Discussion of Presentations on "Commercial Big Data and Official Economic Statistics"

John L. Eltinge
U.S. Bureau of Labor Statistics

Presentation to the Federal Economic Statistics Advisory Committee

June 12, 2015

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

www.bls.gov

Acknowledgements and Disclaimer

The author thanks the organizers and the speakers for the opportunity to discuss the presentations in this session; and Ken Robertson, Rick Clayton, Dave Talan and Akbar Sadeghi for providing the graphs in slides 13 and 14. Some of the comments in the paper are based on conversations with many colleagues over the past decade.

The views expressed here are those of the author and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

BLS

Overview

I. Many thanks to the authors

1. Much food for thought
   – extensive discussion warranted

2. Part of broad & deep societal reconsideration of (statistical) information over decades:

   - expectations on quality, cost, risk, credibility, accountability and access
   - tools to address
   - resource allocation: amount and mechanisms

BLS

Overview (Continued)

II. Context:

A. Five dimensions of design of federal statistical programs: stakeholder information needs, data sources, methodology, production systems, management

B. Data sources:

1. Common definition of “big data” – 4Vs: Volume, velocity, variety and veracity

2. Related: “organic data” and “non-designed data” (Groves, 2012; Couper, 2013; others)

3. Today: Commercial sources (some aggregated?)

Overview (Continued)

III. This discussion: Methodological topics anchored in the need of federal statistical agencies to balance multiple dimensions of

data quality, cost and risk

1. Walk: Concepts and tools for currently available data

   - Largely adapted from previous tools for survey and administrative data

2. Run: Institutional capacity to evaluate, manage and integrate prospective future dynamic flow of datasets

BLS

Overview (Continued)

IV. Topics:

A. Qualitative questions

B. Formal questions on methodology

C. Managerial issues that inform methodological approaches

Throughout: Links with presentations by Drs. Woodward, Silverman and Carroll

A. Qualitative Questions

1. Do these (aggregated) data provide useful, cost-effective and reliable economic information? (Broadly: "fitness for use")

   a. In addition to currently collected data?
      - Improve timeliness & frequency (Silverman, Woodward), quality?
      - Expand statistical product lines?
      - In-depth analyses for special cases (Gelman et al., 2014)

   b. Instead of (some) currently collected data
      - Direct replacement (full pop or specialized subpop)?
      - Basis for imputation, weighting?
      - Reduce costs, burden?

BLS

A. Qualitative Questions (Continued)

2. Consistent with stakeholder expectations (explicit or implicit; cf. RSS 1/19/2015) on

   a. Standard quality measures:

      Relevance, accuracy, timeliness, comparability, coherence, accessibility (e.g., Brackstone, 1999)

   b. Transparency and reproducibility?

   c. In re-considering (2.a) and (2.b), distinguish carefully between
      - fundamental concepts & principles
      - specific implementation

BLS

B. Formal Methodological Questions

Inferential basis provided by a given (aggregated) data source, in conjunction with some or all of our previously collected data? (Including some issues identified by Kirk Wolter in the December 9, 2011 FESAC discussion of web scraping.)

1. New estimands?
   a. Descriptive: Means, totals (and quantiles?)
   b. Relationships: Regression & GLMs; hierarchical models
   c. Full population? Specialized subpopulation(s)?
   d. Improved variable measurements? (Difficult or impossible to collect through surveys?)
      Carroll – medical vs. non-medical debt; new/any account
      Silverman – transactions and balances

BLS

B. Formal Methodological Questions (Continued)

2. Dominant components of uncertainty (bias, variance)?

    a. Variable specification (temporal and cross-sectional aggregation; definitional issues; reference periods)

    b. “Unit problem” (Silverman – user vs. household; Woodward – establishment vs. firm; ENBES, 2014)

B. Formal Methodological Questions (Continued)

c. "Representativeness" issues from:

- (Sub) population coverage:
  Silverman et al. – Table 1, substantive factors that drive differential participation in MintBills (esp age, edu)
  Carroll: credit invisibles/stale/insufficient unscored

- Go beyond coverage rates to examine bias diagnostics

- Representativeness: impact on mean, variance, cyclical effects (Woodward)

- Reasons for under/over-coverage can inform propensity models, prospective adjustments

BLS

B. Formal Methodological Questions (Continued)

c. "Representativeness" issues (continued):

- Unit survival rates, relative to full pop (Woodward)
- Cohort effects - marketing strategy of data (Woodward)
- Changing cohort effects (Woodward)
- Unit-level missingness (including slow reports)

