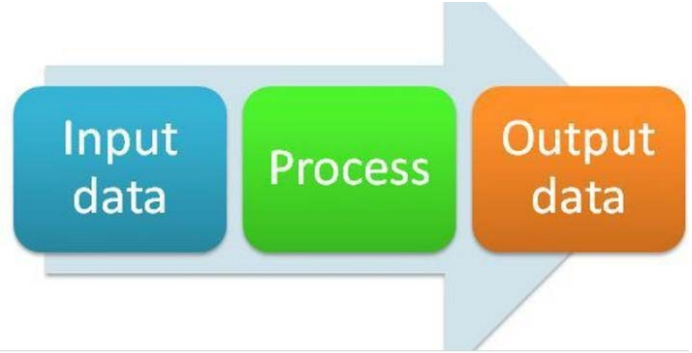
Recap of Workshop 2

Joe Schafer
U.S. Census Bureau
Federal Committee on Statistical Methodology
Joseph.L.Schafer@census.gov

Opinions expressed are those of the author and are not necessarily the views or policies of the United States Census Bureau.

Three Workshops

Reporting on Quality of Integrated Data



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

Prioritization of “Data Processing” 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 SAE)	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

Session 1: Record Linkage

- Main presentation by Rebecca Steorts (Duke University) (40 min)
- Comments and discussion by Bill Winkler (Census Bureau) (10 min)

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 decks" (e.g. hand-matched subsamples), both for supervised learning and for quality evaluation
- Errors in original source data sets, plus mistakes in matching, all add up
- Methodology for assessing how these errors impact final estimates is still in its infancy

Session 2: Harmonization

- Presentations by Ben Reist (Census Bureau), Don Jang (NORC), and Scott Holan (U. Missouri)

Key Ideas

- Using a survey to adjust/improve estimates from administrative records
- Combining data from multiple surveys of similar populations and topics (e.g. college graduates) 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

- Varying quality profiles are often *primary motivation* for combining data sources
- 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

Session 3: Statistical Matching, Modeling, Imputation

- Presentation by Jerry Reiter (Duke U.), with discussion by Ed Mulrow (NORC)

Key Ideas

- Statistical matching (as most have been doing it) makes strong assumptions (e.g. CIA) that are not directly testable
- Moving away from matching to explicit (e.g. regression-based) models makes 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 datasets 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

Session 4: Disclosure Avoidance

- Presentation by Latanya Sweeney (Harvard U.), with discussion by John Abowd (Cornell/Census)

Key Ideas

- Not necessarily a direct tradeoff between data utility and confidentiality protection; “sweet spots” do exist
- Current “best practices,” especially at the state level, are still vulnerable to re-identification
- Need for continuous improvement

Take Away Messages

- Adding random noise is necessary to overcome consequences of database reconstruction theorem; its error properties are quantifiable, and we can be fully transparent about the method
- Risk depends on properties of a given dataset, plus everything else that has already been released (privacy budget)
- Commission on Evidence-Based Policymaking provides sound guidance