www.apdu.org

*Association of Public Data Users*

*NEWSLETTER, March/April 2010, Vol. 33, No. 2*

# *OnTheMap*: Block-level Job Estimates Based on Longitudinally Integrated Employer-Employee Micro-data

John M. Abowd(1)
Cornell University

## Overview

This article discusses the statistical underpinnings of *OnTheMap*, the Census Bureau's graphical statistical system that displays workplace and residential distributions with geographic resolution to the census block level(2). Although the data superficially resemble a small-area estimation system, there are important differences that I attempt to elucidate in this short note.

The most important difference is that *OnTheMap* and its sister product, the Quarterly Workforce Indicators, are built from a comprehensive longitudinally integrated job frame. Hence, the micro-data and the associated tabulations more closely resemble censuses than survey-based estimates. Statistical universes are usually defined in terms of either households or business establishments. Longitudinally integrated job data define their universe in terms of the pair—a business and an individual who is a statutory employee. The most important difference between a job frame and either a household or establishment frame is that the frame itself is the result of behavioral links between the individuals and businesses that form and dissolve as a consequence of economic activity. Hence, even a perfectly implemented, complete job frame—one that contains information on every active employer-employee pair in the economy—is the realization of a process in which many of the unobserved pairs in a given period could have occurred but didn't. This "random graph" feature of the longitudinal job frame provides the conceptual impetus for using probability models to describe the frame and its associated tabulations even though there is no sampling of jobs to form any of the summary estimates.

In preparing the public-use files for *OnTheMap* and the Quarterly Workforce Indicators, the Longitudinal Employer-Household Dynamics (LEHD) program at the Census Bureau makes use of both frequentist (sampling-theoretic) and Bayesian methods, following the "pragmatic" approach advocated by Rubin, Little, and others(3).

**Conventional Small-area Estimation Compared to Geographically Detailed Tabular Summaries**

In conventional small-area estimation there are limited data on the concepts of interest measured at the geographical level of interest, and these are usually survey-based responses(4). For example, small-area poverty estimates from the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program are based on the American Community Survey (since 2005) supplemented with Decennial Census data aggregated to the geographic level of interest, Internal Revenue Service (IRS) data aggregated to the state level, and Supplemental Nutrition Assistance Program (SNAP) data aggregated to the county level. The main poverty measures come from the ACS, which is a sample survey. Because the primary variables of interest are directly measured in sample surveys, conventional small-area estimation uses models to relate measures that can be observed with precision at the appropriate geographic level to the survey measures of interest. Then, these models are used to estimate the small area mean and the reliability of that mean. Continuing with the SAIPE example, the ACS measure of the appropriate income or poverty variable is modeled using the data from the Census of Population, IRS, and SNAP as predictors(5).

In contrast, administrative record-based systems like *OnTheMap* use the universe of outcomes of interest measured directly at the geographical level of interest. Consequently, such systems more closely resemble census tabular summaries than small-area estimates. For example, in *OnTheMap* the main employment outcome is defined by counting all individuals who appear in the job frame at the same employer in the first and second quarters of the year. This variable is then refined to keep one such job when an individual has several with that "primary job" selected as the one that produced the highest earnings in the second quarter of the relevant year. The job is then geo-coded to a residence based on the best available residential address for the data year and to a work place based on the best available workplace address for the second quarter of the calendar year. There is no sampling or model-based estimation of the measure of interest (employment and potential commute patterns). The primary reliability issues revolve around coverage, edits, imputation, and confidentiality protections applied to the underlying database to generate the tabular summaries. There is no classical sampling error, and any random estimation error stems from the properties of the edit and imputation models. Modern confidentiality protections also induce randomness in the published data, by design.