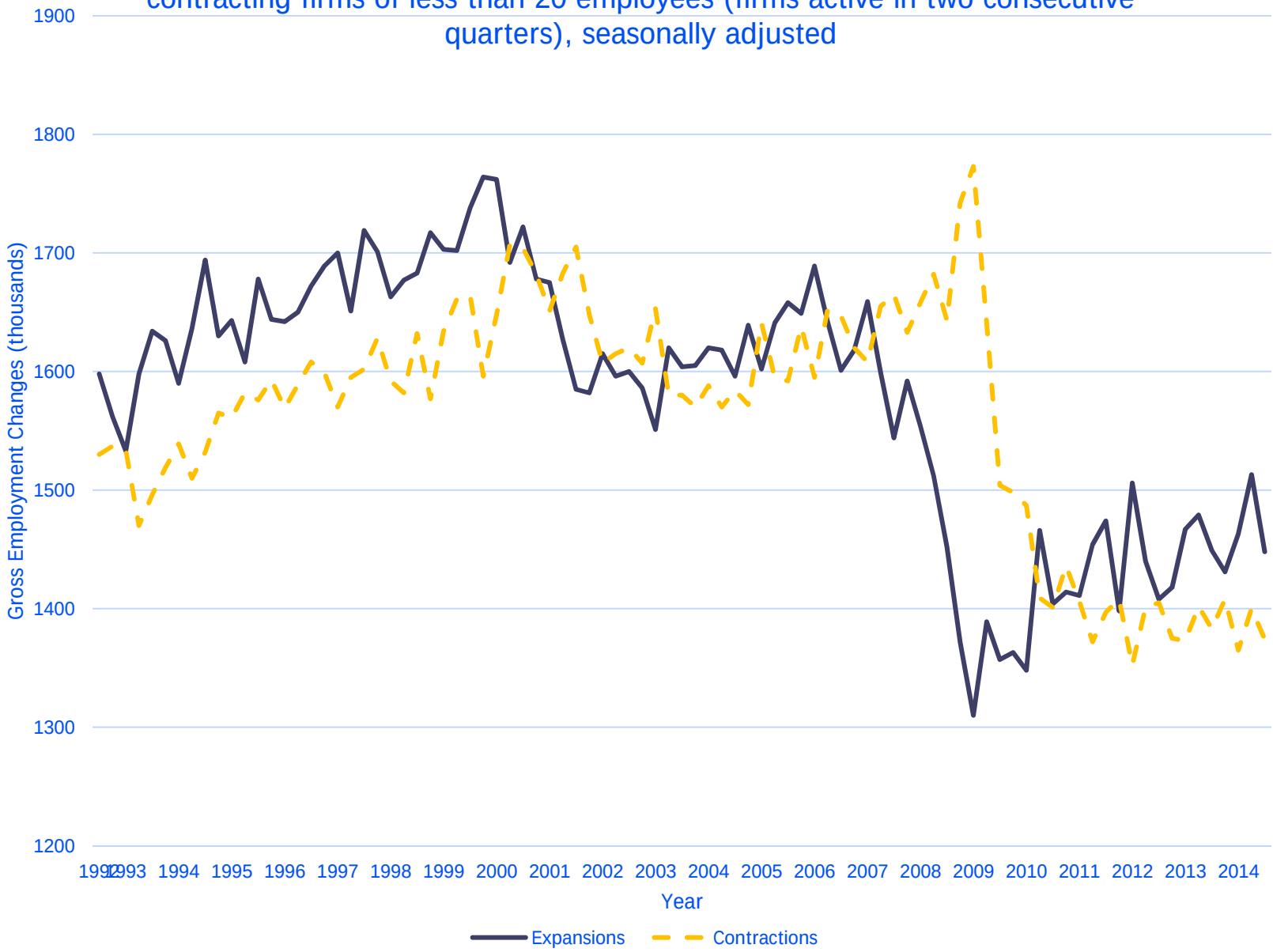
- Item-level missingness (Silverman, Tables 2 & 3 means; Woodward – Quickbooks "backfill" for recent months)

- Incomplete record linkage (external records; across time; consent to link). Especially important for "tall and thin" data sources ("unit rich - variable poor")

BLS

Chart 1: Survival rates: 1994 cohort

Chart 2: Components of gross job gains and gross job losses at expanding and contracting firms of less than 20 employees (firms active in two consecutive quarters), seasonally adjusted



B. Formal Methodological Questions (Continued)

d. Subpopulation classification variables missing or error-prone

e. Sampling error (often reduced if all sources use many units)

f. Measurement errors (per Woodward “Winsorization” of QuickBooks data; and typos in QuickBooks Online)

g. Processing effects (adjustments for quality issues; seasonality – Woodward, Silverman)

h. Extensions from data quality for surveys, administrative records (Davern, 2009, 2010; Zhang, 2009, 2011, 2012; others)

BLS

B. Formal Methodological Questions (Continued)

3. Empirical information & practical tools for quality issues in (B.2)?

   a. Assess qualitatively, quantitatively (bias analyses, variance estimation, sensitivity analyses)

   b. Tools to enhance quality: text analysis, imputation, calibration, small domain estimation

   c. Comparison with quality characteristics of other sources?

   d. Integration with other sources?

