

## **Synthetic Data for Administrative Record Applications at LEHD**

Jeremy Wu, Assistant Division Chief LEHD, Data Integration Division  
John M. Abowd, Distinguished Senior Research Fellow

*This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on methodological or technical issues are those of the authors and not necessarily those of the U.S. Census Bureau.*

## 1. Abstract

The Longitudinal Employer-Household Dynamics Program at the U.S. Census Bureau has developed several synthetic data products including OnTheMap and a partially synthetic version of the Survey of Income and Program Participation linked to Social Security Administration and Internal Revenue Service data. In this paper we propose quality assurance standards that statistical agencies might use for developing and supporting such products. Documentation of inference validity and its relation to the synthetic data method of confidentiality protection is an important feature of our proposal.

## 2. Introduction

Responding to a report (1993) by the Panel on the Confidentiality and Data Access, Rubin (1993) proposed a new research initiative to release “only synthetic micro-data sets” for general public use while retaining the analytical validity of the original data. In the same volume, Little (1993) proposed the approach that has become known as “partially synthetic” micro-data, where only the “sensitive values” are replaced by synthesized data.

Traditional methods (coarsening, top coding, swapping, and cell suppression) deny users access to certain data, which have been masked to protect the confidentiality of the original provider. Businesses and individuals, who supplied the original data either as direct respondents or as the source of administrative records, are entitled to this confidentiality protection according to both legal and ethical standards. The foundation of the American official statistical system is predicated on the trust that citizens place on the stewards of the data to uphold these standards.

While there are a variety of approaches for implementing Rubin’s original proposal, the term “fully synthetic data” is now generally understood within the statistics community to mean that the original confidential data have been used to simulate the values of all variables for the entire population (respondents and non-respondents, whether originally sampled or not). The released data are samples from this synthetic population.

The term “partially synthetic data” is now generally understood to mean that the released data contains a mixture of actual responses and simulated responses. All synthetic data applications combine sophisticated multivariate statistical modeling and computationally intensive simulations based on this modeling. Abowd and

Woodcock (2001) showed that the partially synthetic approach could be reliably used to produce releasable micro-data files based on confidential data integrated from employer, household, and job-level records.

In February 2006, the Census Bureau released a pilot web-based application known as “OnTheMap”<sup>1</sup>. This online mapping and reporting tool allows the user to select an arbitrary geographical region (resolved to the Census block) to show where people live and work. Companion reports on the workers’ age, earnings, and industry distributions, as well as the number of employers, and Quarterly Workforce Indicators (QWI),<sup>2</sup> are resolved to the block group level.

OnTheMap is the first synthetic data product publicly released by the Census Bureau<sup>3</sup>. Version 2.0 of OnTheMap, which will eventually include 42 states, began its release in April 2007. The QWI (released initially in 2003) were the Census Bureau’s first public-use product protected by noise infusion, a somewhat simpler version of the same principles used in synthetic data.

Although OnTheMap was released first, the LEHD Program has been engaged in synthetic data research since 2001. The most ambitious effort has been a project to create a partially synthetic version of the 1990-1996 panels of the Survey of Income and Program Participation (SIPP) that have been linked to Social Security Administration (SSA) benefits data and Internal Revenue Service (IRS) data from the complete longitudinal earnings history of the respondent (annually back to 1950). The design of this file is the product of collaboration between the Census Bureau; SSA/Office of Research, Evaluation, and Statistics; IRS/Statistics of Income and Research; and the Congressional Budget Office. Because the existing versions of the SIPP public use files already include thousands of variables on each respondent household, the version under development for this project synthesizes every variable except for a small

---

<sup>1</sup> Available at <http://lehd.did.census.gov> on July 1, 2007.

<sup>2</sup> These are flagship products for the LEHD Program providing unprecedented information on employment, churning, and earnings that describe the local employment dynamics. For additional details, visit [http://lehd.did.census.gov/led/library/tech\\_user\\_guides.html](http://lehd.did.census.gov/led/library/tech_user_guides.html), available on July 1, 2007.