**Coverage**

The block-level data published in *OnTheMap* are aggregated from the LEHD Infrastructure File System (Abowd et al. 2009). These data constitute

a comprehensive job frame based primarily on Unemployment Insurance wage records, which record the amount paid to a statutory employee by a UI-covered employer during a calendar quarter. The UI wage records and the Quarterly Census of Employment and Wages (QCEW) establishment data are provided to the Census Bureau via the Local Employment Dynamics (LED) federal/state partnership. The UI wage records contain Social Security Numbers (SSNs), which allow administrative record linking to individual/household frames, and state Unemployment Insurance account numbers, which allow administrative record linking to business frames, in particular the QCEW and Census Business Register.

Private employer coverage for non-agricultural jobs is nearly universal. At the business establishment level, the relevant universe is the QCEW, which the Bureau of Labor Statistics estimates to cover over 96% of all wage and salary civilian jobs(6). At the job level, the main coverage discrepancy with the QCEW occurs because federal and USPS employees receive unemployment insurance through a different administrative agency. Federal and USPS employers do not file UI wage records, and are, therefore, not covered by the current LEHD job frame. Efforts to integrate federal employees using their Office of Personnel Management records are underway as part of the congressionally funded FY2010 Local Employment Dynamics initiative.

**Input Data Sources and Vintages**

In addition to UI wage records and QCEW establishment data, which are updated quarterly as part of the LED federal/state partnership between the Census Bureau and state Labor Market Information offices, many other data sources are linked to the basic infrastructure file system. Different vintages of the Master Address File, the Census Bureau's residential address frame; the Personal Characteristics File, a file derived from the Social Security Administration's Numident registry of SSNs; TIGER line and shapefiles, the official descriptions of Census geography; and the Composite Person Record, an output of the Statistical Administrative Records System. Each of these input files is, to a greater or lesser extent, partially dependent upon information from the others. Consequently, maintaining vintage consistency and accuracy is a major challenge to the data integration effort.

The sources listed in the previous paragraph are the major input data integration sources, but many other sources are currently used, and are planned for use in the near future. These include Census 2000, the American Community Survey (from 2005), annual March supplements to the Current Population Survey, all panels of the Survey of Income and Program Participation, the American Housing Survey, the Census Employer and Non-Employer Business Registers, Economic Censuses, Annual Economic

Surveys, additional Statistical Administrative Records System (StARS) data, and administrative data from other agencies. Much of this integration effort is already underway at the Census Bureau.

**Edits**

In principle, every UI wage record should link to valid individual data based on the SSN(7) and to a UI-covered business via the UI account number(8). Detailed analyses of the reliability of these identifier systems have been published elsewhere(9). The primary edits to the identifiers consist of recoding some SSNs that have been determined to be erroneous and recoding some SEINs to enhance the longitudinal integrity of the identifier.

In fact, some of the primary edits occur when the employment counts from the UI wage records are compared to the employment counts for each UI account. This comparison, which is performed each quarter using the complete historical job frame, reveals mismatches that could potentially affect the published data. These mismatches are resolved as follows. First, a knowledgeable staff economist reviews the preliminary data. The analyst's judgment is used to determine if the submitted data are too incomplete to use. If so, the LED partner state normally corrects this problem and resubmits data. Otherwise, either a programming modification is made to resolve the mismatch or the mismatch is deemed "minor" and is resolved via the establishment-level final weight (Abowd et al. 2009).

**Imputation**

Valid personal characteristics (birth date and sex) link directly from other Census Bureau sources for about 95% of the SSNs. The remaining missing data are imputed using multiple imputations from a Bayesian Posterior Predictive Distribution (PPD) estimated from the non-missing data. Best residential address information is linked from StARS(10), which builds a current address for each calendar year by means of probabilistic record linking of IRS, Medicare, and other administrative data.

Valid establishment characteristics link directly from the Employer Characteristics File system, which is a longitudinal frame of establishments based primarily on QCEW records. Missing NAICS codes and workplace addresses are imputed once based on a longitudinal edit that copies missing information from chronologically close QCEW records.

