Findings from the Integrated Data Workshops hosted by the Federal Committee on Statistical Methodology and Washington Statistical Society

Alexandra Brown, University of Maryland

Co-Authors:
Katharine G. Abraham, University of Maryland
Frauke Kreuter, University of Maryland
Andrew Caporaso, University of Maryland

Overarching Questions

- How does integrated data parallel traditional survey data in terms of quality trade-offs?
- Fitness for Use: What particular quality features are going to be important for different data users to understand when considering whether to use a particular data source?
- Communication with Stakeholders: What is the best way to communicate those predominate quality features to stakeholders who may be technical specialists, substantive data users, or the general public?

Workshop 1: Quality of Input Data

Session One: Structured and Administrative Data Sources

- Steven B. Cohen, RTI – "The Utility and Limitations in Administrative Data for Medical Care Expenditure Analysis"
- Michael Berning and David Sheppard, U.S. Bureau of the Census – "Quality of Administrative Records as Source Data"
- Bonnie Murphy and Crystal Konny, Bureau of Labor Statistics – "Quality Considerations for Administrative Data Used for the Producer Price Index (PPI) and Consumer Price Index (CPI) Development"
- Mary Muth, RTI – "Assessment of Commercial Store and Household Scanner Data: Methods, Content, and Cautions"

Workshop 1: Quality of Input Data

Session Two: Less Structured Data Sources

- Dr. Peter Elkin, University of Buffalo – "The Improvement in Sensitivity and Often Specificity when Adding Unstructured to Structured Data"
- David Johnson, USDA NASS – "Data Quality of Satellite Imagery for Studying Complex Systems"
- Subrat Mahapatra, Maryland Department of Transportation – "Sensing Data Quality in Sensor-Based Data"
- Roberto Rigobon, National Bureau of Economic Research & MIT – "Web-scraped Data, Consideration of Quality Issues for Federal Statistics"

Workshop 1: Quality of Input Data

Themes:

- ✓ There are some parallels with traditional survey data, but not a ton
- ✓ Fitness for use is very project specific
- ✓ Institutions assess the quality trade-offs differently
- ✓ Data processing prior to integration is a priority
- ✓ Reduced control, but can increase coverage and timeliness
- ✓ Transparency and communication is important but maybe limited

Workshop 2: Processing Data

Session One: Record Linkage

- Rebecca Steorts, Duke University – "Entity Resolution: Measuring and Reporting Quality"

Session Two: Harmonization of Data Across Sources

- Ben Reist, U.S. Bureau of the Census – "Leveraging Survey Methods to Improve Administrative Record Estimates"
- Don Jang, NORC – "Data Harmonization in Survey Data Integration"
- Scott Holan, University of Missouri – "Recent Advances in Spatial and Spatio-Temporal Change of Support for Official Statistics"

Session Three: Combining Data by Statistical Matching, Imputation and Modeling

- Jerry Reiter, Duke University and U.S. Bureau of the Census – "Blending Data Through Statistical Matching, Modeling, and Imputation"

Session Four: Disclosure Avoidance

- Latanya Sweeney, Harvard University – "The Elusive Sweet Spots of Privacy and Utility"

Workshop 2: Processing Data

Themes:

- ✓ Transparency is important, but data producers need to be careful
  - ✓ Model robustness, evaluation metrics used, comparisons made, work that isn't going well, variables used for matching, model assumptions, reveal potential biases
- ✓ Data harmonization
- ✓ Data privacy
- ✓ Ethical treatment of data

Workshop 3: Quality of Output Data

Session One: Break in Series

- Lynn Langton, Bureau of Justice Statistics – "Identifying and Addressing the Breal (Blip) in Series"

Session Two: Combining Data from Disparate Sources

- Trivellore Raghunathan, University of Michigan – "Combining Information from Multiple Data Sources: Challenges and Opportunities"

Session Three: Frameworks for Assessing Data Quality

- Paul Biemer, RTI – "Assessing and Improving the Accuracy of Estimators from Blended Data"
- John Czajka, Mathematica – "Transparency in the Reporting of Quality for Integrated Data: International Standards"

Session Four: Summary

- Frauke Kreuter, Joint Program in Survey Methodology

Workshop 3: Quality of Output Data

Themes:

- ✓ Transparency of data quality, but at what level and for whom?
- ✓ Combining data from different sources
- ✓ Communication is needed across agencies and across countries
- ✓ Statistical agencies need to think outside of the box to produce the statistics that the public is demanding

Big Takeaways

- ✓ Transparency is important, but what should data producers be transparent about? And who is the target audience?
- ✓ Quality may be project specific.
- ✓ Federal agencies need to communicate with each other to determine a unified path forward.
- ✓ Data matching and harmonization needs to be a focus for data producers and users.
- ✓ Disclosure risks and data privacy cannot be taken lightly.