<sup>3</sup> In the 1990 Census of Population and Housing, the summary tape files used a technique called “blank and impute,” which is now generally recognized as a predecessor to synthetic data (Statistical Policy Working Paper 22, 2005, p. 21).

group of demographic and benefit variables (four variables in all). The remaining 500+ variables on the file are synthetic.

A test version of these data, called the SIPP Synthetic Beta, was approved by the Census Bureau, SSA, and IRS for experimental use by the general research community in May 2007.

Research access to the synthetic data is provided on a Cornell University server known as the Virtual RDC. These data are not confidential but have not been released as a supported public-use product, pending the results of analyses to be conducted by SSA, the Census Bureau, and the general research community regarding the analytical validity of the SIPP Synthetic Beta. We discuss this process in more detail below.

### 3. Technical Background

In developing its public-use products based on synthetic data, the LEHD Program employs a pragmatic approach which Little (2006) has termed “calibrated Bayes.”

Our public-use synthetic data products are based on estimated posterior predictive distributions. The prior contribution is based on “parameters” that are a combination of empirical estimates and constants that are controlled for confidentiality protection. The likelihood contribution is based on the actual confidential data, as well as estimates using a variety of techniques including exact multivariate distributions and approximations. The posterior predictive distributions are used to make multiple draws, known as implicates or replicates, by high-power computers to produce the synthetic data.

LEHD integrates existing data from administrative records, censuses, and surveys to build its infrastructure file system. Multiple imputations are used to replace missing values, and noise infusion is applied strategically to provide confidentiality protection for some public-use statistics based on establishment-level summaries such as QWI. The LEHD program uses only a small amount of suppression and is migrating away from this technique by implementing synthetic data replacements for the suppressed values.

The goal of the LEHD Program is to build a longitudinal national frame of jobs<sup>4</sup> with an associated data infrastructure to support rapid production of timely, rich

---

<sup>4</sup> Currently, the frame is based on state unemployment wage records and the Quarterly Census of Employment and Wage (QCEW) records that cover about 98 percent all private, non-farm employment.

results. The present infrastructure includes the history and characteristics for each worker and each employer covered under the state unemployment insurance wage record system for 45 participating states.<sup>5</sup>

Since a worker may have multiple jobs, synthetic observations in On The Map are generated for both the count of jobs and the count of workers for each unique census block of residence (also called origin), conditional on each census block of workplace (also called destination) and defined combination of age,<sup>6</sup> earnings,<sup>7</sup> and industry<sup>8</sup> association (collectively called characteristics). Minnesota is the only state in the U.S. that codes the worker’s establishment on its unemployment insurance wage records. Consequently, the Minnesota wage record data can be directly integrated with the establishment level data using conventional, identifier-based record linkage methods. For all other states, LEHD uses a statistical model, estimated using information from Minnesota and information from the specific state, to multiply impute the establishment (and therefore destination) for each worker employed in a multi-unit firm.

The synthetic data for OnTheMap are generated for each origin based on a Multinomial model for the origins whose probabilities are drawn from the posterior Dirichlet (multivariate beta) distribution conditional on unique destination and employee-workplace characteristics. The conjugate prior distribution for the probabilities is also a conditional Dirichlet distribution which must have sufficient empirical support and whose parameters are specified for confidentiality protection. The likelihood function is the multinomial distribution conditional on destination and characteristics. Noise is already infused into the destination count of workers from the QWI protection system. Thus, only the origin data are synthesized, and OnTheMap may be described as a

---

<sup>5</sup> As of May 2007, 45 states have signed the Memorandum of Understanding to join the federal-state Local Employment Dynamics partnership, which establishes the source of the state-supplied unemployment wage records. As of May 2007, 41 of those states are supplying production-ready data and four are supplying provisional data.

<sup>6</sup> There are three categories for the age of the worker – up to 30, 31-54, and 55 and above.

<sup>7</sup> There are three categories for the average monthly earnings of the highest paying job of a worker in a quarter – up to \$1,200, \$1,201 – \$3,400, and \$3,401 and above.

