

LEHD: Understanding Its Quality

Local Employment Dynamics

All About Jobs

Jeremy S. Wu

April 30, 2010

I. EXECUTIVE SUMMARY

“By providing good data for good policy, you could contribute to society.”¹

Peter Miller

President, American Association for Public Opinion Research

The importance of good data cannot be overstated. However, few can agree exactly on what data quality actually is, especially for a hybrid program like Longitudinal Employer-Household Dynamics (LEHD) that is not a census, a survey, or a set of administrative records.

As an innovative program, LEHD has already produced cost-effective new data and applications that are unmatched by any other federal statistical systems. However, as the program continues to grow with its 2010 Local Employment Dynamics (LED) budget initiative, it is imperative that we be able to answer simply and clearly the question:

How good are the LEHD data?

This paper describes the underpinnings of the legal and professional requirements concerning information quality, the unique statistical characteristics of the LEHD program, its applicable quality factors, the continuing efforts to improve the quality of LEHD data, and the upcoming challenges. Pursuing a common understanding of LEHD’s quality will help to build a shared vision for the program in the coming decades.

II. APPLICABLE QUALITY LAWS AND CONCEPTS

According to the National Research Council,² commitment to Quality and Professional Standards of Practice is a fundamental practice for a Federal statistical agency. Section 515³ of the Treasury and General Government Appropriations Act for Fiscal Year 2001, commonly

¹ Washington Post, *Groves brings scholarly depth to bear in leading census, winning over critics*, March 31, 2010. Available at <http://www.washingtonpost.com/wp-dyn/content/article/2010/03/30/AR2010033003675.html> on April 4, 2010.

² Martin, Margaret E; Straf, Miron L.; and Citro, Constance F. (2005). *Principles and Practices for a Federal Statistical Agency*. Committee on National Statistics, Division of Behavioral and Social Sciences and Education, National Research Council of the National Academies. The National Academies Press, Washington, DC.

³ Available at <http://www.fws.gov/informationquality/section515.html> on March 6, 2010.

known as the Information Quality Act, directs the Office of Management and Budget (OMB) to issue government-wide policy and procedural guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information that Federal agencies disseminate.

OMB issued the corrected final guidelines⁴ in 2002. The Census Bureau defines Data Quality⁵ in 2006 as “fitness for use” according to the needs of its customers, which include all branches and levels of the Federal government, state and local governments, and the public. As foundation and operating principles, the Census Bureau further defines six dimensions of data quality: relevance, accuracy, timeliness, accessibility, interpretability, and transparency.

The Census Bureau has also established quality performance principles.⁶ These cover: the definition of data quality; the development of concepts and methods; the planning and design of surveys and other means of collecting data; the collection, processing, editing, and analysis of data; the production of estimates and projections; the establishment of review procedures; and the dissemination of statistical information products to the public. In turn, the Census Bureau quality standards define the implementation of the performance principles.⁷

III. LEHD AND ITS GROWTH

The predominant activities of the Longitudinal Employer-Household Dynamics (LEHD) program are to collect, compile, process, and disseminate information for a 21st century statistical system on the dynamics of the U.S. labor market.

Under the Local Employment Dynamics (LED) partnership, the emerging LEHD data infrastructure is a longitudinal national frame of jobs, linking workers with businesses for which they work over time. The LED state partners provide historical and ongoing unemployment insurance wage records and business records as the foundation for the LEHD.

Unlike other statistical frames whose primary purpose is to facilitate random sampling, the comprehensive LEHD infrastructure has made it possible to create innovative data such as Quarterly Workforce Indicators⁸ (QWI) and OnTheMap.⁹ A new data line is also under development, on the dynamics of job flows. These data have not been available before in such quantity, magnitude, and refined geographical resolution.

LEHD is unique because it is a longitudinal data system. It relies heavily on the processing power and storage capacity of modern information technology to leverage existing data sources and integrate them into a comprehensive, near-census statistical system. However, LEHD is not a census, a statistical survey, or a set of administrative records. It is a system of records that are collected not according to probability sampling. As a result, LEHD data do not possess typical

⁴ Available at <http://www.whitehouse.gov/omb/assets/omb/fedreg/reproducible2.pdf> on March 6, 2010.