BLS

B. Formal Methodological Questions (Continued)

3. Empirical information (continued)

   e. Direct comparison with current microdata & estimates?
      Carroll: Compare Consumer Credit Panel demographics with Census Bureau profiles

   f. Temporal and cross-sectional stability of (a)-(e)?

   g. Formal inference for estimands?
      - Confidence intervals
      - Highly exploratory analyses (Madigan et al., 2014 caution)

BLS

B. Formal Methodological Questions (Continued)

4. Design issues

a. Fault-tolerant designs to manage risks (quality, continuity, other) – adapt approaches from engineering: Denning (1976), Zhang et al. (2005), Monkman & Schagaev (2013)

b. Supplementary survey to address quality issues in (3) ("bridge the gaps"):
- Standard ratio or regression estimators ("commercial" source provides auxiliary mean or total)
- Small area estimation with area-level models (Rao, 2003); "commercial means" provide area-level predictors
- Variants on the Current Employment Survey link-relative estimator (e.g., Woodward, 2015)

B. Formal Methodological Questions (Continued)

c. Optimal or robust designs of supplementary surveys?
(cf. Sarndal and Lundstrom, 2010; Marker, 2001)

i. Entirely new design?

- Carroll: Use of CCP as a sampling frame?
  Consider use of multiple-frame approaches to account for acknowledged coverage issues in CCP
- Similarities w/address-based sampling (Iannacchione, 2011; Valliant et al., 2014; West et al., 2015; others)

ii. Adapt current surveys

- Expand aggregation patterns to align with commercial-source aggregation patterns
- Expand to collect “commercial” X at microdata level

B. Formal Methodological Questions (Continued)

5. Relative information value of aggregated data or microdata

   a. Aggregated data only
      - Stand-alone usage or integration through regression estimation, area-level small domain estimators per (B.4))
      - Limited quality evaluation options (comparison with other aggregates; assessment of variable definitions)
      - Limited subpopulation breakouts

   b. Aggregated data mostly; microdata on a limited basis for model development and evaluation, other diagnostics)

   c. Microdata on a continuous (confidentiality protected) basis

BLS

B. Formal Methodological Questions (Continued)

6. Cost structures

   a. Data acquisition
      - Frame, administrative or commercial file(s):
        Level of aggregation (finer level ⇨ higher cost)?
      - Sampling & interview, if any (screening, full instrument)
      - Record linkage (and related cleaning – Winkler, 2009)

   b. Data edit, imputation, integration

   c. Production systems (design, implementation, maintenance)

   d. Additional components for fault-tolerant design (backups)

BLS

C. Management Approaches to Support Robust Development and Implementation

1. Legal, regulatory and contractual issues: source continuity and access (Silverman); quality; intellectual property (Fuller, 2014); informed consent

2. Resource allocation

   a. Federal statistics: Capital intensive (mostly intangible)

   b. Volatile features of input data sources
      - Increase depreciation rate on source-specific investments
      - Increased costs
      - Must target investments accordingly

   c. Human resources: skills, expectations, risk tolerance

C. Management Approaches (Continued)

3. Two approaches to use of a technology to expand product lines:

   a. Focus on information needs of a few stakeholders

      - Clear idea of intended uses; priorities on substantive and statistical features – incremental value of greater temporal & cross-sectional refinement, improved variables, improved quality
      - Spinoff benefits for others possible

   b. "Build it and they will come"

Blend of (a) and (b) complicated by "public goods" phenomena and limitations on market mechanisms

BLS

Summary

1. Three papers provide good illustrations of the opportunities and challenges in use of commercial data sources for official statistics.

2. Part of broad & deep societal reconsideration of (statistical) information over decades:

   - expectations on quality, cost, risk, credibility, accountability and access
   - tools to address
   - resource allocation: amount and mechanisms

BLS

Contact Information

John L. Eltinge
Associate Commissioner
Office of Survey Methods Research
www.bls.gov/ore
202-691-7404
eltinge.john@bls.gov

BLS