FCSM Working Group on Transparent Quality Reporting

Wendy L. Martinez

UNECE Workshop on Statistical Data Editing
September 2018

Disclaimer

- The findings and views expressed here are those of the author and do not necessarily reflect the policies of the Bureau of Labor Statistics (BLS) or the Federal Government.

- Acknowledge:
  - FCSM Working Group on Transparent Quality Reporting in the Integration of Multiple Data Sources and Workshop Speakers
  - Mark Prell (Economic Research Service), Jennifer Parker (National Center for Health Statistics), Chris Chapman (National Center for Education Statistics), Joe Schafer (Census Bureau), and John Eltinge (Census Bureau)

BLS

About the FCSM

- The U.S. Federal Committee on Statistical Methodology (FCSM) is an interagency committee dedicated to improving the quality of federal statistics. The mission of the FCSM is to:
  - Advise the Interagency Council on Statistical Policy (ICSP) on methodological and statistical issues that affect the quality of federal data;
  - Compile, assess, and disseminate information on statistical or survey methods and practices for federal statistical programs;
  - Provide recommendations on issues of statistical methodology such as … and dissemination of information that affect federal statistical programs and improve data quality, including transparency, timeliness, accuracy, relevance, utility, accessibility, and cost effectiveness

BLS

Subgroup on Data Quality

- Convened by U.S. Chief Statistician
- **Objective**:

  Review and synthesize work that may inform quality standards for use with statistical information products and services that are based on multiple data sources.

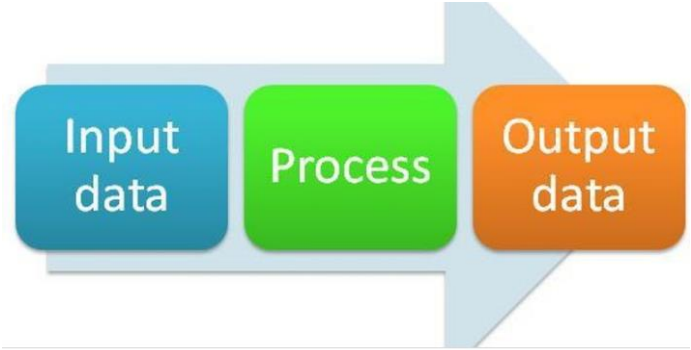
BLS

Questions to be Addressed

- In context of integrated data, what should be communicated to users of the final data products?
- **Fitness for use**:
  - Quality features when deciding to use a data source
  - Quality features to understand strengths and weaknesses of final product
- **Communication**: Best way to communicate quality features to diverse audience

BLS

Three Workshops



Workshop 1: Quality of Input Data

December 1, 2017

Workshop 2: Quality of Data Processing

January 25, 2018

Workshop 3: Quality of Output Data / Synthesis

February 26, 2018

BLS

WORKSHOP 1

DATA INPUTS

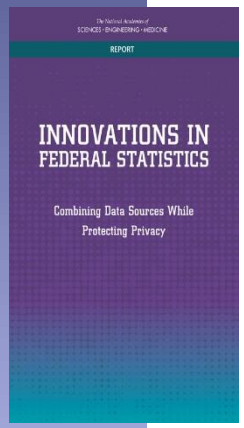
Focus of Workshop 1

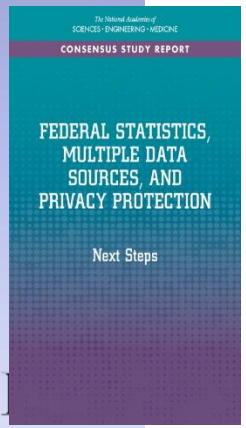
- Organizing principle of workshop 1 was to identify data quality standards and issues for non-sample survey data.
- Statistical agencies have a lot of guidance for quality metrics to report on survey data.
- Less guidance is available on quality metrics and reporting practices for administrative data and unstructured data
- Some guidance for the workshop was drawn from the CNSTAT report: "Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy"

BLS

Inputs from Expert Panels

- US National Academy of Sciences
  - Agencies should adopt a broader framework for statistical information than total survey error to include additional dimensions that better capture user needs, such as timeliness, relevance, accuracy, accessibility, coherence, integrity, privacy, transparency, and interpretability.
  - Agencies should outline and evaluate the strengths and weaknesses of alternative data sources on the basis of a comprehensive quality framework, and, if possible, quantify the quality attributes and make them transparent to users.
  - … and should focus more attention on the tradeoffs between different quality aspects… rather than focusing on accuracy.