⁵ Available at http://www.census.gov/quality/P01-0_v1.3_Definition_of_Quality.pdf on March 6, 2010.

⁶ Available at http://www.census.gov/quality/quality_guidelines.htm on March 6, 2010.

⁷ Available at http://www.census.gov/quality/quality_standards.htm on March 6, 2010.

⁸ Available at <http://lehd.did.census.gov/led/datatools/qwiapp.html> on March 6, 2010.

⁹ Available at <http://lehdmap4.did.census.gov/themap4/> on March 6, 2010.

statistical properties, and many of the existing quality standards based on probability sampling do not apply.

While LEHD benefits from the quality practices of the various existing data sources, it also suffers from their shortcomings, which LEHD itself cannot easily correct. While state-of-the-art methods have been introduced to protect confidentiality and retain analytical validity for the new data, these methods are not well understood or adequately transparent for more-detailed review of theoretical or empirical support.¹⁰

Beginning with fiscal year (FY) 2010, LEHD is receiving requested appropriations¹¹ to begin the transformation of LEHD from a research pilot into a core, multidisciplinary operational program in the Census Bureau. The purpose of the LED initiative is to “provide federal, state, and local policymakers and planners, businesses, private sector decision makers, and Congress with comprehensive and timely national, state, and local information on the dynamic nature of businesses and their workers.”¹²

Although LEHD data are intended as labor market information, they are also applicable for workforce and economic development, emergency management, transportation planning, and education information.

LEHD is no longer a research pilot—instead, LEHD data are being used for community grant applications and to assess the impact of individual disasters on property and lives. It is therefore imperative that LEHD responsibly ensure that its data are of acceptable quality so that they can be used reliably and confidently for grant applications, protection of lives and properties, and policy-making. Although explicit, applicable quality standards may not exist for integrated data at this time, data created and released by LEHD are now subject to the Census Bureau’s operating principles of relevance, accuracy, timeliness, accessibility, interpretability, and transparency.

IV. QUALITY FACTORS FOR LEHD

While quality has many aspects and interpretations, an Ishikawa diagram is a tool commonly used to identify factors that can cause an overall effect. “Each cause or reason for imperfection is a source for variation. Causes are usually grouped into major categories to identify the sources of variation.”¹³ An Ishikawa diagram may also be known as a cause-and-effect diagram due to its intended purpose, or a fishbone diagram due to its appearance.

¹⁰ Feedback from December 4, 2009 quarterly meeting of the Council of Professional Associations on Federal Statistics (COPAFS), whose website is located at <http://www.copafs.org/>, available on March 15, 2010.

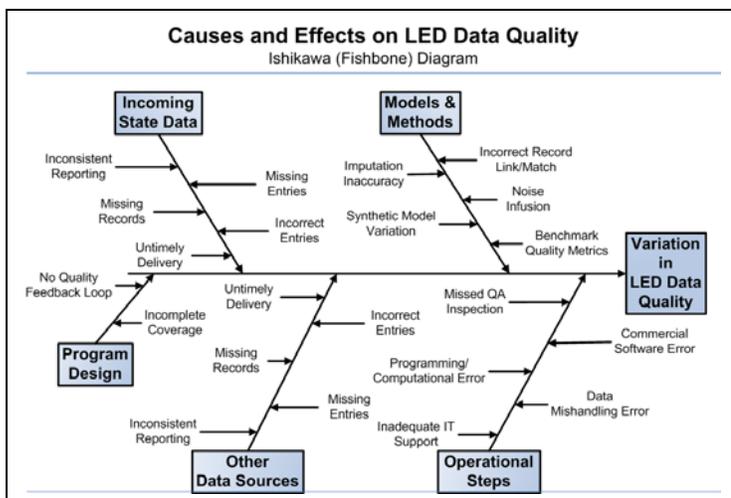
¹¹ FY2010 Congressional Budget Justification for the Census Bureau, available at <http://www.osec.doc.gov/bmi/budget/10CJ/Census%20FY%202010%20Congressional.pdf> on March 6, 2010.