<sup>8</sup> There are 20 categories based on the 2-digit sector level under the North American Industry Classification System (NAICS), available at <http://www.census.gov/epcd/www/naics.html> on July 1, 2007.

partially synthetic data product, using the current standard nomenclature.

Ten implicates were drawn from the posterior predictive distribution for each origin-destination pair, and the first implicate was used for the graphical implementation of OnTheMap on the Census Bureau's web site.

The SIPP Synthetic Beta and the proposed administrative records enhancements to future SIPP released are based on a more elaborate synthetic data system that works variable by variable to create the synthetic values. The system is based on sequential regression multivariate imputation (SRMI, Raghunathan, Reiter and Rubin, 2003) as applied to the case of partially synthetic data with missing values (Reiter 2004). The estimation system stratifies the analysis based on non-synthesized variables and other variables specified by the analyst building the synthetic file. The imputation/synthesizing software automatically builds a posterior predictive distribution for each variable, conditional on all other values in the data. First, the missing data are completed using SRMI. Then the synthetic implicates are built using the same software.

The value of synthetic data lies in its ability to provide public use micro-data that preserve analytical validity and provide confidentiality protection.

#### 4. Evaluation Criteria

As a state-of-the-art innovative approach, integrated data and synthetic data do not have quality standards that are comparable to census and survey data. This situation is perhaps similar to the period when the concept of random sampling was introduced at the end of the nineteenth century (Wu 1995), but contemporary statistical theories and methods to support the applications of random sampling did not get fully established in the international statistical community until more than 30 years later.

Development of the LEHD products follows the general guidelines of the Data Quality Act (2001) and the Census Bureau (2006) on utility, objectivity, and integrity. In particular, LEHD products have been developed to fit the needs of its users.

The creation of a voluntary federal-state partnership known as Local Employment Dynamics (LED) was based on the recognition and principle that the state partners supply their wage records on workers and firms and the Census Bureau builds an integrated data infrastructure to create unprecedented new statistics about local employment dynamics that include both economic and demographic information. Since the beginning of 5 states in 2000, LED has now grown to 45 state partners (as of

May 2007). The state of New York enacted legislation<sup>9</sup> to allow for data sharing with the Census Bureau in order to join LED, as did Michigan and Rhode Island. The Employment and Training Administration (ETA) of the U.S. Department of Labor provided major funding support for OnTheMap and its successor, OnTheMap Version 2, as well as the expansion of the infrastructure of LED to a national program. The Brookings Institution identifies LEHD as a top priority for its Federal Data Agenda.<sup>10</sup>

LEHD was designed to use only existing data that have already been collected; therefore, it does not impose additional burden on the original respondent. Processing these data is a sophisticated, multi-threaded operation using about 6,500 hours of computing time in each 3-month cycle to (a) integrate all input data with other Census Bureau data sources and produce the flagship product known as Quarterly Workforce Indicators (QWI), and (b) update the LEHD integrated data infrastructure, which currently has more than 6 billion records and growing.

With LEHD as a highly automated operation, the cost of data processing is a fraction of a penny per record although, in over-simplified terms, the annual LEHD volume is more than 20 times that of the decennial census<sup>11</sup> over a 10-year period.

The inaugural version of OnTheMap was publicly released in 2006 after 18 months of intensive design, development, test, and evaluation guided by teams and beta testers inside and outside the Census Bureau. It was reviewed and approved by the Census Bureau Disclosure Review Board and verified for Section 508 compliance prior to its release.

The measures of analytic validity for OnTheMap are based on comparing the synthetic commuting distances with the distances computed from the underlying confidential data. Abowd, Andersson, and Roemer (2006) provide detailed evidence of the analytic validity of OnTheMap data.

---

<sup>9</sup> New York State Assembly bill was A11619 and New York State Senate bill was S08072, available at <http://assembly.state.ny.us/leg/> on September 10, 2006.

<sup>10</sup> Available at <http://www.brookings.edu/metro/federaldata.htm> on July 1, 2007.