Data Quality Dimensions

Dimension	# cited	Dimension	# cited	Dimension	# cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

Source: Wand and Wang (1996)

BLS

Multiple Sources of Data

	Data Source	
	Government	Private-Sector
Structured	censuses probability surveys	academic surveys market research surveys
	administrative records	commercial transactions bank and credit card records medical records
	other: traffic sensors weather sensor water quality sensors	e-commerce mobile phone location GPS
Semi-structured	web-scraped quantitative data web logs	logs, web logs text messages and e-mail
Unstructured	satellite images traffic videos blogs and comments	Facebook pictures and videos Internet searches

Source: Groves et al., *Innovations in Federal Statistics* (2017)

Key Ideas

- Data quality dimensions developed for survey data quality are applicable to non-survey data (admin data, commercial transactions, web-scraping)
- Statistical agencies already consider dimensions other than accuracy for non-survey data.
- New concepts of data quality might be needed to describe integrated data (especially unstructured, like text).

BLS

Key Ideas

- Key factors to consider on input data quality
  - Original purpose of collections
  - Level of documentation available
  - Issues with private data providers needing to protect intellectual property and stakeholder confidentiality, and not being consistent over time
- Issues with respect to admin data
  - Purpose of the original collection
  - Technology used to collect data and limitations of the technology
  - How technology changes affect data comparability over time
  - Speed at which data are available

BLS

WORKSHOP 2

DATA PROCESSING

Prioritization of Topics

Topic	Priority (L/H)
1. Record linkage	**H**
2. Multiple frames	L
3. Statistical matching / data fusion	**H**
4. Combining aggregate statistics or estimates (as in small area estimation)	L
5. Dimension reduction / feature extraction	L
6. Harmonization across data sources	**H**
7. Edit and imputation	L
8. Adjusting for representativeness	L
9. Estimation	L
10. Disclosure avoidance	**H**
11. Provenance / curation of metadata	L

BLS

Record Linkage

Key Ideas

- Techniques for "entity resolution" with noisy identifiers
- Computationally intensive
- Traditional methods (e.g. Fellegi-Sunter) become intractable with multiple data sets
- Difficulties abound, yet many agencies are already doing it, even in large scale projects

Take Away Messages

- Well established quality metrics do exist (e.g. precision, recall)
- Importance of high quality "truth sets", for supervised learning and for quality evaluation
- Errors in original source data, plus mistakes in matching, all add up
- Methods for assessing how errors impact final estimates is still in its infancy

BLS

Harmonization

Key Ideas

- Using a survey to adjust/improve estimates from administrative records
- Combining data from multiple surveys of similar populations and topics to add value to data products
- Modeling techniques for "change of support" to generate estimates for different levels of aggregation in space and time

Take Away Messages

- Harmonization is hard work, but can be made simpler if survey designers plan for it
- Estimates for different levels of aggregation may have very different quality characteristics, even if the data sources are the same (MAUP), but theory exists for how to minimize the error
- Need high-quality data sets where true matches are known

BLS

Matching, Modeling & Imputation

Key Ideas

- Current statistical matching makes strong assumptions that are not directly testable
- Moving away from matching to explicit (e.g. regression-based) models doesn't solve that problem, but makes it easier to perform sensitivity analyses
- Explicit models allow us to use auxiliary datasets as "glue" to estimate those inestimables