¹² Exhibit 300: Capital Asset Plan and Business Case Summary, page 1, available at http://ocio.os.doc.gov/s/groups/public/@doc/@os/@ocio/@oitpp/documents/content/prod01_008010.pdf on March 6, 2010.

¹³ Ishikawa diagram, available at http://en.wikipedia.org/wiki/Ishikawa_diagram on March 6, 2010.

An Ishikawa diagram for LEHD was developed several years ago. The major categories of causes have been identified in the latest version of the Ishikawa diagram to the right as:

- Program Design,
- Incoming State Data,
- Other Data Sources,
- Models and Methods, and
- Operational Steps.



Notably absent from this diagram are the causes of sampling variations, because LEHD has only non-sampling variations.

The distinction between state-supplied and other data is intended to reflect their respective potentials for corrective or follow-up actions: some corrective actions are possible for state-supplied data, while practically none are possible for other data sources.

Measurement issues are currently encapsulated as benchmark quality metrics in Models and Methods. Given its significance, it may however be more appropriate to recast Measurement Issues as a major category of causes for LEHD in the future.

V. CONTINUING IMPROVEMENT EFFORTS

To the extent practicable, LEHD has used the Ishikawa diagram to make continuous improvements according to quality management principles advocated by W. Edwards Deming.¹⁴ These improvements and immediate plans are discussed below.

5.1 Program Design

5.1.1 Completeness (Partnership). A national LED Partnership will cover 50 states, the District of Columbia, Puerto Rico, and the U.S. Virgin Islands that participate in both the Unemployment Insurance (UI) and Quarterly Census of Earnings and Wages (QCEW) programs.

The current completeness rate is 51 out of 53 potential state partners, or 96.2 percent. Massachusetts has agreed to join LED when a national agreement is ready for the state's approval and signature. New Hampshire is undergoing legislative change that will enable it to share data with the Census Bureau and join LED. A national agreement has been prepared and completed the Department of Commerce review; it is intended to extend the partnership uniformly for 10 years until the year 2020.

¹⁴ W. Edwards Deming (1988). *Out of the Crisis*. ISBN 0-911379-01-0. Massachusetts Institute of Technology, Cambridge, Massachusetts.

5.1.2 Completeness (Production). Under the LED agreement a state partner supplies its historical and ongoing data files to join Regular Production, such that the state's data will be integrated into the LED data infrastructure.

Among the 51 LED partner states, 47 or 92.2 percent are in Regular Production. Experimental Production for Connecticut has been completed; Connecticut is expected to respond by June on joining Regular Production. Release of Washington, DC data was scheduled to pair with the release of Federal employment data in 2010. Start-up for Puerto Rico has been expected since November 2009. The U.S. Virgin Islands are awaiting the supply of their complete historical files.

5.1.3 Completeness (Content). According to the Bureau of Labor Statistics (BLS),¹⁵ QCEW covers about 98 percent of U.S. jobs. In addition, “employment data under the QCEW program represent the number of covered workers who worked during, or received pay for, the pay period including the 12th of the month. Excluded are members of the armed forces, the self-employed, proprietors, domestic workers, unpaid family workers, and railroad workers covered by the railroad unemployment insurance system....The QCEW program does provide partial information on agricultural industries and employees in private households.”¹⁶ Federal civilian workers, U.S. postal workers, and some undocumented workers are also known to be missing from the current LED infrastructure.

Addition of Federal workers and the self-employed is part of the LEHD work plan under the FY2010 budget initiative. The work plan also includes the addition of race, ethnicity, and education to the existing worker demographic profile of age and gender. The current LED data infrastructure focuses only on the employed; therefore, some data gaps exist for workers who became unemployed, returned to school, gave up work, retired, or otherwise disappeared from the labor force.

