Data Center. Researchers can use such a site to gain familiarity with the dataset structure, develop code, and estimate analytical models. Once the analysis is developed as far as feasible on this site, the models can be re-estimated at a Research Data Center on the "gold standard" file. The comparison of the two sets of estimates can be distributed as widely as possible – each analysis will provide an increment to the common body of knowledge as to what works and what doesn't. This section describes the three-layered approach

i) Synthetic data

A great deal of attention has recently been paid to the potential of using synthetic data as an alternative approach to releasing public use data files. (see Muraldhiar and Sarathy, and Abowd and Woodcock for reviews). In this paper we concentrate on the approach used by the latter: developing samples composed of draws from the posterior predictive distribution of the confidential data, given some conventionally disclosure-controlled data

The intuition behind this approach is straightforward. (We describe this for the linear case for ease of exposition.) The actual micro data, *Y,* - are replaced by a scientifically valid replacement. While one obvious candidate would be the predicted *Y*, conditional on everything we observe, this would not incorporate the inherent variability of the micro-data – particularly, outliers and unusual cases. The clear solution to this is to begin with the distribution of the predicted *Y*'s plus the residuals, the posterior predictive distribution Practically speaking, the approach generates a prediction and a residual for each *Y* variable 10 times - the "implicates." Statistical models using the data can then average the results from all ten implicates. Standard error formulas are available and simple to compute. The method has been in general use in statistics for more than 25 years for handling missing data and has recently been formalized for the synthetic data problem by Reiter (2002, 2003a and 2003b) and Ragunathan et al. (2003).

One example of this approach is to create a micro dataset of a universe of employees at a point in time in which the variables to be disclosure proofed include the person identifier (artificial), the sex, the county of residence and the date of birth

(year:month:day) In this example sex and county of residence are disclosure controlled using conventional methods.

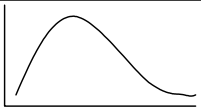
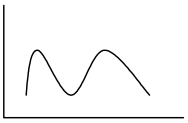
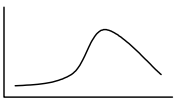
	County A	County B	…
Male	 Date of Birth Distribution	 Date of Birth Distribution	
Female	 Date of Birth Distribution	 Date of Birth Distribution	

Figure 1

The disclosure proofing approach is to form the table: sex by county of residence, and call each cell size $n(i,j)$. Then for each cell create $n(i,j)$ records with the appropriate values of sex and county of Residence. For each record, treat date of birth as an item missing value, given sex and county of residence. Then, estimate the posterior predictive distribution for date of birth, given sex and county of Residence and create a sample of $n(i,j)$ values from this posterior predictive distribution, one for each synthetic record. Repeat the procedure M times for each cell. The resulting synthetic data would exactly reproduce conditional distribution of birth dates in the sex by county of residence table by the properties of the posterior predictive distribution.[2]

What are the advantages of this approach? The synthetic data are inference valid. They contain the same statistical information as the micro data, as it is summarized in the

[2] In practice, the number of variables in each cell could be quite large.

posterior predictive distribution. Every record in the synthetic data set contains an imputed date of birth, hence, no record corresponds to a person in the population from which the original data were collected. In addition, the univariate distribution of date of birth within sex by county of residence cell can be reproduced to an arbitrary level of precision and the effect of disclosure protection on data quality can be measured. Finally, the multiple synthetic data implicates are not identical so the analyst can use the between implicate variation to measure the extent to which confidentiality protection made the inferences less precise.

