Overview of Today’s Workshop

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
What “Data Processing” Entails in the Context of Integrated Data

1. Record Linkage
   Exact match, probabilistic match, privacy-preserving record linkage

2. Using Multiple Frames
   Drawing samples from two or more frames to improve coverage or reduce costs

3. Statistical Matching / Data Fusion
   Joining two or more non-overlapping samples by variables shared in common, then applying modeling or imputation techniques to handle the missing values

4. Models for Combining Aggregate Statistics or Estimates
   Combining estimates from different sources at national, subnational or subpopulation levels, as in Small-Area Estimation
What “Data Processing” Entails in the Context of Integrated Data

5. Dimension Reduction / Feature Extraction
   Techniques for summarizing unstructured data (e.g. images, freeform text)

6. Harmonization
   Combining information across datasets in the presence of mode effects, differing definitions or granularities (e.g. variables measured at differing time periods or levels of geography)

7. Edit and Imputation
   Other types of cleaning after data sources are combined

8. Adjusting for Representativeness
   Making combined data more representative of the intended population (reweighting, benchmarking, poststratification, calibration, ...)
What “Data Processing” Entails in the Context of Integrated Data

9. Estimation
   Computing estimates of population quantities and associated measures of uncertainty

10. Disclosure Avoidance
    Techniques for preventing re-identification or de-anonymization of individual records

11. Provenance and Curation of Metadata
    Preserving information about data sources, dictionaries, audit trails
Prioritizing the Topics

Which of these 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)?

• not already covered in Workshop 1?
<table>
<thead>
<tr>
<th>Topic</th>
<th>Priority (L/H)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Record linkage</td>
<td>H</td>
</tr>
<tr>
<td>2. Multiple frames</td>
<td>L</td>
</tr>
<tr>
<td>3. Statistical matching / data fusion</td>
<td>H</td>
</tr>
<tr>
<td>4. Combining aggregate statistics or estimates (as in SAE)</td>
<td>L</td>
</tr>
<tr>
<td>5. Dimension reduction / feature extraction</td>
<td>L</td>
</tr>
<tr>
<td>6. Harmonization across data sources</td>
<td>H</td>
</tr>
<tr>
<td>7. Edit and imputation</td>
<td>L</td>
</tr>
<tr>
<td>8. Adjusting for representativeness</td>
<td>L</td>
</tr>
<tr>
<td>9. Estimation</td>
<td>L</td>
</tr>
<tr>
<td>10. Disclosure avoidance</td>
<td>H</td>
</tr>
<tr>
<td>11. Provenance / curation of metadata</td>
<td>L</td>
</tr>
</tbody>
</table>
Features of Workshop

- Speakers from academia, federal agencies, research firms
- Many are participating from remote locations (WebEx)
- Presentations may be less formal than typical seminar
- Focus less on methods, more on quality issues
- Extra time in each session for questions, comments, open discussion
- Rapporteurs from JPSM will synthesize what we learn
Session 1: Record Linkage

- Main presentation by Rebecca Steorts (Duke University) (40 min)
- Comments and discussion by Bill Winkler (Census Bureau) (10 min)
- Questions, comments, discussion (25 min)