5.1.4 Quality Loops. Data quality starts at the origin of data supply. However, existing data sources are designed to serve their original purposes, not necessarily for the best interests of LEHD. A typical example is the “place of work” data element from state-supplied administrative records. The payroll office or the headquarters of a school district is sufficient to serve the purpose of the UI system, but it does not reflect the individual school locations where teachers actually work. When the results are displayed at the census block level in OnTheMap, they appear to be anomalous. Several pilot efforts have been made to establish quality loops, including engaging local metropolitan planning organizations for “place of work” refinement, and developing audit reports with state partners to ensure file integrity during transmission. However, an identified data quality problem may not be corrected at the origin of data supply. Another example of the lack of an active quality loop is the reported incorrect spelling of a street name in the TIGER¹⁷ files, which can be remedied only by the Census Bureau Geography Division.

¹⁵ Available at <http://www.bls.gov/cew/> on March 20, 2010.

¹⁶ Available at <http://www.bls.gov/cew/cewover.htm> on March 20, 2010.

¹⁷ Topologically Integrated Geographic Encoding and Referencing (TIGER) file, available at <http://www.census.gov/geo/www/tiger/> on March 20, 2010.

5.2 State-Supplied Data

Standard Operating Procedures (SOP) 3000¹⁸ requires a state partner to supply four types of files in standard formats to the Census Bureau:

- UI wage records, historical and ongoing quarterly;
- QCEW records, historical and ongoing quarterly;
- Workforce Investment Board (WIB) definitions, one-time and as needed with changes; and
- Longitudinal Data Base (LDB) files for bridging Standard Industrial Classification (SIC) and North American Industry Classification System (NAICS), one time and if applicable.

5.2.1 Quality Factors. Data supplied by the state partners may suffer from inconsistent reporting, missing records, untimely delivery, missing entries, and incorrect entries. LEHD may ask the state partners to resubmit data files as corrective action, or to skip Regular Production until adequate corrective actions have been applied.

5.2.2 Standard Metrics. How good are the state-supplied data? Standard quality metrics have not been fully established under this category, but they include:

- **Initial Acceptance.** Data quality must be at an acceptable level before a state partner can join Regular Production of Quarterly Workforce Indicators (QWI). Among the 51 existing partners, 4 have not reached this level: Connecticut, District of Columbia, Puerto Rico, and Virgin Islands. The other 47 state partners are under Regular Production.
- **Continuing Acceptance.** Regular Production of QWI is skipped for a state partner if the state-supplied data for the current production cycle is missing, late, or deemed to be of unacceptable quality.
- **Measurement Methods.** A partial indicator of quality for the state-supplied data is the number and percentage of skipped production since 2006Q2:

	FY06 (No Quarter1)	FY07	FY08	FY09
Eligible Production States¹⁹	101	158	178	187
Total – Skipped	10	12	5	7
Due to missing/late data	7	3	0	2
Due to quality issues	3	9	5	5
Total – Skipped Percent	9.9%	7.6%	2.8%	3.7%

¹⁸ SOP3000, Joining the Local Employment Dynamics Partnership, available at <http://lehd.did.census.gov/led/partnersonly/sop.html> on March 20, 2010.

¹⁹ Sum of eligible Regular Production states during specified fiscal year.

One analyst uses ad hoc methods resembling control charts to detect extreme values and abnormal patterns as the quality assurance (QA) process during the Regular Production of QWI. However, the LEHD control charts are not based on statistical criteria. They produce “cautions” and “warnings,” but have no specific rules or goals as in Six Sigma²⁰ on follow-up actions, which are primarily invoked based on the judgment of the lone QA Analyst or the LEHD Program Manager.

Although continuous efforts have been made to improve the quality of the state-supplied data, progress cannot be measured quantitatively because QA data were not retained for analysis. Initial efforts have recently started to assemble and retain historical QA data to analyze the historical trends of “cautions” and “warnings” for the production of QWI, adequacy of file submissions, presence or absence of top firms in file submissions, and concordance of worker and employer links. These efforts will be used to develop quantitative metrics on the goodness of the state-supplied data. Two contractor statistical analysts are also being recruited at this time.