The most important imputation in the LEHD Infrastructure File System, and the one with the most serious implications for block-level public-use data in *OnTheMap*, is the method used to resolve missing establishment identifier information on the wage records for employers that operate multiple

establishments within the same state. For most states, this imputation affects 30% to 40% of all jobs. For each job from a multi-unit employer, ten imputed establishments are sampled from a PPD that conditions on the entire demography of establishments and employee histories at the multi-unit UI account being considered. In addition, the PPD uses information about the relative odds of different commute distances (estimated from Minnesota data) to adjust the posterior probabilities. Most of the variation in posterior probabilities comes from the differences in the employment levels of the various establishments within the UI account. Workplace characteristics—NAICS code and address—are taken from the imputed establishment's data in the appropriate period(11). All ten imputed values are used in preparing the public-use data.

Once the integrated longitudinal data have acquired residence and workplace addresses, these addresses are geo-coded. More than 80% of the addresses are geo-coded to a rooftop latitude and longitude. Those that cannot be geo-coded to at least a census tract are treated as missing. The missing data model for geo-codes in *OnTheMap* is also a PPD that uses both workplace and residence geo-codes from the complete data as conditioning information(12).

**Confidentiality Protection**

The confidential micro-data that underlie *OnTheMap* can be summarized by a tabulation that gives the count of all jobs, all private jobs, primary jobs, and primary private jobs for each residence block, workplace block, and value of the conditioning variables—age groups, earnings groups, and industry groups. The residence block represents the residence geo-code of the employee. The workplace block and industry group come from the employing establishment. Age groups are coded from the individual's linked data. Earnings groups are coded directly from the UI wage record in the frame.

This tabulation is too disclosive to publish in this form. Confidentiality protections are applied in two steps. First, the block-level employment counts are protected using the system developed for the Quarterly Workforce Indicators(13). When this block-level employment count would have been suppressed under the QWI system, it is imputed using a PPD that conditions on the QWI suppression rules and the distribution of confidential suppressed values. Then, the residence block tabulations for each workplace block are synthesized using a PPD that conditions on the confidential tabulations and prior information that summarizes the confidentiality protection parameters. The residence block tabulation provides formal privacy protection by implementing probabilistic differential privacy(14). For computational tractability and improved analytical validity, the

residence block PPDs are estimated using a coarsened domain that places the block in a SuperPUMA, PUMA or tract depending upon the distance from the workplace. Given the synthetic data at the coarsened geo-code, blocks are sampled using a PPD that conditions only on the Census 2000 population counts for blocks within the coarsened geo-code.

**Public-use Data**

The public-use data consist of tabulations that contain the employment counts for each workplace block with positive confidentiality-protected employment and each residence block with positive synthetic employment. Summary tabulations for each workplace block (aggregating all residence blocks) and each residence block (aggregating all workplace blocks) are also produced. The data are accessible from the Census Bureau's graphical mapping application or by direct download.

**Measures of Reliability**

The reliability of the *OnTheMap* data was assessed by comparing the confidentiality protected data (labeled "posterior" in the comparisons below) with the raw unprotected data (labeled "likelihood" in the comparisons below). The reliability measures answer the question: How similar are the public-use data to the confidential data when many block-level comparisons are made? To perform this assessment, every workplace block was coded into one of three size groups: employment of 1-9 persons, 10-99 persons, or 100+ persons. This coding was done separately for each of the 27 unique combinations of age group (14-30, 31-54, 55+), monthly earnings (1-1,250, 1,251-3,333, 3,334+), and industry group (goods producing, trade-transportation-utilities, all others). For each workplace block, residence blocks were coded into five directions (same as workplace, northeast, southeast, southwest, northwest) in reference to the latitude and longitude of the workplace block and eight distances in miles (0, <1, 1-3.9, 4-9.9, 10-24.9, 25-99.9, 100-499.9, 500+). Eliminating structural zeroes there are 29 cells possible for each of the 27 sub-populations.