<sup>11</sup> There will be approximately 300 million individual records for the 2010 census, which is conducted every 10 years. There are about 150 million workers in the nation whose records are processed four times a year by LEHD.

Confidentiality protection is assessed using a measure that is comparable to the “swap rate” in systems that are based on data swapping. We compare the relative difference between the synthetic count and the actual confidential count of jobs or workers for all origin-destination pairs and their aggregates.

Abowd et al (2006) introduced this “Reclassification Index,” which varies from 0 to 1. If the counts in synthetic and confidential data were identical in all cases, the reclassification index would equal to 0. The index may also be interpreted as the proportion of workers that need to be reallocated across origins in the synthetic data in order to replicate the actual data. The analysis of the reclassification index shows that for small geographic areas, there is considerable reclassification required to reconstruct the confidential data (often more than 50% of the cases must be reallocated) whereas in large geographic area only a trivial percentage need reclassification (usually less than 2%).

Since its release, OnTheMap has stimulated strong interest and discussion not only about its use for workforce and economic development, but also for transportation planning, emergency preparedness and response, and military base realignment. Originally funded for 12 pilot states by ETA, there were 17 states participating in OnTheMap Version 1. ETA supported the Census Bureau to expand the application to 42 states in 2007. OnTheMap Version 2.0 (3 states) was released in April 15, 2007. Version 2.1 (13 additional states) was released on June 1, 2007.

OnTheMap public use data are currently made available through the LED state partners, the Census Bureau, and the Cornell University. In particular, all 10 implicates of the OnTheMap data for Version 1.0 for Oregon and Texas were available from the Cornell University VirtualRDC for use and evaluation<sup>12</sup> by registered users. Three (3) implicates for all Version 2 states are being released under the same terms on the VirtualRDC. Users are encouraged to submit emails<sup>13</sup> to comment on the program and report bugs<sup>14</sup> in the application, as well as share their findings and results through listservs.

---

<sup>12</sup> Available at <http://vrdc.ciser.cornell.edu/onthemap/doc/index.html> on September 10, 2006.

<sup>13</sup> The standard email address is [did.local.employment.dynamics@census.gov](mailto:did.local.employment.dynamics@census.gov), and it is monitored directly by the program manager.

<sup>14</sup> Known and unresolved bugs about OnTheMap are posted at <http://lehd.did.census.gov/led/datatools/onthemap.html>, available on July 1, 2007.

OnTheMap is a relative simple application of the synthetic data approach. For the SIPP Synthetic Beta and future SIPP applications both the analytical validity studies and the confidentiality protection analysis are much more complicated.

Analytical validity is assessed by comparing the complete univariate distribution of each synthetic variable to its confidential counterpart. Discrete variables match exactly for the overall sample and the sub-samples represented by the unsynthesized variables. Continuous variables match every percentile from 1 to 99 exactly. Multivariate analytical validity is assessed using multi-way contingency tables with up to four interactions, covariance matrices, regression analyses, micro-simulation, and propensity score methods.

There is no prevailing standard for analytical validity of such products.

The Census Bureau and SSA invited researchers to use the SIPP synthetic beta files on the Cornell VirtualRDC. Application documents and codebooks have been established at [www.sipp.census.gov/sipp](http://www.sipp.census.gov/sipp). Approved analyses will also be run on the confidential version of the integrated SIPP/SSA/IRS data.

The Census Bureau and SSA recognize that developing standards for the analytical validity of synthetic data is essential to their success. This is among the reasons for releasing the product in beta format for extensive testing by users who did not participate in the product’s development.

Confidentiality protection in these more complicated partially synthetic applications is based on attempting to re-identify the confidential source record for each synthetic record. Probabilistic record linking and distance matching were used for the re-identification studies. The goal was to have very low overall correct re-identification rates and to have the “best match” case be a false re-identification as often as it is a true re-identification. Both of these standards have been met.

## 5. Ongoing Development

Given its brief history, the development and practical use of synthetic data is just beginning. The supporting theories and the measures for evaluating its usefulness and value will undoubtedly grow and evolve.