5.3 Other Data Sources

LEHD integrates state-supplied data with additional data sources to create its core infrastructure and data products. These additional data sources include:

- **U.S. Census Bureau**
 - 2008 and 2007 Master Address File
 - 2008 and 2004 TIGER/Line Shapefiles
 - 2007 Statistical Administrative Records Systems (StARS)²¹ Personal Characteristics File
 - 2007 American Community Survey Place of Work file, ongoing quarterly
 - 2006 Geographical Reference Files
 - 2002, 1997, 1992, and 1987 Economic Censuses
 - 2002 American Housing Survey
 - 2001 Survey of Income and Program Participation (SIPP)
 - 2001 Business Register
 - 2000 and 1990 Decennial Censuses of Population and Housing
 - 2000 Census Transportation Planning Package
 - 1999 StARS Composite Person Record
 - 1997 Current Population Survey (CPS)
 - Annual surveys of manufacturing, service, trade, transportation and communications industries, unknown vintages

²⁰ Basic history and definition available at http://en.wikipedia.org/wiki/Six_Sigma on March 20, 2010.

²¹ 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 at <http://nisl05.niss.org/affiliates/dqworkshop/papers/judson-background.pdf> on March 20, 2010.

- **Bureau of Labor Statistics, U.S. Department of Labor**
 - Quarterly Census of Employment and Wages, ongoing quarterly
- **Employment and Training Administration, U.S. Department of Labor**
 - WIRED Region definitions
- **U.S. Department of Education**
 - 2007 Integrated Postsecondary Education Data System
 - 2006-2007 Common Core of Data
- **U.S. Department of Transportation**
 - 2008 National Transportation Atlas
- **U.S. Office of Personnel Management**
 - Central Personnel Data File, ongoing quarterly
- **Pitney Bowes**
 - Code 1 Plus Address Update, ongoing monthly

5.3.1 Quality Factors. LED also encounters issues with these data sources, involving varying degrees of inconsistent reporting, missing records, untimely delivery, missing entries, and incorrect entries. However, LEHD is very limited by the corrective actions that the original data source can undertake.

5.3.2 Standard Metrics. How good are the data from these other sources? Data supplied by other existing sources are subject to the quality practices of the respective programs and how LEHD makes use of the data. LEHD usually accepts these data sources as presented. The intent of the additional data sources is to provide additional demographic and geographic data elements, continuing updates, coverage expansion, and benchmark comparisons.

Standard quality metrics have not been established under this category; available documents tend to be limited or outdated. A major needed effort is to create an inventory of their statistical profiles, quality attributes, vintage and use in the LEHD infrastructure and data products, and update practices. The LED data's longitudinal nature means that proper vintage is a major consideration and challenge to the LED data quality.

LEHD has no current work plans to integrate data from the 2007 or 2012 Economic Census, the 2010 Decennial Census, the annual American Community Survey, CPS, and SIPP. Their importance for the continuing quality of the longitudinal LEHD infrastructure has not been studied or assessed.

The Geo-coded Address List (GAL) is a key file system in the LEHD infrastructure, containing unique commercial and residential addresses in a state, geo-coded to the census block level and latitude/longitude coordinates. GAL is considered severely outdated due to (a) the commercial termination of the address un-duplication software, (b) inability to update address information on a longitudinal and consistent basis for the LEHD infrastructure, and (c) lack of a clear decision

tree and criteria to choose among multiple address sources. No quantitative measure exists of the goodness of the current GAL information.

Carl Anderson, most recently Director of Geographic Information Systems for the Fulton County Government in the state of Georgia, started work as a contractor Geographer in April. His primary assignment is to enhance the GAL process and develop goodness measures.

5.4 Models and Methods

LEHD employs a combination of innovative model-based approaches and observational studies to build its core infrastructure and data products. The econometric/statistical models and computational methods serve these primary purposes:

- **Linkage of records** for demographic information based on exact or probabilistic matching of records according to the Protected Identification Key. Records are also linked for geographic information based on unduplicated, standardized matching of geocodes for residential or work places.
- **Imputation** to replace missing values, including non-response values for age and gender, allocation of workers to multi-site firms, and cross-walks for public sector offices or SIC-NAICS transition.
- **Infusion of noise**, applied to estimates of all workplace-level measures to protect confidentiality.
- **Generation of synthetic data** to protect confidentiality and retain analytical validity at the census block level for the OnTheMap application.

