Input Data Quality for Integrated Data Products

Chris Chapman
National Center for Education Statistics (NCES), U.S. Department of Education

FedCASIC April 19th, 2018

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

This presentation is intended to promote ideas. The views expressed are part of ongoing research and do not necessarily reflect the position of the U.S. Department of Education

National Academies – Committee on National Statistics (CNSTAT) Recommendations

- CNSTAT released "Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps" in 2017
  https://www.nap.edu/download/24893
- Two recommendations helped inform FCSM work
  (6.1) Federal statistical 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.

CNSTAT Recommendations

- (6-2) Federal statistical 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.

  Agencies should focus more attention on the tradeoffs between different quality aspects, such as, trading precision for timeliness and granularity, rather than focusing primarily on accuracy.

Reviewed a wide range of quality frameworks

- FCSM reviewed extant data reporting frameworks to evaluate approaches to providing information beyond total-survey-error-related information
- Many similarities across them
- Common theme was to provide information needed to secondary data users to evaluate fitness for use for their purposes

Structured initiative around 3 integration topics

- What to report to help consumers evaluate quality of data that are being integrated
- What to report to help consumers evaluate the quality of how data were integrated
- What to report to help consumers evaluate the quality of the resulting integrated data product

Input data quality reporting

- Goal is to consider what information to provide users to evaluate data quality, and not to improve evaluation techniques
- We have a wealth of reporting metrics for survey data
- We lack consistent reporting metrics for non-survey data

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)

Common concepts

- Discussed administrative record data, semi-structured data like quantitative data from web scraping, and unstructured data like those from medical images
- Theme was that these types of input data were evaluated in terms of quality like survey data quality

Common concepts

- All of the presentations stressed the importance of providing end users with information about why the data were collected
- Purpose of the collection was considered central for end-users to evaluate fitness for use for their own work

Significant questions

- Are there unique data quality issues for non-survey data that lack analogies in survey metrics?
- Related questions center around what to report about data quality when data lack survey-industry standard documentation
- How to convey information about data from private sector data when vendors need to protect trade secrets

Significant questions

- More broadly, what information should be provided beyond that needed to evaluate total survey error?
  - CNSTAT provided excellent recommendations and we identified other important data quality reporting dimensions
  - How do agencies work to report on more of these other data quality dimensions?

Multiple dimensions in use

Quality report for ESS Labor Force Survey 2015 (2017)

Ch 3. Relevance

Ch 4. Accuracy

Ch 5. Timeliness

Ch 6. Accessibility and Clarity

Ch 7. Comparability

Ch 8. Coherence