In practical terms, an important additional value of such inference-valid synthetic data is that multiple public use files can be created from the same underlying data - targeted at different audiences. For example, some users of business data (such as transportation agencies) are particularly interested in geographic detail, while others are interested in industry detail (such as industry analysts). Providing both levels of detail on the same data set immediately re-identifies important businesses. However, inference-valid synthetic data could be used to produce two separate data sets that can not be re-linked for such re-identification. For example, a demographic dataset such as the Survey of Income and Program Participation (discussed below) has at least two important user constituencies. One constituency is interested in modeling the participation in welfare programs that are state-specific, with state specific qualification criteria – in which case geography is critical. Another constituency is interested in modeling retirement decisions – in which case date of birth is critical. It is well known that including both of these items in a file is a serious disclosure risk (Sweeney) – so the ability to release two separate files, each targeted to their distinct audiences, is extremely attractive. More

complex relationships can also be modeled (see for example, Woodcock 2003, or Abowd's presentation on http://www.urban.org/nsfpresentations/index.html).

ii) Restricted access Data Center – the "Virtual" RDC

If multiple users can access the same dataset, and build on an existing database infrastructure, there are numerous advantages. Results can be replicated or expanded – which is a critical condition for scientific validity. Researchers can use existing datasets to condition their analyses in different ways, with different foci, which develops a broader understanding of the generalizability of results. In addition, the common use of similar dataset builds a common body of knowledge, as has been the case with public use files such as the Public Use MicroSample for the Decennial Census and the Current Population Survey.[3]

One increasingly important approach for facilitating researcher access is to maintain the data in a secure, restricted-access environment, but widely distribute synthetic data through a restricted-access remote site. Because the simulated data can be used at less secure sites than the statistical agency itself, researchers can develop an understanding of the structure of the datasets and use simulated data to develop code and estimate basic relationships before sending the code to the an official secure site to estimate the underlying relationships from the actual confidential data.

[3] Indeed, a very powerful case for this approach has been made by Soete and ter Weel, 2003.

Figure 2

The National Institute on Aging and the National Science Foundation have already invested in the development and provision of simulated data to selected users at the Cornell Restricted Access Data Center (CRADC). Researcher access to these data is supervised by the CRADC. The purpose of the CRADC is to provide external researchers with access to the simulated data as well as access to the basic data research tools. Computing resources for facilitating research on synthetic data are provided on the CRADC nodes. These resources include SAS, Stata, Matlab, Fortran V6, GLIM, Genstat, Gauss, eViews, ASREML, data conversion software StatTransfer, as well as

tools such as TextPad, Microsoft Office, Scientific Workplace, and Adobe Acrobat. An example of what the site looks like from a researcher's office is provided in Figure 2.

iii) Research Data Centers

An important component of developing a new confidentiality protection system is to develop a research data center (RDC) network in which the quality of the new data product can be tested. The more sites that are available and accessible, the greater the ability of the scientific community to build the core common body of knowledge necessary for the acceptance and use of the new data product.

The existence of such a network is, of course, critical whether or not synthetic data approaches are adopted. An important consequence of the increasing threat of re-identification is that more and more noise is being added to public use datasets – with analytical consequences that would be unknown without access to the underlying confidential data. Since noise addition may seriously bias estimated coefficients, researchers might, for example, incorrectly conclude that earnings differentials by race and sex had vanished over time – rather than realizing that more noise had been added over time!

The basic structure of the RDC network in the United States is well known, and described in both Dunne (2001) and on the Center for Economic Studies website (www.ces.census.gov). Briefly, RDCs enable external researchers to access micro-data under strict security protocols. All researchers must become Special Sworn Status employees of the Census Bureau (which involves fingerprinting, an FBI check, and an oath to protect the confidentiality of respondents – which, if broken, subjects the researcher to the penalty of a $250,000 fine and/or 5 years in jail). The researcher must

document which files will be accessed, which variables used, and for which period of time. The researcher must also demonstrate that the predominant purpose of the research is to improve Census Bureau censuses, surveys and inter-censal population estimates, and provide a post-project certification that this has been achieved.

However, the RDC network still imposes substantial access costs. To date, there are seven physical RDC sites in the U.S. – so the vast majority of researchers are not physically close enough to be able to easily access micro data. Even researchers who are reasonably close to an RDC must commit substantial amounts of time away from their own office and research materials. Clearly, the combination of a "Virtual" RDC with a real RDC network reduces many of these costs.