The combination of these models and methods provide the distinct capability for LEHD to produce the QWI and OnTheMap data for public use.

5.4.1 Quality Factors. Innovation is key to the success of LEHD, although it is not easily quantified. The LEHD approach does not follow strictly the Bayes or the Frequentist paradigm. The LEHD approach matches most closely to what Roderick Little described as the Pragmatists “who don’t have a clear philosophy and choose what seems to work.”²² Little further quoted Donald Rubin on the Bayes/Frequentist compromise:

“The applied statistician should be Bayesian in principle and calibrated to the real world in practice – appropriate frequency calculations help to define such a tie ... frequency calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test; here the

²² *Calibrated Bayes. A Bayes/Frequentist Roadmap*, Presidential Invited Address by Roderick Little to the 2005 Joint Statistical Meetings, available at <http://astor.som.jhmi.edu/~sining/BM/articles/rodlittle.pdf> on March 20, 2010.

technique is the calibration of Bayesian probabilities to the frequencies of actual events.²³

LEHD is predominantly a model-based approach using observational studies for which the common considerations of model quality include:

- theoretical foundation of the model,
- sensitivity or validity of the underlying assumptions, and
- empirical support.

5.4.2 Standard Metrics. How good are the models and methods LEHD uses? LEHD's innovativeness is unmatched by any federal statistical system in recent history. Successfully employing the pragmatic modeling approach and formally defining privacy in the statistical system are among the many creative approaches and practices that LEHD has taken. The Kullback-Leibler (K-L) divergence, or its symmetric variant, has been suggested as a quality metric for LEHD imputation and synthetic models. On the other hand, the underlying statistical properties of the pragmatic approaches and models are basically unknown. Despite the large amount of data, large-sample asymptotic theories do not necessarily apply, because of LEHD's emphasis on refined geographies such as the census block level. A key model assumption of ignorability, or "missing at random," for the imputation models remains to be validated by empirical evidence, especially when the missing value rates are high.

The Transportation Research Board has begun to make limited evaluations of OnTheMap data against benchmark results from the American Community Survey²⁴ and on the appropriate use of Bayesian techniques for data synthesis.²⁵

The expected addition of Professor Xiao-Li Meng, Chair of the Department of Statistics at Harvard University, as a Distinguished Senior Research Fellow²⁶ in late summer of 2010 will provide strong impetus for LEHD to build on its theoretical foundations and strengthen statistical quality and research—not only for LEHD, but also for the statistical profession. This fruitful and productive collaboration between the Census Bureau and academia has notably helped to establish and sustain the LEHD program, as well as creating a source of recruitment for new staff and talents. Professor Meng's participation is expected to be part-time; Alex Blocker, an intern from Harvard University, and at least two senior-level Mathematical Statisticians will also be

²³ Rubin, Donald B. (1984). *Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician*. *Annals of Statistics* 12, pp. 1151-1172.

²⁴ Cambridge Systematics, Inc. (2009). Enhancing The American Community Survey Data as A Source for Home-to-Work Flows, NCHRP Project 08-36, Task 81, National Cooperative Highway Research Program, Transportation Research Board, available at http://onlinepubs.trb.org/onlinepubs/nchrp/docs/NCHRP08-36%2881%29_FR.pdf on March 20, 2010.

²⁵ 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 at [http://www.trb.org/NotesDocs/NCHRP08-36\(71\)_FR.pdf](http://www.trb.org/NotesDocs/NCHRP08-36(71)_FR.pdf) on March 20, 2010.

²⁶ Professor John Abowd of Cornell University and Professor John Haltiwanger of University of Maryland are current Distinguished Senior Research Fellows to LEHD.

needed to support this activity. The vacant Quality and Statistics Branch Chief position will also be filled.

The most fundamental quality metrics for record linkage and imputations include, for example:

- How frequently is probabilistic record match possible and correct?
- How do the match rates differ between exact and probabilistic record matching?
- How frequently is the age or gender of a worker correctly imputed? How much do their match rates vary by subgroups?
- How much can the match rate vary by state, or by more-refined geography such as county or census block?