Table 1 summarizes the analytical validity of the *OnTheMap* data for a typical year for a small, medium, and large state, stratified by the employment level in the block using the Integrated Mean Squared Error (IMSE) as the measure of reliability. Consider the largest of these numbers, 0.02354, the IMSE for a workplace block with very little employment (1-9 persons) in a small state. This IMSE implies a maximum bias of 15% in the estimated distribution of residences for this block (on average) or a maximum estimation variability of the same magnitude(15). The smallest of these numbers, 0.00002 for a block with 100 or more employees in a large state has essentially no bias or estimation variability (<0.5% maximum).

**Table 1: Integrated Mean Squared Error (Posterior v Likelihood)**

| *Workplace block employment* | *All* | *1-9* | *10-99* | *100+* |
|---|---|---|---|---|
| *Small State* | 0.00104 | 0.02354 | 0.00093 | 0.00003 |
| *Medium State* | 0.00051 | 0.01327 | 0.00054 | 0.00002 |
| *Large State* | 0.00052 | 0.01527 | 0.00063 | 0.00002 |

Table 2 summarizes the analytical validity of the *OnTheMap* data for a typical year using the Kullback-Leibler (KL) divergence as the measure of reliability. Consider the largest of these numbers, 0.13662, the KL for a workplace block with very little employment (1-9 persons) in a small state. This KL implies an average divergence of approximately 14% between the estimated distributions (posterior and likelihood) of residences for this block(16). The smallest of these numbers, 0.00092 for a block with 100 or more employees in a large state has essentially no divergence (<0.1%).

**Table 2: Kullback-Leibler Divergence (Posterior from Likelihood)**

| *Workplace block employment* | *All* | *1-9* | *10-99* | *100+* |
|---|---|---|---|---|
| *Small State* | 0.01210 | 0.13662 | 0.02219 | 0.00215 |
| *Medium State* | 0.00807 | 0.11340 | 0.01650 | 0.00116 |
| *Large State* | 0.00715 | 0.11779 | 0.01565 | 0.00092 |

These results suggest that some caution should be taken when studying workplace areas that are very thinly employed (<10 workers). Otherwise, there is very little discrepancy between the public-use data and the underlying confidential data based on global comparisons of the edited confidential data to the posterior predictive distribution used to create the publication data.

Both the IMSE and the KL divergence are global measures of reliability—that is, they integrate over residential locations, for a given workplace location, and then over workplace locations that meet the stratification

criteria (workplace size). Neither one accounts for the fact that the published data are sampled from the relevant posterior distribution. External validity checks suggest that both of these measures correctly summarize the overall *OnTheMap* reliability; however, as with sampling variability, there can be discrepancies, sometimes large ones, for specific areas(17). Good statistical practice given the current evidence is that users should try to combine *OnTheMap* data with other data sources in order to assess the reliability of any particular analysis.

Users should also remember that, as in any statistical system, there can be quality variation in the input data that can compromise any analysis. Such quality variation may be difficult to detect since most of the administrative records underlying the major data sources are confidential. Even if knowledgeable civil servants can detect these anomalies using confidential records that they are authorized to view, there may not be an appropriate publication or correction venue due to the explicit, and overriding, confidentiality protections provided by both federal and state laws. This is not a new problem for statistical agencies, but it is particularly troublesome for a public-use product that is so heavily dependent on administrative records.

Many of the modeling assumptions that underlie *OnTheMap* and the Quarterly Workforce Indicators would benefit from additional formal statistical analysis. Resource limitations during the LEHD program's first decade, and continuing computing resource limitations, have made the production of a thorough quality profile difficult to accomplish. This isn't unusual for innovative Census Bureau programs. The Survey of Income and Program Participation released a quality profile for the 1984 panel in the late 1980s that is comparable to the LEHD Infrastructure technical paper first released in 2006(18). A full quality profile for the SIPP did not appear until 1998, more than a decade and a half after the first panel(19).