III. Applications

One initial application of this approach has been to develop a public use file based on a match between the Survey of Income and Program Participation (SIPP) and detailed administrative earnings histories. A major goal of this new public use data set is to increase access by researchers studying retirement and disability issues to this new micro-data source – given that accurate information on individual earnings histories, combined with demographic information, are critical inputs to modeling the retirement decision. The production of this data set is considered one of Census's major Title 13 uses of the administrative data.

The development of this micro-data file involved the establishment of a committee with the policy, disclosure, survey improvement and production, and research staff from three agencies that are data custodians - the Census Bureau, the Social Security Administration, and the Internal Revenue Services – and two agencies that are data users

- the Congressional Budget Office and the Social Security Administration. The committee had to make a number of decisions - summarized in the following steps:

1. The focus of the target audience (which was retirement and disability researchers).
2. Identify key variables necessary to model outcomes of interest.
3. Investigate feasibility of traditional disclosure-proofing methods (primarily coarsening and item suppression).
4. Create an internal-use "Gold-Standard" dataset with two functions: determine that the appropriate research-required set of variables has been included and act as a benchmark for the eventual public-use version.
5. Create a synthetic data set from the "Gold-Standard" file.
6. Test the synthetic dataset to ascertain that it prevents re-identification of SIPP individuals and preserves statistical relationships among variables through a restricted access site.
7. Disseminate the new products.

The first four steps have been taken, and the first version of the synthetic file is expected to be created by October – the researchers in the group will then begin to evaluate the quality of the synthetic file.

Another application of this approach has been to create a synthetic dataset for French longitudinally linked employer data.[4] The quality of this approach is evident in Figure 3 – using French data, Abowd and Woodcock show that there is almost no difference between results estimated using synthetic (masked) data and real data.

[4] Abowd and Woodcock (2001); Abowd and Woodcock, 2003. They call the technique discussed herein "masked data." Note that no public use micro data products were released as a part of this research.

Figure 1

Summary

The continued distribution of public use data-files is clearly threatened by the increased re-identification risk associated with both technological advances in linking software and the widespread availability of administrative records. It is clear that new approaches to developing public use data files must be investigated. This paper suggests the adoption of a three-tiered approach that combines both technical and non-technical approaches. The technical approach – the creation of synthetic datasets – could, in principle, permit the creation of multiple public use datasets from a single underlying confidential file that could be customized for multiple different constituencies. The non-technical approach is to combine the use of an already well accepted RDC network with that of a "Virtual" RDC to both reduce the access costs

and develop a common body of knowledge about the quality of the results generated from the analysis of synthetic data files relative to that from confidential micro-data. While the initial results have been quite promising, more extensive research is ongoing.

References

Abowd, John M. and Simon Woodcock, "Disclosure Limitation in Longitudinal Linked Data," in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001), 215-277.

Abowd, John M and Simon Woodcock, "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data", mimeo, Cornell University, 2003

Dunne, Timothy, "Issues in the establishment and management of secure research sites" in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001).

Muralidhar, Krishnamurthy and Rathindra Sarathy, "Application of the Two-step Data Shuffle to the 1993 AHS Data: A Report on the Feasibility of Applying Data Shuffling for Microdata Release," research report prepared for the Census Bureau (June 2002).

Raghunathan, T.E., J.P. Reiter, and D.B. Rubin, "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics*, 19 (2003): 1-19.

Reiter, J. (2002) "Satisfying Disclosure Restrictions with Synthetic Data Sets." *Journal of Official Statistics,* 18, pp. 531-544.

Reiter, J. (2003a) "Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study." *Journal of the Royal Statistical Society, Series A.*

Reiter, J. (2003b) "Inference for Partially Synthetic, Public Use Microdata Sets." *Survey Methodology.*

Soete, Luc and Bas ter Weel, "ICT and Access to Research Data: An Economic Review", Maastricty Economic Research Institute on Innovation and Technology, mimeo, June 2003

Yancey, William E., William E. Winkler and Robert H. Creecy, "Disclosure Risk Assessment in Perturbative Microdata Protection," Research Report Series Statistics 2002-01, available online at http://www.census.gov/srd/papers/pdf/rrs2002-01.pdf, cited June 11, 2003.