Through a 3-year grant to Cornell University as a coordinating institution with the Census Bureau as the prime subcontractor, the National Science Foundation (SES #0427889) encourages innovative, high-payoff

research and education to develop public-use synthetic data under the Census Bureau Research Data Center system, and to help facilitate collaboration to help design and test these products. The LEHD Program benefits from and contributes to this activity.

The broad and growing interest surrounding integrated data and synthetic data offer several opportunities for the Census Bureau, covering workforce development, transportation planning, economic dislocation and development, emergency preparedness and response, in addition to aging research, small business policy analysis, and program evaluation.

LEHD is also identified in the Census Bureau Strategic Plan<sup>15</sup> as a performance measurement tool on improvements to data coding, processing, and analysis. The transportation community, under the leadership of the American Association of State Highway and Transportation Officials and the U.S. Department of Transportation (DOT), have also planned to compare LEHD data with the American Community Survey, the decennial census, and commercial data sources as part of its development of the next Census Transportation Planning Package. The ability to make longitudinal checks and edits of time series data is one of the inherent benefits of LEHD.

A particularly noteworthy outcome since the release of OnTheMap is the developing collaboration of the state labor market information (LMI) offices, the state DOT, and the local metropolitan planning organizations (MPO). As the state DOT and MPO have a growing desire for the origin-destination data, they also have better local knowledge about place of work. These agencies for at least 4 states have begun exchange of ideas to form a feed-forward and feedback loop to enhance data quality. This growing trend can benefit LEHD substantially by reducing the need for imputation and improving the overall data quality.

### Acknowledgements

The authors would like to thank Don Rubin and Al Tupek for their valuable comments and suggestions. In addition, the many current and past LEHD staff members have been particularly dedicated and made extraordinary contributions to the program.

---

<sup>15</sup> Objective 1.4: Produce new information using existing data sources by developing cutting-edge techniques and promoting knowledge sharing, available at <http://www.census.gov/main/www/strategicplan/strategicplan.html#1-4> on September 10, 2006.

### References

- Abowd, J.M., Andersson, F, and Roemer, M.I. (2006). "Disclosure Avoidance and Analytical Validity in "On The Map" –A Synthetic Data Application from the U.S. Census Bureau's LEHD Program" (October 18 draft).
- Census Bureau (2006). "Census Bureau Section 515 Information Quality Guidelines," Available at <http://www.census.gov/quality/>, cited on September 10, 2006.
- Duncan, G.T., de Wolf, V.A., Jabine, T.B., and Straf, M.L. (1993). "Report of the Panel on Confidentiality and Data Access," *Journal of Official Statistics*, 9, 271-274.
- Federal Committee on Statistical Methodology, "Report on Statistical Disclosure Limitation Methodology," Working paper 22, 2005 ([http://www.fcsm.gov/working-papers/SPWP22\\_rev.pdf](http://www.fcsm.gov/working-papers/SPWP22_rev.pdf)) Cited on June 8, 2007.
- Little, R (1993). "Statistical Analysis of Masked Data," *Journal of Official Statistics*, 9, 407-426.
- Little, R (2006). "Calibrated Bayes: A Bayes/Frequentist Roadmap," *The American Statistician*, 60, 213-223.
- Office of Management and Budget (2001). "Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies," [http://www.whitehouse.gov/omb/fedreg/final\\_information\\_quality\\_guidelines.html](http://www.whitehouse.gov/omb/fedreg/final_information_quality_guidelines.html), cited on September 10, 2006.
- Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics*, 19, 1-16.
- Reiter, J.P. (2004). "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation," *Survey Methodology*, 30, 235-242.
- Rubin, D.B. (1993). "Discussion: Statistical Disclosure Limitation," *Journal of Official Statistics*, 9, 461-468.
- Wu, J. (1995). "One Hundred Years of Sampling," special invited paper in *Sampling Theory and Practice*. State Statistical Bureau of China, ISBN 7-5037-1670-3, China Statistical Publisher, Beijing, China.