**Conclusions**

*OnTheMap* demonstrates that reliable block-level data can be published from longitudinally integrated employer-employee data with provable confidentiality protection. The primary benefit of this publication is not the block-level tabulations themselves but rather the flexibility that block-level publication provides to users who want to define their own analysis areas. It is important to recognize the explicit tradeoffs in such a system. Using traditional methods, many of the block level reports would have been suppressed from publication. This would have converted the problem to one that more closely resembles traditional small-area estimation because users (or the Census Bureau) would have had to model and estimate the suppressed data. Using modern probabilistic confidentiality protection

allows *OnTheMap* to publish noisy data for every block that are, on average, not substantially biased nor substantially noisy for workplace blocks with moderate (10-99) or high (100+) employment levels. These assessments are only valid when averaged over many analyses constructed from varying block-level summaries. Much additional work is required to provide a more detailed quality profile.

**References**


Abowd, J. M., B. Stephens, and L. Vilhuber (2006) "Confidentiality Protection in the Census Bureau Quarterly Workforce Indicators," LEHD Technical Paper 2006-02 available online at http://lehd.did.census.gov/led/library/techpapers/tp-2006-02.pdf (cited on March 26, 2010).

Abowd, J M., B. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer, and S. Woodcock (2009) "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators" in T. Dunne, J.B. Jensen and M.J. Roberts, eds., *Producer Dynamics: New Evidence from Micro Data* (Chicago: University of Chicago Press for the National Bureau of Economic Research), pp. 149-230 available online at http://www.vrdc.cornell.edu/news/?p=88 (cited on March 26, 2010).

Abowd, J M., B. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer, and S. Woodcock (2006) "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators," LEHD Technical Paper 2006-01 available online at http://lehd.did.census.gov/led/library/%20techpapers/tp-2006-01.pdf (cited on March 30, 2010)

Abowd, J. M. and L.Vilhuber (2005) "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers," *Journal of Business and Economics Statistics*, Vol. 23, No. 2 (April): 133-152, JBES Joint Statistical Meetings invited paper with discussion and "Rejoinder" (April): 162-165.

Bell, W. R. (2008) "Examining Sensitivity of Small-Area Inference to Uncertainty about Sampling Error Variances," Proceedings of the Survey Research Methods Section, American Statistical Association, available online at http://www.amstat.org/sections/SRMS/%20proceedings/y2008f.html (cited on March 25, 2010).

Benedetto, G., J. Haltiwanger, J. Lane, and K. L. McKinney (2007) "Using Worker Flows to Measure Firm Dynamics," *Journal of Business and*

*Economic Statistics*, Vol. 25, No. 3 (July): 299-313.

Cambridge Systematics, Inc. (2009) *Enhancing The American Community Survey Data as A Source for Hone-to-Work Flows*, NCHRP Project 08-36, Task 81, National Cooperative Highway Research Program, Transportation Research Board, available online at http://onlinepubs.trb.org/onlinepubs/nchrp/docs/NCHRP08-36(81)_FR.pdf%20(cited on March 28, 2010).

Census Bureau (1998) SIPP Quality Profile, SIPP Working Paper No. 230, available online at http://www.census.gov/sipp/workpapr/wp230.pdf (cited on March 30, 2010).

Fay, R. E. and Herriot, R. A. (1979) "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, 74, 269-277.

Fienberg, S. E.; and Love, T. (2009) *Disclosure Avoidance Techniques to Improve ACS Data Availability for Transportation Planners*, NCHRP Project 08-36, Task 71, National Cooperative Highway Research Program, Transportation Research Board, available online at http://onlinepubs.trb.org/onlinepubs/archive/NotesDocs/NCHRP08-36%2871%29_FR.pdf (cited on March 28, 2010).

King, K., R. Petroni, and R. Singh (undated) "Quality Profile for the Survey of Income and Program Participation," Survey of Income and Program Participation Working Paper No. 30, available online at http://www.census.gov/sipp/workpapr/wp30.pdf (cited on March 30, 2010).