These empirical results for LEHD tend to be sporadic, outdated, or to have not been habitually monitored and analyzed over time. In general, statistical profiles or measures of goodness of fit for Bayesian models to the observed data²⁷ have not been fully established for LED data.

5.5 Operational Steps

LEHD relies heavily on advanced information technology for data processing, storage, and analysis. Regular Production of the QWI is defined by a series of up to 16 concurrent and sequential processes. The SAS computer programs for these processes were typically developed by economists, and subsequently reviewed and enhanced by the SAS programming staff when the latter became more established and available. Total wall-clock processing time in the most recent quarter exceeded 9,000 hours for the production of QWI of 45 states. The amount of public-use output data is measured in terabytes. The QWIs are disseminated to the LED state partner by DVD and the public via the Internet.

Given its highly sensitive nature, LEHD cannot afford to make serious missteps in handling or releasing its data. Violations are subject to imprisonment and/or fines under Title 13 and Title 26. Bureau-wide policies and practices, training, standard operating procedures, physical restrictions, state-of-the-art methodologies, and management controls are additional standard safeguards of LEHD data.

5.5.1 Quality Factors. The three primary quality factors are:

- **Custom programming.** The custom programs may contain inefficiencies, incorrect codes, or improper implementation of algorithms. LEHD has started to formalize a code review and change process in accordance with Capability Maturity Model Integration (CMMI)²⁸ principles.

²⁷ Gelman, Andrew; Meng, Xiao-Li; and Stern, Hal. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica* 6, pp. 733-807, available at <http://www3.stat.sinica.edu.tw/statistica/oldpdf/A6n41.pdf> on March 20, 2010.

²⁸ Available at <http://www.sei.cmu.edu/cmmi/> on April 12, 2010.

- **Human errors.** Practically every operational step has a human component that is subject to errors. Communications, coordination, management controls, and staff vigilance have helped to minimize their occurrence.
- **Commercial software.** Off-the-shelf commercial software such as SAS and Group1 is not immune from errors, although the likelihood is considered low.

5.5.2 Quality Metrics. How good are the LEHD operational steps? There are again few quantitative measures, but the occurrence of human and commercial software errors has been infrequent, and the magnitudes of their impact have been small when they occurred.

The code review and change process is being enhanced as additional resources from the 2010 LED budget initiative become available, including the installation of version control and ticket tracking systems. A Production Code Manager vacancy has been approved by the Human Resources Division and the Information Technology Directorate; it will be announced soon.

VI. UPCOMING CHALLENGES

How good are the LEHD data?

There is currently no concise and clear quantitative answer to this simple question. The eventual response will depend on:

- What quality metrics do we use?
- When do we think we have a data quality problem?
- What can be done when a data quality problem is identified?

However, several upcoming challenges suggest that there is urgency for the timely development of quality metrics for LEHD:

1. Completion of research for the addition of race, ethnicity, education, and federal workers to the LEHD infrastructure is expected to occur on May 1, 2010. Whether each of these data elements or data sources is of reasonable quality to be accepted for the next phase of implementation has to be decided.
2. An Economic Directorate Quality Audit has been scheduled to start for LEHD on December 6, 2010, with document presentation to begin on November 1. The Quality Audit Program will assess the LEHD program's compliance with the Standards set by the Office of Management and Budget.

As the LEHD user base has grown over time, LEHD users have also become a source to identify data quality problems through their inquiries and complaints. Rapidly developing data

visualization tools such as IBM Many Eyes²⁹ and Google Fusion Table³⁰ and Public Data³¹ also allow rapid problem identification, despite the large amount of LEHD data.

In general, the trend of publishing more government information online and the need to improve the quality of government information will be continued, if not accelerated, by the Open Government Initiative.³²

²⁹ Available at <http://manyeyes.alphaworks.ibm.com/manyeyes/> on April 3, 2010.

³⁰ Available at <http://tables.googlelabs.com/public/tour/tour1.html> on April 3, 2010.

³¹ Available at <http://www.google.com/publicdata/home> on April 3, 2010.

³² Available at www.whitehouse.gov/omb/assets/memoranda_2010/m10-06.pdf on April 12, 2010.