Take Away Messages

- Bayesian multivariate models are a promising theoretical framework for combining data sets in this (non record-linkage) realm
- These models do not yet incorporate our understanding of different quality profiles of different data sources
- These techniques can be expanded to do so; this is a promising area for future research

BLS

WORKSHOP 3

DATA OUTPUTS

BLS

European Perspective

- Chap 3: Relevance
- Chap 4: Accuracy
- Chap 5: Timeliness
- Chap 6: Accessibility & Clarity
- Chap 7: Comparability
- Chap 8: Coherence

https://ec.europa.eu/eurostat/web/lfs/publications/quality-reporting

Mathematica Paper – Key Findings

- **Transparency in the Reporting of Quality for Integrated Data: A Review of International Standards and Guidelines,**
- Only one national statistical organization—Stats NZ—has developed a quality framework explicitly designed to address integrated data.
- Eurostat's quality standards and guidelines, which apply to most of Europe and are perhaps the most extensive, deal with integrated data to a much more limited degree and instead focus on quality more generally.
- Many of the quality assurance frameworks and the associated standards and guidelines reviewed in this report are associated with extensive prescriptions for quality assessments and their communication to data users in detailed quality reports. The volume and types of information requested in Eurostat quality reports bears substantial resemblance to what was included in the quality profiles prepared by a number of U.S. federal agencies in the 1990s and early 2000s but, for multiple reasons, not continued.
- Efforts to deal with quality aspects of administrative data are much farther along than efforts to deal with the quality of other forms of Big Data.

BLS

WRAP-UP

Research Topics

- Develop tools/methods to
  - Deal with limitations of alternative data sources
  - Handle errors in non-survey data sources
  - Improve assessment and modeling of record linkage quality
  - Assess overall level of uncertainty in estimates based on integrated data

BLS

Research Topics

- Related to the previous areas, work on the following:
  - Development of open-source or open-access software addressing the previous items
  - Development of standardized toolsets for creating integrated-data products

BLS

Summary Messages

- Data producers must be transparent about each step:
  - Original need to collect data
  - Harmonization steps
  - Matching procedures
  - Models used and assumptions
  - Evaluation techniques used
  - How privacy was maintained
- Decisions captured in metadata – users can judge utility

BLS

More Workshops

- Metadata – September 14
- Geospatial Data & Data Quality – October 26
- Sensitivity Analysis – October or November
- You are welcome to join via WebEx – send email – contact info to follow

BLS

Links to Workshops

- Links to slides

http://washstat.org/presentations/

- Links to videos – see playlists

https://www.youtube.com/channel/UCblmtGTydPN4978pSy55b3A/playlists

BLS

Questions for Discussion

- What are your experiences working with a data quality criterion and methodological tools or approaches?
- For what applications have you found them to be useful?
- What feedback have you received from your stakeholders?
- Please share recommendations, software used, tech reports, and case studies.

BLS

References

- Beyond Accuracy: What Data Quality Means to Data Consumers Author(s): Richard Y. Wang and Diane M. Strong Source: *Journal of Management Information Systems*, Vol. 12, No. 4 (Spring, 1996), pp. 5-33. http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf
- **Transparency in the Reporting of Quality for Integrated Data: A Review of International Standards and Guidelines,** John L. Czajka and Mathew Stange , https://www.mathematica-mpr.com/our-publications-and-findings/publications/transparency-in-the-reporting-of-quality-for-integrated-data-a-review-of-international-standards
- National Academies of Sciences, Engineering, and Medicine. 2017. *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24893 or http://nap.edu/24893
- National Academies of Sciences, Engineering, and Medicine. 2017. *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24652

BLS

Contact Information

Wendy Martinez

Office of Survey Methods Research
Bureau of Labor Statistics
00-1-202-691-7400
martinez.wendy@bls.gov

Prioritizing the Topics

What topics are

- substantially more complicated or qualitatively different when combining multiple data sources?
- less familiar to statisticians and methodologists?
- not well covered by existing standards for quality and transparency?
- not as well covered by existing literature (e.g. on Small Area Estimation or Total Survey Error)?

BLS