Little, R (2006) "Calibrated Bayes: A Bayes/Frequentist Roadmap," The American Statistician, Vol. 60, pp. 213-223.

Machanavajjhala, A. D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber (2008) "Privacy: Theory Meets Practice On the Map," *International Conference on Data Engineering (ICDE) 2008*, pp. 277-286.

Rao, J.N.K. (2003), *Small Area Estimation* (Hoboken, New Jersey: John Wiley).

Rubin, D. B. (1984) "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," Annals of Statistics, Vol 12, pp. 1151-1172.

Wu, J. and J. M. Abowd (2008) "Synthetic Data for Administrative Record Applications at LEHD," Census Bureau technical report available online at

http://lehd.did.census.gov/led/library/presentations/Wu-Abowd-20070831.pdf (cited March 30, 2010).

Comments or questions? Please feel free to contact the author at john.abowd@cornell.edu.

Notes:

*1. I am grateful to Jeremy Wu, ADC for the Longitudinal Employer-Household Dynamics program in the Center for Economic Studies at the Census Bureau for many helpful comments on this article. This program is funded as part of the Local Employment Dynamics initiative in the Census Bureau's 2010 budget. The opinions in this article are those of the author alone and do not represent the opinions of the U.S. Census Bureau or any of the sponsors of the LEHD program. Funding from the National Science Foundation is gratefully acknowledged.*

*2. See http://lehdmap4.did.census.gov/themap4/ (cited on March 25, 2010).*

*3. See Rubin (1984), Little (2006), and Wu and Abowd (2008).*

*4. See Fay and Herriot (1979), Rao (2003), and Bell (2008).*

*5. This is the view from 10,000 meters. See http://www.census.gov/did/www/saipe/index.html (cited on March 25, 2010) for details.*

*6. See http://www.bls.gov/opub/hom/homch5_b.htm (cited on March 25, 2010).*

*7. The Census Bureau actually removes the SSNs from the confidential data files, replacing them with an internally generated confidential identifier known as a Protected Identification Key (PIK) that is a one-time-pad encryption of the SSN.*

*8. The LEHD Infrastructure file system uses a multi-state unique recoding of the UI account number called a State Employer Identification Number (SEIN) internally.*

*9. See Abowd and Vilhuber (2005) and Benedetto et al. (2007).*

*10. See http://www.census.gov/sipp/DEWS/CNSTAT_01-26-*

*07SallyObenski.ppt (cited on March 25, 2010) and "The Statistical Administrative Records System: System Design, Successes, and Challenges (StARS)," incorporating data from seven major Federal databases: the Internal Revenue Service (IRS) 1040 Master File, IRS Information Returns file, Selective Service registration file, Medicare Enrollment Database file, Indian Health Service patient file, Housing and Urban Development Tenant Rental Assistance System file, and the Social Security Administration Numident file, available online at http://nisla05.niss.org/affiliates/dqworkshop/papers/judson-background.pdf (cited on March 28, 2010).*

*11. There are more details in Abowd et al. (2009).*

*12. OnTheMap uses a missing geo-code imputation model that is more sophisticated than the one used in the Quarterly Workforce Indicators, which uses only employer data to condition the PPD.*

*13. See Abowd, Stephens and Vilhuber (2006).*

*14. See Machanavajjala et al. (2008) for the algorithm and F. Andersson http://www.vrdc.cornell.edu/news/wp-content/uploads/2009/08/1-5-Andersson.pdf (cited on March 26, 2010) for the implementation details.*

*15. The IMSE is the integrated sum of the squared bias and the variance of the estimated proportions.*

*16. The KL is reported in log points. Percentage deviations can be calculated as 100 x (exp(KL)-1).*

*17. See Feinberg and Love (2009) and Cambridge Systematics (2009).*

*18. See King, Petroni and Singh (undated) and Abowd et al. (2006).*

*19. See Census Bureau (1998).*

Go Back